Nutzung von anonymisierten Einzelangaben aus Daten der amtlichen Statistik

Bedingungen und Möglichkeiten

Band 5 der Schriftenreihe
Forum der Bundesstatistik

Herausgeber: Statistisches Bundesamt, Wiesbaden
Verlag: W. Kohlhammer, Stuttgart und Mainz

CIP-Kurztitelaufnahme der Deutschen Bibliothek

Nutzung von anonymisierten Einzelangaben aus Daten der amtlichen Statistik: Bedingungen und Möglichkeiten / Hrsg.: Statist. Bundesamt, Wiesbaden.
– Stuttgart; Mainz: Kohlhammer, 1987.
(Schriftenreihe Forum der Bundesstatistik; Bd. 5)
ISBN 3-17-003338-7

NE: Deutschland <Bundesrepublik> /
Statistisches Bundesamt; GT

Erschienen im April 1987

Preis: DM 16,50
Bestellnummer: 1030405-87900
ISBN ~~3-17-003338-7~~
9783824604168

Vorwort

Unter den rechtlichen Bedingungen, die insbesondere das Urteil des Bundesverfassungsgerichts vom 15. Dezember 1983 zur Volkszählung geschaffen hat, gehört die Auflösung des Spannungsverhältnisses zwischen Statistikgeheimnis und damit Sicherung der Anonymität statistischer Einzeldaten auf der einen Seite und dem berechtigten Bedürfnis der Wissenschaft zu möglichst intensiver Nutzung der von der amtlichen Statistik erhobenen Informationen zu den schwierigen, aber vordringlichen Aufgaben, die die amtliche Statistik zu lösen hat.

Im Rahmen dieser Arbeiten veranstaltete daher das Statistische Bundesamt zusammen mit der Arbeitsgemeinschaft Sozialwissenschaftlicher Institute (ASI) vom 3. bis 5. März 1986 in Wiesbaden ein wissenschaftliches Kolloquium über die Bedingungen und Möglichkeiten der Nutzung von anonymisierten Einzelangaben aus Daten der amtlichen Statistik. Ziel der Tagung war es, den gegenwärtigen Stand der Forschung auf dem Gebiet der Anonymisierung von Einzelangaben sichtbar zu machen und die Umsetzung der theoretischen Erkenntnisse in die Praxis, unter Berücksichtigung der datenschutzrechtlichen Rahmenbedingungen, zu fördern.

Wissenschaftler verschiedener Disziplinen aus Universitäten und Instituten, Statistiker und Datenschutzbeauftragte erörterten die Problematik unter den verschiedensten Gesichtspunkten, Teilnehmer aus Großbritannien und den Vereinigten Staaten brachten die Erfahrungen aus ihren Ländern mit ein.

Dank gilt insbesondere Herrn Prof. Dr. Allerbeck, der mit viel Engagement und Sachkunde das Kolloquium moderierte und in Zusammenarbeit mit dem Institut für Methodenforschung im Statistischen Bundesamt die Last der Vorbereitung dieser Veranstaltung getragen hat. Dank gilt auch der Stiftung Volkswagenwerk für die finanzielle Förderung des Kolloquiums. Vielmals danken möchte ich ferner allen Referenten und Teilnehmern dieses Kolloquiums. Sie alle haben der Veranstaltung zum Erfolg verholfen.

Die vorliegende Veröffentlichung enthält neben den Referaten, die auf dem Kolloquium gehalten wurden, eine umfassende Darstellung der abschließenden Podiumsdiskussion.

Wiesbaden, im April 1987

Der Präsident des Statistischen Bundesamtes

Egon Hölder

Inhalt

Seite

Prof. Dr. Klaus Allerbeck
Universität Frankfurt

Einführung in das Thema

Fortschritte der Akzeptanz der Statistik

Statistische Zählungen — wie vor allem Volkszählungen — und Kontroversen um die Tätigkeit der Statistik sind kein Kind der Moderne, sondern schon aus biblischen Zeiten bekannt.[1]) Kennzeichen der Moderne ist allerdings, daß die Erhebung von statistischen Informationen als Merkmal und Voraussetzung gesellschaftlichen Fortschritts galt.[2]) In Europa ist die Entwicklung der Statistik im Gefolge der Aufklärung und mit dem Bemühen um eine aktive, gesellschaftsgestaltende Politik eng verbunden; als nur eine Illustration hierfür ein Zitat aus dem Schreiben von König Max II. (31. 10. 1856) an den Vorstand des Statistischen Bureaus Bayerns, welches u. a. besagte:

> „Ich wünsche eine Statistik Bayerns über die Hauptzweige der National- und Volkswirtschaft und der inneren Verwaltung, um bemessen zu können, in welcher Beziehung man in Bayern Ursache hat, zufrieden zu seyn und in welcher anderen Nachhülfe notwendig wäre."

Abgesehen von der Sprache des vorigen Jahrhunderts scheint dies auch heute noch fast aktuell — in mehrfacher Hinsicht. Der Schwerpunkt der Frage richtet sich auf Zustände, nicht auf Bewegungen. Das Interesse an Untersuchungen des Wandels war gering. Dies ist nicht trivial. Denn auch heute kann unsere Zufriedenheit mit den Aussagen über Zustände größer sein als mit den Daten über Veränderungen, sei es über Veränderungen von Kollektiven oder von einzelnen.

Untersuchungen über Veränderungen, auch heute noch die Ausnahme, begannen im Bereich der Bevölkerungsentwicklung, wobei es um die Veränderungen von Kollektiven wie der Gesamtbevölkerung ging. Sie hatten Ergebnisse, die die Zeitgenossen erstaunten. Für uns ist heute das Bevölkerungswachstum dieses Jahrtausends eine selbstverständliche Tatsache. Aber für Montesquieu[3]) schien das genaue Gegenteil richtig: ihm erschien ein Rückgang der Bevölkerung auf weniger als ein Zehntel der Weltbevölkerung im Altertum selbstverständlich, und auch zu Zeiten von Karl dem Großen hatte nach seiner Einschätzung Europa mehr Einwohner als zu seiner Zeit. Montesquieu fand in Frankreich mit diesen Aussagen Zuspruch.[4]) Erst durch Zählungen und Vergleiche konnten die Tatsachen ermittelt werden. Doch Vergleiche waren nicht das Hauptziel damaliger Erhebungen, und sie sind es auch heute vielfach nicht.

Tatsächlich sind heute die Zählungen als solche, mit ihren Ergebnissen für große Aggregate, wie die Bevölkerung der Bundesrepublik Deutschland oder eines großen Bundeslands,

---

[1]) 2. Samuel 24.
[2]) David Glass, Numbering the people, Farnborough 1973.
[3]) Montesquieu, Lettres Persanes, CXII, 1764, Bd. 5, pp. 286—289 und ders. „De l'Esprit des Lois", Buch 23, Kap. 24, Oeuvres . . . Bd. III, S. 23 ff.
[4]) Vgl. Glass, a.a.O., S. 21.

ganz unumstritten. Umstritten sind die Individualdaten, die in diese Zählungen eingehen und eingehen müssen: ist ihre Anonymität, ihre absolute Geheimhaltung, tatsächlich gesichert? Kann der einzelne sicher sein, daß seine Angaben in denen der Masse anderer Merkmalsträger gleichsam untergehen und daß sie nicht in irgendeiner Weise für den Verwaltungsvollzug genutzt werden können?

Vergleichbarkeit und Verknüpfbarkeit statistischer Daten

Historisch hat es also eine Problemverschiebung gegeben: die Aggregatdaten sind mittlerweile allerorten erwünscht — vielleicht noch nicht so wie in den USA, wo der zehnjährliche Census in der Verfassung verankert ist —, aber die Individualdaten gelten heute in vielen Augen als besonders schützenswert. Die Frage ist indes: Schutz vor wem, und wie? Dieses Symposion behandelt den Zugang der Wissenschaft zu anonymisierten Individualdaten.

Aus der Sicht der Sozialforschung hat dies zwei Aspekte: die Möglichkeiten des Vergleichs und die Möglichkeiten der Verknüpfung, welche beide möglichst „reine" und vielfach Individualdaten voraussetzen. In einer Zeit, in der die Datenverarbeitung weit fortgeschritten ist, stellt sich die Frage, ob durch die Verknüpfung von Daten nicht ursprünglich „harmlose" Informationen zu Tatbeständen werden, die der Geheimhaltung unterliegen müssen.

Erhaltung der Einzelinformation

Die Forderung der Wissenschaft nach Erhaltung von möglichst vielen Informationen, auch um spätere, bei Datenerhebung nicht antizipierte Analysen zu ermöglichen, ist nicht neu. Das Verbot zu frühzeitiger Zusammenfassung von Einzeleingaben galt schon immer. Dies belegt die Aussage von Mayr's vor 90 Jahren, der als „allgemeine Regel fest(hielt) . . . , in den Urtabellen möglichst weit in der Auseinanderhaltung zu gehen; Zusammenlegungen sind später leicht bewerkstelligt, während eine später als wünschenswert erkannte Auseinanderhaltung, welche bei der Ausbeutung übersehen worden ist, nachträglich bei großen statistischen Erhebungen entweder gar nicht mehr oder nur mit unverhältnismäßigen Opfern durchgeführt werden kann. In wissenschaftlicher Beziehung kommen hier namentlich auch die Rücksichten der vergleichenden Statistik in Betracht".[5]) Für „Zusammenfassungen der reichgegliederten Ergebnisse in gleichartige Rahmen"[6]) ist möglichst wenig Aggregation zwingende Voraussetzung.

Heute gilt dies noch verstärkt gegenüber 1895, dem Erscheinungsdatum von Georg von Mayrs „Statistik und Gesellschaftslehre".

Aus der Sicht der Wissenschaft, und der der Arbeitsgemeinschaft Sozialwissenschaftlicher Institute, ist völlig klar, daß heute die Fortschritte der Datenverarbeitung und der Daten-

[5]) Georg von Mayr, Statistik und Gesellschaftslehre, Bd. 1, Theoretische Statistik, Freiburg und Leipzig 1895, S. 69.
[6]) Von Mayr, a.a.O., S. 69.

banktechnologie genutzt werden müssen, um dringliche Fragen wissenschaftlich zu erkunden, um so mehr, als jetzt das technische Instrumentarium zur Verfügung steht. Dazu sind auch die Daten erforderlich, welche vielfach zwar für Einzelfälle verknüpfbar sind, indes mit dem Ziel, sinnvolle Aussagen zu ermöglichen, welche sich gerade nicht auf Einzelfälle beziehen.

Zielsetzungen der Sozialforschung mit Datenschutz übereinstimmend

Wenn die Sozialforschung sich über die Anforderung des Datenschutzes äußert, muß sie dabei voranstellen, daß ihre Zielsetzung mit der der heutigen Vorstellung des Datenschutzes identisch ist. Die Sozialwissenschaften haben, so merkwürdig dies für den Laien klingen mag, kein Interesse an einzelnen Individuen. Sobald die Daten einmal erhoben sind, mittels eines Fragebogens zum Beispiel, interessiert der Befragte als einzelner Befragter nicht mehr; ja, wenn es sich um eine Stichprobenerhebung handelt, interessierte der einzelne Befragte nicht einmal bei der Befragung selbst als konkreter einzelner, sondern nur als ein zufällig ausgewählter Repräsentant eines Kollektivs, dem er angehörte. Die Aussagen der Wissenschaft beziehen sich, wenn sie mit statistischen Methoden arbeiten, nicht auf einzelne, namentlich oder sonst irgendwie identifizierbar Befragte. Es besteht kein Interesse daran zu wissen, welche bestimmte Person arm ist; es besteht dagegen Interesse daran, aussagen zu können, wieviel Prozent derjenigen, die zu einem bestimmten Stichtag als arm klassifiziert wurden, noch ein Jahr darauf als arm zu gelten hatten. Die Sozialwissenschaften haben keinerlei Interesse daran, einzelne gläserne Menschen entstehen zu lassen. Die Interessenlage der Sozialwissenschaften ist so, daß sie von ihren Zielen her einen Konflikt mit dem Datenschutz gar nicht haben.

Statistikgeheimnis als Anfang des Datenschutzes

Statistik wie Sozialwissenschaften haben eine langjährige Tradition, die Identität ihrer Befragten und ihre Anonymität zu schützen. Die Tradition des Statistikgeheimnisses, welche auch ihren gesetzlichen Niederschlag gefunden hat, ist älter als das Auftauchen der Vorstellung des Datenschutzes. In der Umfrageforschung gibt es eine Tradition, die so alt ist wie die Umfrageforschung selbst, den Befragten die vollständige Anonymität ihrer Angaben zuzusichern. Auch hier also Deckungsgleichheit von Forschungspraxis und Anforderungen des Datenschutzes. Wenn dies so ist, wo liegt dann ein Problem in dem Dreieck von Sozialwissenschaften, amtlicher Statistik und Datenschutz?

Technischer Fortschritt erleichtert Datenschutz

Mit den Gefahren, die durch die moderne Datenverarbeitung entstanden sind, gibt es mittlerweile nicht zuletzt durch die Mikroelektronik auch Schutztechniken, welche den Datenschutz sichern. Die Vorstellung, Bereitstellung von Daten für wissenschaftliche Forschung bedeutete deren Bereithaltung in Mehr-Benutzer Timesharingsystemen, wie sie für Universitätsrechenzentren charakteristisch sind, mit unwägbaren Folgen für den Datenschutz,

traf allenfalls für die Vergangenheit allgemein zu; heute ist sie überholt. Heute ist es möglich, zumindest die Merkmale, welche der Verknüpfung dienen (also die Identifikation ermöglichen könnten), auf Disketten vollständig aufzubewahren und buchstäblich hinter Schloß und Riegel oder im Panzerschrank aufzubewahren, so daß alle erdenklichen Zugriffshindernisse geschaffen sein können. Auch können auf Mikrocomputern inzwischen elaborierte, der Kryptographie entstammende Verschlüsselungstechniken verwandt werden, die es gewährleisten, daß nur Befugte Datenbestände zur Kenntnis nehmen können. Sinnreiche Kombinationen solcher technischer Vorkehrungen sind möglich und werden auch im Bereich der Wissenschaft praktiziert, so daß es heute an den technischen Voraussetzungen nicht fehlt, um auch innerhalb der wissenschaftlichen Forschung einen umfassenden gesicherten Datenschutz zu praktizieren, selbst dann, wenn es sich um sensitive Merkmale handelt. — Dies sind relativ neue technische Entwicklungen, welche nur illustrieren, wie rasch der technische Fortschritt in diesem Bereich ist und Fragen und Probleme obsolet werden läßt, welche der Siegeszug der Groß-EDV geschaffen hatte.

Schutz für Angaben einzelner

Wesentliche Schwierigkeiten in diesem Dreieck entstehen durch die großen Schwierigkeiten angemessener rechtlicher Normierung, die Anspruch auf eine gewisse Dauerhaftigkeit haben muß. Dieser Band berichtet im Detail über diese Schwierigkeiten und die Bemühungen, angemessene rechtliche Regelungen zu finden. Dies ist ein aus Gründen der Systematik der Gesetzgebung schwieriges Unterfangen. Die Ratio der Datenschutzgesetzgebung ist ja, den gläsernen Menschen zu verhindern, der dadurch entstehen könnte, daß eine Vielzahl von Informationen über einzelne befugt oder unbefugt zusammengefaßt werden könnte. Jedwede Weitergabe von Daten zu Zwecken, für die sie nicht erhoben wurden und die denen, die die Auskünfte gaben, nicht bekannt waren, muß verpönt sein. Die Möglichkeiten eines Registerabgleichs sind bekannt, wenn sie oft auch überschätzt werden; es besteht die Gefahr eines Herrschaftswissens neuer Art. Da in der modernen Gesellschaft der Datenschatten eines Individuums immer länger wird, ist es von größter Bedeutung, Freiheiten trotz dieses Schattens zu bewahren. Es muß gesichert sein, daß niemand, keine Person, keine Institution, auf datentechnischem Weg über Individuen irgend etwas erfahren kann, das ihnen nicht ausdrücklich selbst anvertraut wurde. Die entstehende (partielle) Unkenntnis ist gewollt und richtig.

Informationsbedarf über Kollektive

Die moderne Gesellschaft hat indes einen immensen Wissensbedarf über Kollektive. Immer weniger Planungsentscheidungen können aufgrund persönlicher Erkenntnisse getroffen werden. Immer mehr Entscheidungen benötigen zutreffende Daten aus eigenen Erhebungen oder dem Verwaltungsvollzug („prozeßproduzierte Daten") als Grundlage. Entscheidungen über Maßnahmen, die sich — um nur einige Beispiele zu nennen — auf Altersversorgung, Arbeitslosigkeit, Armut, Krankheiten usw. beziehen, können verantwortlich nur auf der Grundlage der Kenntnis der entsprechenden kollektiven Informationen getroffen werden. Wie aber kann dieser enorm gewachsene Wissensbedarf über Kollektive mit dem Informationsschutz für einzelne in Einklang gebracht werden?

Handelt es sich dabei nicht gleichsam um die Quadratur des Kreises? Dies könnte so scheinen. Das Bemühen um diese „Quadratur des Kreises" war der Gegenstand dieses Symposions. Die Gestaltung des Symposions war darauf bedacht, den Schwierigkeiten des Gegenstandes gerecht zu werden. Die Erfahrungen des Auslands und der Fachleute verschiedener Fächer wurden eingebracht; auch der neue Forschungszweig der Deanonymisierungsforschung wurde bedacht und ist in diesem Band mit den Diskussionen zwischen den Experten der verschiedenen Fachgebiete dokumentiert. Nur in solch fachübergreifender Diskussion, die weder das Detail der Aufgabe noch die Herausforderung der je anderen Disziplin scheut, können die notwendigen Lösungen gefunden werden.

Prof. Dr. Duane F. Alwin
University of Michigan, Ann Arbor, Institute of Social Survey

Possibilities and Prospects for Anonymized Public Use Samples: National Data Resources in the Social Sciences*)

Introduction

Questions associated with the access of empirically-oriented social scientists to large-scale bodies of anonymized micro-level survey data are of considerable importance and substantial interest. Access to such micro-level data varies considerably from country to country.[1]) Access to micro-data in the U. S. is widespread and has been for some time, so much so that my U. S. colleagues in the social sciences frequently take such data resources for granted, without much knowledge of the relevant issues behind data dissemination and protection of confidentiality of individual data. Understandably, access to these types of data differs in the German context, and while circumstances are necessarily different, our common interest in the complex issues that lie behind access to such data motivate their serious consideration.By sharing experiences and discussing common concerns, unresolved research issues in the field of anonymization of data will be clarified and the dissemination of anonymized data bases will be facilitated.

In this paper I present a discussion of several issues related to the dissemination of anonymized microdata among social scientists in the U.S. I do this through reference primarily to three large-scale social science projects in the United States: The General Social Survey (GSS), developed by James A. Davis of Harvard University and conducted by Davis and Tom W. Smith of the National Opinion Research Center (NORC) based at the University of Chicago; the Panel Study of Income Dynamics (PSID), developed by James Morgan and his colleagues at the Institute for Social Research at the University of Michigan; and Michigan's American National Election Studies (ANES), developed by Angus Campbell, Philip Converse, Donald Stokes, Warren Miller and their colleagues. These are not official government surveys, but the issues surrounding their widespread use by the social science community are pertinent to the discussion of the dissemination of anonymized data bases.

The perspective I offer on these issues is one of a social science researcher, one who has made frequent use of public use data, and one who has been involved in a consultative role with the GSS project from its early years.[2]) This perspective is perhaps best informed on is-

*) The author benefited from conversations with Lawrence Cox, James Davis, Greg Duncan, Mark Abrahamson, David McMillan, Graham Kalton, and Daniel Kasprzyk in preparing this presentation.
[1]) See David H. Flaherty, Privacy and Government Data Banks: An International Perspective. London: Mansell, 1979. For an earlier discussion of data protection legislation in North America and Europe, see Ulrich Dammann, Otto Mallman and Spiros Simitis (Eds.) Data Protection Legislation: An International Documentation. Frankfurt am Main: Albert Metzner Verlag GmbH, 1977.
[2]) The author is currently Chair of the GSS Board of Overseers.

sues of the value of data resources of this kind and aspects of the demand for such data in the social and economic sciences. This perspective is somewhat less informed on the technical, legal and ethical issues of data protection and anonymization, but some such issues are briefly discussed in the following.

Definitions

The term "Public Use Sample" conveys more than one meaning. Perhaps the most popular usage involves reference to micro-level (personal or dwelling unit) data from governmental surveys, as well as samples from censuses, available to research scientists in the social and economic sciences through one or another set of mechanisms. Examples of such public use data from the U. S. are The Current Population Survey (CPS), The Survey of Income and Program Participation (SIPP), The Health Interview Survey (HIS) and the 1 in 100 and 1 in 1,000 public use samples from selected decennial censuses. The mechanisms for making such data accessible to social scientists vary from survey to survey, but provisions exist for review and dissemination in most federal surveys. One such mechanism is committee review at the agency level. Typically, decisions regarding access to government survey data files is determined at the agency level in the sense that guidelines and review procedures rest with the agency-level interpretation of government statutes regarding data protection.[3]

A second usage of the term "Public Use Sample" refers to any large-scale sample survey data set in the public domain, accessible to social scientists through data archives. Here, although the collection and management of such survey data sets may be funded by the government, the data are obtained for non-governmental purposes and are principally collected in service of research inquiry rather than governmental need for information. Such samples are typically national in scope and often are massive in coverage of micro-level information. Three examples from the U.S. are The Panel Study of Income Dynamics (PSID) and The American National Election Study (ANES), conducted by the University of Michigan's Institute for Social Research, and The General Social Survey (GSS) conducted by the NORC.[4] All of these data sets are national in scope and cover many years. All are supported by the National Science Foundation (NSF).[5] The mechanisms for making such data accessible to social scientists are less well-defined. Provisions existent in federal guidelines typically concern guarantees of confidentiality by the organizations collecting data, and funding

[3] The U. S. Census Bureau's release of micro-level data is governed by congressional legislation, "Title 13, United States Code Census", December 31, 1976. For an interesting set of papers dealing with current issues of the confidentiality of data in Federal Surveys, see L. Cox, B. Johnson, S. McDonald, D. Nelson, and V. Vasquez, "Confidentiality Issues at the Census Bureau", Pp. 199–218 in Proceedings of the First Annual Research Conference, Bureau of the Census. T. Plewes, "Confidentiality: Principles and Practice". Pp. 219–226, Ibid W. Griffith, "Discussion: Confidentiality Issues in Federal Statistics". Pp. 227–231, Ibid.

[4] There are other examples: The monthly national surveys of consumer sentiment carried out at The University of Michigan; the annual national surveys of U. S. high school students also conducted at The University of Michigan; the national longitudinal studies of high school graduates carried out by NORC; and the national longitudinal studies of labor force participation, conducted by Herbert Parnes of Ohio State University. These studies, however, have received considerably less widespread use among researchers in the social sciences.

[5] In fact, the NSF has given the three surveys referred to here a special status of national data resources, which credits not only their importance to the social science community, but which also acknowledges their special administrative status.

agencies (e. g. NSF and HHS) typically rely on the local institutional review boards dealing with the protection of humans subjects to carry out the provisions of the relevant regulations. Inevitably, an important source of control over the protection of the anonymity of data relies on standards of professional ethics of social science disciplines.[6])

These two categories of Public Use Samples differ in their subject matter, their auspices, in the nature of their management and control, and in the mechanisms for implementing assurances of confidentiality. There are nonetheless many common issues related to their use and dissemination. In what follows I first describe three major data resources in the social sciences: the PSID, the ANES, and the GSS and describe their major foci. Then I discuss the arguments in favor of the availability of such data resources, including those surveys collected without research purposes necessarily in mind. And finally, my discussion raises several issues linked to data protection and confidentiality, which must be satisfied in order to facilitate access to micro-level data, whatever their source. My conclusion will be that it is possible to balance the public's right to privacy and the researcher's need for data. In short, ways of doing research are possible, while protecting the rights and welfare of individuals. Let us now examine three prominent public use data bases in the U. S. and look at their record of disbursing data and at the same time protecting the confidentiality of respondent data.

The General Social Survey

Since 1972 the National Science Foundation (NSF) has supported the National Data Program for the Social Sciences, a NORC-based program for social indicator research and data diffusion. The major focus of this program has been the GSS, a (nearly) annual national survey of roughly 1 500 respondents. A major goal of this project has been to provide the social science community with large-scale, substantively important data of high quality. The GSS has been conducted in 12 years between 1972 and 1985.[7]) Current funding provides for surveys in 1987 and 1988. Continuation of the GSS beyond 1988 will be reviewed by the NSF for a new 5-year cycle of funding in late 1986. The German equivalent of the GSS is the ALLBUS which has existed since 1982.

The GSS data include a wide range of variables, touching on many areas of current interest to social scientists. Over 700 different variables are in the GSS cumulative data file. These include standard socio-economic and demographic variables, and a range of attitudes, self-reports of behavior and personal evaluations. Attitude items cover several broad topic areas, such as attitudes to abortion, crime and punishment, sex roles, foreign affairs, institutional leadership, national spending priorities, race relations, religious beliefs, taxation and income redistribution, tolerance, violence, the work ethic and child-rearing orientations. Behavioral reports include measures of drinking and smoking, gun ownership, organizational memberships, political affiliation and voting, social interaction patterns, major life events and life stress. Personal evaluations include measures of alienation or anomie, general sub-

[6]) See "Revised ASA Code of Ethics", Footnotes, August 1980, p. 12. See also National Research Council "Protecting individual privacy in evaluation research", 1975.
[7]) The GSS has been run annually from 1972 to 1978, in 1980, and from 1982 to 1985.

jective well-being, and satisfaction with aspects of one's job, marriage, finances, family, friends, place of residence, health and leisure time.

One of the unique characteristics of the GSS is its repeated cross-section design, permitting the study of time-trends in its various attitude measures. The GSS replicates such items regularly, according to a previously specified rotation scheme. In addition, many of the GSS items were selected from earlier NORC, Gallup, and SRC surveys, allowing trend analysis beyond the scope of the GSS on many of the topics covered by the survey.

Methodological experimentation is another unique feature of the GSS. Through the use of randomized experiments (or split ballots), special scales, and test-retest measurement, the GSS data allow the assessment of some sources of errors in attitudinal measurement. For example, GSS surveys have been a source of data on the effects of question wording and context.[8]) The GSS Technical Report series documents many of the findings of this methodological work.

Within the past three years the GSS has changed its format to include special one-time modules focusing on important substantive issues in the social sciences. Each of these modules is designed to use about one-fourth of the GSS interview. The first of these topical supplements was part of the 1985 survey and focused on the topic of social networks, the 1986 survey will contain a vignette study of the feminization of poverty, and the 1987 module will contain a replication of Verba and Nie's 1967 study of socio-political participation in American society. The topical module design of current GSS surveys is an enhancement of the basic design that permits innovation and the development of new measurement approaches, in addition to strict replication.

The Panel Study of Income Dynamics

In 1968, after the federal government had launched its "war on poverty", a group of economists at the University of Michigan's Institute for Social Research launched a longitudinal study designed to answer some fundamental questions about the nature of poverty and its persistence over the life course of individuals.[9]) Beginning with a sample of about 5000 households, the PSID project has obtained annual interviews with members of these original households. The study followed, not only members of the original households, but split-off households formed when adult children leave home or when couples divorced. By the 18th wave of data (obtained in 1985) the sample consists of roughly 6800 families, wherein one primary adult is interviewed each year.

---

[8]) See Howard Schuman and Stanley Presser, Questions and Answers in Attitude Surveys. New York: Academic Press, 1981.

[9]) This summary is based on the description of the PSID presented by Greg J. Duncan and James N. Morgan, "The Panel Study of Income Dynamics". Pp. 50–71 in Glen H. Elder, Jr. (ed.), Life Course Dynamics: Trajectories and Transitions, 1968–1980. Ithaca and London: Cornell University Press, 1985. See also James N. Morgan, Jonathan Dickinson, Katherine Dickinson, Jacob Benus, and Greg J. Duncan, Five Thousand American Families – Patterns of Economic Progress. Vol. 1. Ann Arbor, MI: Institute for Social Research, 1974. Page references in this section of the paper are to the Duncan and Morgan (1985) paper.

The information obtained by the PSID project involves both economic and non-economic variables, and since the study follows those individuals who leave the original households to create new households, the data set provides the capability of linking individual histories to their families of origin (p. 51). The PSID's focus on sources and amounts of income is a major feature of the data collection effort, which requires the collection of substantial information that helps the analyst understand the income data, such as household composition, employment and periods of unemployment, occupation, housing, expenditures, geographic mobility, disability, health and other background information (p. 55). The project has also obtained data on other topics on a one-shot basis, for example, on cognitive ability, achievement motivation, home production, and a variety of attitudes and behavior patterns linked to economic success (pp. 55-59). Important for demographic analysis, the PSID chronicles the histories of events in the lives of families, e. g. birth histories, periods of unemployment, residential mobility, job and employment changes, changes in marital status and household composition. There are about 600 variables added to the family data record each year.[10])

The principal value of the PSID is its panel design, permitting the identification of gross change at the individual level and the analysis of persistence and change in socio-economic experiences of individual households. In addition, except for in- and out-migration from the U. S., the PSID sample is representative of U. S. families each year (since split-off families are interviewed), and study staff indicate that "the cross-sectional analysis of the most recent interviewing wave is nearly equivalent to an analysis with a fresh cross-sectional data set such as the Census Bureau's Current Population Survey" (p. 60). Finally, according to Elder, the PSID is uniquely suited to the analysis of life course dynamics because it permits the investigation of trajectories over the life course and the experiences induced through ecological transitions.[11])

The American National Election Studies

Every two years since 1952 the Survey Research Center/Center for Political Studies at the University of Michigan's Institute for Social Research have interviewed a representative cross-section of Americans to track national political participation.[12]) On the years of presidential elections, a sample is interviewed before the fall election and is reinterviewed immediately afterward. Only post-election interviews are conducted in congressional election years.

Originally, these election-year studies were concerned primarily with the understanding of electoral behavior, but with time their scope broadened to include a wide range of socio-

[10]) Beginning in 1974, the PSID project published an annual volume of findings (through 1984) and in 1983 a User's Guide to the PSID was published, which summarizes the important aspects of the project. Data tapes for the PSID are available through the ICPSR data archive.

[11]) See Glen H. Elder, Jr., "Perspectives on the life course". Pp. 23–49 in Glen H. Elder, Jr. (ed.), Life Course Dynamics: Trajectories and Transitions, 1968–1980. Ithaca and London: Cornell University Press, 1985.

[12]) The basis for this discussion is Warren E. Miller, "Long-terms support for the American National Election Studies", A proposal to the National Science Foundation, January, 1984.

political behavior and attitudes, and these studies now represent a national data resource of considerable importance to the social and behavioral sciences. In addition to trends in political partisanship and factors linked to political participation and electoral behavior, these studies have permitted the investigation of a breadth of topics linked to basic issues in processes of attitude formation and change and the dynamics of change in aggregate public opinion.

The ANES actually combine both the benefits of a repeated cross-sectional study design and those connected to the panel design, for on three separate occasions the initial sample was reinterviewed later. In the 1956 study, a substantial subsample was reinterviewed in 1958 and 1960. In the 1972 study, a large subsample was reinterviewed in 1974 and 1976. And in the 1980 study, a subsample was reinterviewed three times during the election campaign.

As a data resource for the social science community, the ANES provides a distinct model for the design of research in that conceptual development in the ANES is based on a collaborative model. The community of active users shapes the direction of the data base and are heavily involved in the design of each successive election-year study. This model has become institutionalized, and beyond a basic core of repeating questions, instrumentation of current research needs in the field is shaped by a committee acting upon behalf of the wider research community.

The ANES project has also been a leader in methodological innovation. The project, as noted, has incorporated more than one type of research design. In the 1984 survey, for example, the ANES included a design consisting of weekly interviews with small national population samples throughout the preelection period. The ANES has also experimented with the use of innovative data collection techniques, including the 1982 test of the viability of computer-assisted telephone interviewing for large and complex data gathering efforts.

The Case for Public Use Samples in the Social Sciences

Large-scale public use samples are indispensable for an empirically oriented social science. The availability of such sample data is a critical aspect of the development of an empirically-based set of social science disciplines contributing to knowledge of human behavior and social policy effects. There are three essential functions such large-scale data bases serve for social scientists. One is monitoring and quantifying social trends and their causes; the second is that of testing leading hypotheses regarding social processes and human behavior; and the third is the potential for international collaboration and cross-national comparisons.[13]) Let me discuss each of these briefly, providing some examples from the research literature where it is appropriate to do so.

[13]) There are other important functions served by public use data bases, including methodological developments and the stimulation of new areas for research, but I restrict my present comments to those listed here.

Monitoring trends and their causes

The existence of social change provides a compelling basis for the interest of social scientists in monitoring social trends in both behavior and attitudes. In Toward Social Reporting: Next Steps (1969), Otis Dudley Duncan states that the "improved capability and capacity to measure social change is fundamental to progress in the field of social reporting." In the past two decades there has been an emergence of great interest in documenting social change through the use of replicated survey designs.[14]) This represents a clear benefit of large-scale public use data sets, such as those discussed here. In some instances such studies replicate the measures taken in earlier one-time surveys. In others, such as the GSS, measures are repeated annually or nearly so.

The GSS is one current exemplar of the repeated cross-section design in the social sciences. Although the time period covered by the GSS is short – twelve surveys since 1972 – it is possible to place it in the context of earlier Gallup and SRC surveys using the same items. These data permit the analysis of patterns of variation over time within age cohorts. A review of the major themes learned from research with the GSS and other attitude surveys include the following:

(1) a gradual increase in social liberalism on most issues over the past 25 years;
(2) net aggregate attitude change has affected all parts of society equally, with conversion as important as replacement;
(3) intergenerational transmittances shape attitudes and behaviors as well as socioeconomic status;
(4) membership in subcultures has a significant effect on a range of attitudes;
(5) occupational mobility has little effect on a wide variety of attitude measures;
(6) education is one of the most persistent influences on sociopolitical attitudes;
(7) differences between some religio-ethnic groups are narrowing.[15])

[14]) See Eleanor Bernet Sheldon and Wilbert E. Moore (eds.), Indicators of Social Change: Concepts and Measurements. New York: Russell Sage Foundation, 1968. Angus Campbell and Philip Converse, The Human Meaning of Social Change. New York: Russell Sage Foundation, 1972. Otis Dudley Duncan, Howard Schuman and Beverly Duncan, Social Change in a Metropolitan Community. New York: Russell Sage Foundation, 1973. Kenneth C. Land and Seymour Spilerman (eds.), Social Indicator Models. New York: Russell Sage Foundation, 1975. Elizabeth Martin, "Surveys as Social Indicators: Problems in Monitoring Trends". Pp. 677–743 in Peter H. Rossi, James D. Wright and Andy B. Anderson, Handbook of Survey Research. New York: Academic Press, 1983.

[15]) It is possible to summarize the findings regarding social trends only briefly. Key references are: J. Davis (1978), "Trends in NORC General Social Survey Items 1972–1977", GSS Technical Report No. 9. J. Davis (1980) "Conservative weather in a liberalizing climate: change in selected NORC General Social Survey Items, 1972–1978", Social Forces 58: 1129–1156. J. Davis (1982) "Achievement variables and class cultures: family, schooling, job and fourty-nine dependent variables in the cumulative GSS", American Sociological Review 47: 569–586. J. Davis (1975) "Communism, conformity, cohorts, and categories: American tolerance in 1954 and 1972–73", American Journal of Sociology 81: 491–513. J. Davis "New money, an old man/lady and 'two's company: 'subjective welfare in the NORC General Social Surveys, 1972–1982", Unpublished GSS Report. J. Davis "Up and down opportunity's ladder", Public Opinion 5: 11–15, 48–51. J. Davis and T. W. Smith (1982), "Have we learned anything from the General Social Survey?" Social Indicators Newsletter 17: 1–2, 8–11. J. Davis (1976), "Background characteristics in the U. S. adult population 1952–1973: a survey-metric model". Social Science Research 5: 349–83. J. Davis (1983) "Counting your change for a ten: America from 1972 to 1982 as reflected in the NORC General Social Survey", GSS Technical Report No. 43.

Testing leading hypotheses regarding social processes

There is little question that one of the major benefits of large Public Use Samples on relevant social and economic variables is the ability to examine relationships among many variables at once. The national scope of these surveys and their large sample sizes permits the examination of multivariate models of social processes. Perhaps the most famous American example of this is the now-classic study of intergenerational and intragenerational occupational mobility by Peter Blau and Otis Dudley Duncan, called Occupational Changes in a Generation, based on a March, 1962 Current Population Survey (CPS).[16]) The German counterpart to this study was conducted by Karl Ulrich Mayer and Walter Müller based on a 1970 micro-census, and the American study was replicated by Robert Hauser and David Featherman in the 1973 CPS.

One of the most important substantive findings of the PSID project, which is possible only because of its panel design, is its questioning of popular conceptions of "the poor" as a relatively homogeneous, stable group of families whose children inherit their life circumstances.[17]) This is particularly interesting given the common "culture of poverty" theories of intra- and inter-generational economic stability. These PSID revelations regarding the relative lack of stability in family income are accompanied by pertinent analysis of the factors associated with change, such as changing family composition and labor force activity/participation.

The ANES have been the mainstay for the analysis of electoral behavior for most of the field of political science since its inception.[18]) As noted, the data base has also been important in the analysis of issues tied to public opinion and attitude change. One of the central debates within American political science during recent years involves the solidity of Americans political attitudes. Using ANES data from the 1960s and 1970s, Philip Converse contributed a widely influential argument that political attitudes vary widely in terms of the degree to which they represent stable and potent action-directing dispositions.[19]) These data revealed that some types of attitudes, such as attitudes toward political parties and political actors, are reflected in stable and affectively strong orientations, while others, notably specific government policies, were remote from persons' everyday lives and relatively unstable for most persons, not reflecting therefore very strong ideological dispositions. Converse's and Knoke and Hout's systematic analyses of cohort vs. period effects on party identification represent another important contribution made possible by the ANES.[20])

---

[16]) See Peter Blau and Otis Dudley Duncan, The American Occupational Structure. New York: Wiley, 1967. The GSS has contributed further confirming evidence of these American studies of mobility.

[17]) See Greg J. Duncan, Years of Poverty, Years of Plenty. Ann Arbor, MI: Institute for Social Research, 1984; Glen H. Elder, Jr., Life Course Dynamics: Trajectories and Transitions, 1968–1980. Ithaca: Cornell University Press, 1985.

[18]) See e. g. Angus Campbell, Philip E. Converse, Warren E. Miller and Donald E. Stokes, The American Voter, New York: Wiley, 1960.

[19]) See Philip E. Converse, "The nature of belief systems in mass publics". In D. Apter (ed.), Ideology and Discontent. New York: Free Press, 1964; and "Attitudes and non-attitudes: continuation of a dialog". In E. R. Tufte (ed.), The Quantitative Analysis of Social Problems. Reading, MA: Addison-Wesley, 1970.

[20]) Philip E. Converse, The Dynamics of Party Support: Cohort Analyzing Party Identification. Beverly Hills, CA: SAGE,1976. David Knoke and Michael Hout, "Social and demographic factors in American political party affiliations, 1952–1972." American Sociological Review, 1974, 39: 700–13.

The Potential for Cross-National Comparisons

The value of comparative research is well-established and well-documented. The knowledge to be gained from cross-national comparisons represents another benefit of large-scale public use samples. The success of such work depends, however, on the willingness of collaborators to establish close equivalences of design across countries.[21]) This has happened with the GSS. The first international collaboration with the GSS was with the Zentrum für Umfragen, Methoden und Analysen (ZUMA) in the 1982 German National Social Survey (ALLBUS), and since then social scientists in a number of countries have embarked on the GSS model, the British Social Attitude Study in England, Wales and Scotland, the annual surveys conducted by the Australian National University, and others. The existence of comparable cross-national public use data sets offers an opportunity for the exchange of findings and replication of the same hypotheses/theories in varying social contexts, permitting greater understanding of system-level differences.

In order to facilitate such international comparisons, the GSS data are available to the international social science community through the data archives at the Roper Center and ICPSR at Michigan in the United States, the Social Science Research Council Survey Archive, University of Essex in England, the Zentralarchiv für Empirische Sozialforschung, University of Cologne in Germany. In addition the NORC staff are willing to provide data and codebooks to foreign scholars on an individual basis. Both the PSID and the ANES have international components, and these data are available through international data archives.

Assessing Demand for Public Use Samples

Is there a demand for public use data bases? One of the critical pieces of my argument is that there is an existing demand for public use samples, of both types, within the social science research community. Obviously, the answer one provides to such a question depends on social and historical circumstances, and any general answer is thereby limited to particular situations sharing experiences in common. I am equipped only to assess the demand for such data in my own country (and am even somewhat limited in doing this), although there are certainly commonalities in our experiences.

For the past two decades statistical agencies in the U. S. have been under increasing pressure for access to microdata for research and statistical purposes.[22]) Economists, sociologists and demographers have for several years depended upon public use samples prepared by the Bureau of Census for research purposes. Research scholars and their students make regular use of Census microdata. In 1979 Flaherty (p. 293) estimated that over two hundred research scholars were using public use data from the Census Bureau, and he cites one informed observer who suggested that "at least one thousand graduate and undergraduate students were using the public use samples from the census, and that at least ten

[21]) See E. Martin (see footnote 16) for a discussion of procedural versus functional equivalence.
[22]) This discussion relies heavily on Flaherty (1979: 292–300) cited above.

thousand more students were aware of their existence". He estimated that perhaps one-hundred sociological articles had been prepared on the basis of analyses of public use samples from the Census. This was nearly a decade ago, and it was projected then that the demand for data was expected to increase. More recent information is most certainly available from the Census Bureau. Similar pressures have been experienced by the National Center for Health Statistics and the Social Security Administration, although the Census Bureau has pioneered in responding to the demand for data dissemination and it provides one of the more powerful illustrations of the benefits derived from access to microdata.

The demand for large-scale non-governmental public use data is somewhat more difficult to assess because such data are widely accessible through data archives and I am unaware of attempts to measure data usage of this type. Thus, information in this area is limited in general. In the specific case of the GSS there is unusually good documentation of publications using the data base. The GSS staff reports, on the basis of a list compiled from a canvass of major users and of social science journals, and a computer-assisted search of social science abstract and citation collections, that the use of the GSS for purposes of research and publication is substantial.[23])

Comparable use figures exist for ANES. Election study estimates for total primary and substantial election data use include some 82 books, 580 journal articles or book chapters, 156 conference papers, 134 doctoral theses, and an additional 136 publications making occasional use of the data. Use counts of this sort clearly under-estimate the demand for such data resources, since considerable use, such as that involving training, is not reflected in these indicators.

There are other indications of widespread use of these data resources, and I conclude from these sketchy indicators, as well as more impressionistic information, that within modern social science there is considerable demand for public use data. I turn now to issues of privacy, confidentiality and data protection under the regime of widespread dissemination of public use data.

Privacy, Confidentiality and Data Protection

The social historian Barrington Moore says that "privacy cannot be the dominant value in any society". Man has to live in society, and social concerns have to take precedence.[24]) His analysis of ancient Greece and ancient China suggest that the words "private" and "public"

[23]) See T. W. Smith and M. Ward, Annotated Bibliography of Papers Using the General Social Surveys, 5th edition, NORC (May, 1984). The authors list over 1,000 known uses since 1972 as registered in publications and completed research papers. They note that "the list of located papers is believed to fall far short of the total uses of the General Social Surveys and especially tends to overlook books, presented papers and very recent works".

[24]) Privacy: Studies in Social and Cultural History. London: M. E. Sharpe, Inc., 1984, pp. 274–75. For additional discussions of privacy issues, see David M. OBrien, Privacy, Law and Public Policy. New York: Praeger, 1979. Ferdinand D. Schoeman (ed.), Philosophical Dimensions of Privacy: An Anthology. Cambridge: Cambridge University Press, 1984.

existed, with the former conveying some hint of the "antisocial". And while these and other ancient societies demonstrate the priority of social concerns, not all such concerns always took precedence. Indeed, Barrington Mooreś conclusion is that a major accomplishment of Western civilization is the role of the concept of privacy in the questioning of social concerns, that is, the "sense of rights protecting the individual against public authority".

Riecken and Boruch draw a distinction between the issues of privacy and confidentiality.[25]) Issues of privacy occur during the process of observation or data gathering due to the intrusion of the observer (in the case of survey data, the interviewer) on the physical isolation of the individual or his/her private activities, attitudes and subjective evaluations. Especially when the data gathering process requests that the individual divulge information they do not care to share with others is the question of privacy an issue. Normal survey procedures respect the right of the individual to not respond to given questions, if s/he so desires. The issue of confidentiality arises logically subsequent to the collection of data when requests are made to divulge information gained in the survey in a way that may potentially identify the individual and link his/her identity to such information.

There is, then, obviously a tension between the privacy rights of individuals and what might be considered the public good, that is, governments need for information and the availability of this information for research investigations. Discussions of these issues from the perspective of the rights of citizens to privacy raise concerns regarding the motives of government and the responsibilities connected with assurances of confidentiality. There should be little question that these privacy issues are extremely important and the principle of confidentiality should not be compromised in the dissemination of data.

Discussions of these issues from the perspective of social scientists with a need for public use data raise several issues regarding confidentiality of information. First, what are the risks of disclosure in public use samples and what types of information are critical to producing such risks? Second, what are the effective review mechanisms for assuring that guarantees of the privacy of survey records are implemented? Third, what are the state-of-the-art techniques for reducing or removing the risk of disclosure of identifiable private information?

Risks of Disclosure

The concern with assuring the confidentiality of individual information contained in public use datasets is reflected in policies governing research involving human beings. For example, the focus of the privacy legislation has been:

(1) to distinguish between the research and administrative use of records,

(2) to specify carefully the circumstances under which records can flow from administrative to research use, and

(3) to prohibit the reverse flow from research to administrative use.[26])

[25]) Henry W. Riecken and Robert F. Boruch (1974), Social Experimentation: A Method for Planning and Evaluating Social Intervention. New York: Academic Press (pp. 255–69). See also Robert F. Boruch, Assuring the Confidentiality of Social Research Data, 1979.

[26]) Bradford Gray, "Developing issues in the ethics of social research on health", Health Survey Research Methods, 2nd biennial conference, Williamsburg, Virginia. National Center for Health Statistics Research. U. S. Dept. of Health, Education and Welfare, 1977. DHEW Publication No. (PHS) 79–3207.

Further, policies aimed at the protection of human subjects in research, although developed primarily for the protection of individuals in bio-medical research, require that surveys supported by relevant government agencies, such as the NSF, protect the confidentiality of their respondents.[27])

Public Use Samples released for research use in the U. S. are universally anonymized in the sense that individual identifiers are not disbursed in machine-readable form. Identifying information in the form of names, social security numbers and street addresses is removed from the records appearing in public use data sets. Thus, under these procedures there is no risk of identification of individual persons by name and address by researchers obtaining public data sets. The records linking individuals to the data in non-governmental research are typically either destroyed, or stored separately under the supervision of the organization collecting the data. What, then, are the areas of concern involving risks to the disclosure of data pertaining to individuals from the release of public use microdata?

One potential concern in non-governmental surveys is the protection of the confidentiality of research data against the legal process. The possibility that criminal courts may subpoena research records on individuals is a threat to the confidentiality of respondents which researchers may not be able to legally protect. Riecken and Boruch (1974) note that while researchers have a moral responsibility and professional obligation to honor promises of confidentiality, there may be circumstances in which the researcher may find it difficult, or impossible, to do so. Although many of these circumstances are only hypothetical, there have been several cases in the U. S. in which legislative committees and judicial authorities have attempted to secure information obtained through the research process.[28]) Riecken and Boruch (1974, p. 256) cite one case in which government authorities have been successful in obtaining information from researchers under threat of imprisonment. Although these are rare instances, they do remind us of the potential for breaches of confidentiality. These potential risks have encouraged researchers to remove all identifying information, or to store record linkages outside of the country, and to push for legislative relief.

A second concern in the area of confidentiality is that researchers or members of their staffs using anonymized data will obtain the identity of individuals through some inductive use of geographical and other information in the file. In cases where the identity of specific respondents could be inferred from unique constellations of responses in the statistical record, great care and supervision must be exercised. Although I know of the existence of no such instances of this in the U. S., the potential for it to be realized causes the purveyors of Public Use Samples to be extremely cautious in reviewing the coded variables in the file. The GSS, for example, codes only geographic region of the U. S. (9 categories) in their cumulative file and does not provide a state code. Unless one knew more about the sampling frame for a given survey it would be virtually impossible for a given respondent to be identified without

[27]) See the Federal Register, "Final Regulations Amending Basic HHS Policy for the Protection of Human Research Subjects", vol. 46, no. 16, pp. 8366–8392.

[28]) Bradford Gray (1977) cites one study which documented at least 17 subpoenas issued for research records and 26 other instances of governmental demands for information.

more detailed geographical information. Other surveys do provide state of residence, e. g. the ANES, and in some instances more detailed geographic information, e. g. the PSID, which increases the likelihood that a given respondent's identity could be revealed.[29] These are very unlikely possibilities, but those agencies and organizations disbursing public use data must exercise care in reviewing their likelihood.

Even though extreme efforts are often taken to protect the confidentiality of data from individuals in public use samples, the public's perception is often different. A 1979 national survey in the U. S. sponsored by the National Academy of Sciences provides some rather startling conclusions.[30] They asked the following question: "Individual survey records identified by names and addresses are kept in files of the United States Bureau of the Census. These records contain information on such things as occupation, income, race and age. Do you happen to know whether these records are public so that anyone who might want to see them can, or are they not open to the public?"[31] Only 35 % of respondents correctly reported that such files were not public; 18 % said they were open; and 46 % did not know whether or not they were. Probing further regarding the access of other government agencies to the Census records, the investigators found that only 5 % of all respondents really believed that the Census records were completely confidential. 80 % said they believed either that Census records were not confidential or that other government agencies could obtain information from the Census Bureau if they *really tried*.

Clearly, the public perception of the anonymity of Census records is apparently quite disparate from legal requirements and actual practice. In addition to pointing to the need for greater education of the public regarding the statutes regulating guarantees of confidentiality and agency practices, the public's misperception of this situation forms an even stronger rationale for the exercise of great care in protecting the confidentiality of individual data.

Removing the Risks

In order to insure that disclosure of confidential information does not happen, substantial responsibility for review and surveillance lies with the agency or organization disbursing the data. The mechanism of agency review panels, which monitor the type of information released is in practice an important safeguard. In such instances decisions of what information to release are decentralized in that guidelines and review procedures rest with the agency-level interpretation of government statutes regarding protection of confidentiality. In the case of non-governmental public use data obtained through government support, there are relevant regulations pertaining to guarantees of confidentiality. But in these cases the funding agencies typically rely on the local institutional review boards dealing with the protec-

[29] Although the PSID data are released in anonymized form, county and state of residence are coded, and "environmental information reported by respondents is supplemented with county-level data about unemployment levels, unskilled wage rates and labor market demand conditions obtained from the employment security officials of the state" (p. 55).

[30] See Privacy and Confidentiality as Factors in Survey Response. National Academy of Sciences, Washington D. C., 1979.

[31] It is, of course, a crime to disclose such records to anyone outside the Census Bureau; they are not available to other federal agencies.

tion of human subjects. As mentioned early in the paper, an important source of control over the protection of the confidentiality of data relies on standards of professional ethics.

In addition to the review mechanisms referred to here, it is also important to briefly consider the alternatives which such review panels have in terms of reducing or removing potential risks to data disclosure. There are several known mechanical techniques for further anonymizing individual data, beyond that provided by removing individual identifying information.[32]) One solution is simply the deletion of any potentially identifying information. If individuals are likely to be identified from knowledge of the Census tract in which they reside in combination with other variables, then the tract information may conceivably be removed from the statistical record. Another solution is to recode the offending variable(s) into cruder categories, so that the public variables have less detail than the raw data files. A third solution involves the introduction of a known amount of random perturbation into the data, so that while individual records are erroneous, statistical estimation can correct for error levels. And finally, data may be released in "micro-aggregates", wherein data are averaged over small sets and data are released only for these "synthetic" individuals rather than for the unique individuals that comprise the micro-aggregate.

In addition to these mechanical solutions to the problem of protecting the confidentiality of data, Riecken and Boruch (pp. 259–61) argue that other solutions, which are legal in nature, may ultimately be necessary, especially for the protection of data in non-governmental surveys. One such solution is for research investigators to have a type of "testimonial privilege", of the sort enjoyed by priests and lawyers. This would prevent violations of privacy resulting from judicial or congressional action. A second legal solution is to legislatively impose severe legal penalties for researchers or their staff members who disclose the identity of respondents. Finally, it is clear that research investigators and their staffs require continual education regarding ethical standards of confidentiality and the legal and governmental regulations governing the use of data and the protection of confidentiality.

[32]) This discussion relies heavily on Riecken and Boruch (1974).

Dr. Lawrence H. Cox
Bureau of the Census, Washington, D. C.

The Practice of the Bureau of the Census with the Disclosure of Anonymized Microdata

The United States Statistical System

The statistical system of the United States is decentralized. The responsibility to design statistical surveys, collect data, and release statistical summaries or microdata is shared among several Federal agencies, including the Bureau of Labor Statistics (BLS), the Energy Information Administration, the Social Security Administration, the Internal Revenue Service, the National Center for Health Statistics (NCHS), the National Center for Education Statistics, the Statistical Reporting Service (Department of Agriculture), the Bureau of Justice Statistics, and the Bureau of the Census. The Bureau of the Census is the Nation's principal statistical agency, in that it is the largest, it frequently collects data for other agencies, and it enjoys nonreciprocal access to certain data held by them.

In a decentralized statistical system, the concept of microdata dissemination must be carefully defined. Microdata may be shared between statistical agencies for purposes including reduction in respondent burden, improving the quality of hard-to-collect data, and validation of collected information from independent sources. In such cases, it may be necessary to share identifiable data or, at least, data which are sufficiently rich in detail as not to qualify as being anonymized.

Data sharing is a major concern to the U. S. statistical system, but is beyond the scope of this paper. For purposes here suffice it to say that a one-way street model of data sharing is employed in the United States, whereby the Bureau of the Census has acces to certain microdata held by other agencies, but the reverse is never true — identifiable Census data are tightly protected by law from release to outsiders (even the courts). When we speak here of releasing Bureau of the Census microdata, we refer either direct release of a microdata file as a public use file or selective release to a third party. Typically, the third party is the sponsor of the survey from which the microdata file is derived, and often the sponsor intends to disseminate these data further. With few exceptions as provided by law, released microdata files are always anonymized.

Microdata Released by Other (Non-Census) United States Statistical Agencies

In this section we briefly describe the principal non-Census released microdata files and the means employed to reduce their disclosure risk. The material contained herein is derived from U. S. Department of Commerce (1978) and informal discussions with representatives of these agencies. Mugge (1984) provided a detailed description of the situation for the NCHS.

The Social Security Administration has developed the Continuous Work History Sample covering all persons holding a social security number. From this file a 1 % sample, called the Longitudinal Employee-Employer Data (LEED) file, is constructed for public use. Variables contained on the LEED file include age, race, sex and earnings data for individuals, as well as data on individuals' employers going back to 1957. From its first release in 1962, the LEED file was released publicly under a restricted use agreement specifying the purposes of use and prohibiting unauthorized sharing of the file or use of the file in matching studies aimed at identifying individuals. Because the LEED file contained earnings data on individuals obtained from the Internal Revenue Service, public release of the file was suspended as a consequence of the passage of the Tax Reform Act of 1976, wherein strict data release criteria for data obtained from tax returns were set.

The Social Security Administration also releases microdata files, such as from the Retirement History Survey, for public use without restrictions on their use. These files represent small samples (less than 1 %), contain only limited geographic information, and are subject to suppression of unusual variables or combinations of variables prior to release.

The Internal Revenue Service releases for public use the Tax Revenue Model for National Estimates without restrictions on its use. This file represents less than a 0.2 % sample of all tax returns, although the sampling fraction is higher for certain subgroups. The Tax Model for State Estimates, including approximately 0.3 % of all returns identified at the State level, is available to State tax agencies. Both files are samples of individual tax returns, stripped of individual identifiers, containing 150 data items from each return.

The National Center for Education Statistics releases a public use microdata file representing a 0.7 % sample of the National high school graduating class of 1972 without restrictions. Stripped of identifying information and containing only region and urban/rural designations geographically, this file contains information such as grade point average, class rank, national test results and area of study, as well as student-provided data on family background, attitudes and plans for the future. Follow-up information provided by respondents is added to the file periodically.

The NCHS releases public use microdata files from many of its surveys and statistical programs, including the Health Interview Survey, the Health and Nutrition Examination Surveys, the National Ambulatory Medical Care Survey and the Hospital Discharge Survey. One file is particularly unique — the file on natality — which contains a 50 % (in some States, 100 %) sample of birth records. Users of all files are required to sign a statement that the file will be used only for statistical research or reporting purposes. NCHS strips all directly identifying information, limits geographic detail to areas with population at least 100 000 persons, and deletes or collapses categories which present rare characteristics. NCHS is confident that these disclosure-avoidance measures are adequate because of the lack of suitable matching files available outside the agency.

The BLS releases a small number of special-order microdata files for limited distribution, sometimes under restricted use. Although BLS sponsors and technically is the releasing agent for public use microdata files from the Consumer Expenditures Survey, these data are

collected, processed and subjected to disclosure-avoidance by the Bureau of the Census. BLS has no continuing public-use microdata release program of its own. This is because BLS obtains much of its data from State agencies under strict restricted use agreements.

Microdata Released by the Bureau of the Census

The Bureau of the Census operates within a legal framework favorable to its role as the Nation's principal data collector but correspondingly strict in terms of its responsibility to preserve respondent confidentiality. It is worthwhile to present the relevant portions of that legislation here.

From Title 13, U. S. Code (as amended December 31, 1976):

Sec. 8 (b) . . . The Secretary may furnish copies of tabulations and other statistical materials which do not disclose the information reported by, or on behalf of, any particular respondent, and may make special statistical compilations and surveys, for departments, agencies, and establishments of the Federal Government, the government of the District of Columbia, the government of any possession or area (including political subdivisions thereof) referred to in section 191 (a) of this title, State or local agencies, or other public and private persons and agencies . . .

. . . (c) In no case shall information furnished under this section be used to the detriment of any respondent or other person to whom such information relates, except in the prosecution of alleged violations of this title.

Sec. 9. Information as confidential; exception

(a) Neither the Secretary, nor any other officer or employee of the Department of Commerce or bureau or agency thereof, may, except as provided in section 8 of this title —

(1) use the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied; or

(2) make any publication whereby the data furnished by any particular establishment or individual under this title can be identified; or

(3) permit anyone other than the sworn officers and employees of the Department or bureau or agency thereof to examine the individual reports.

. . .

(b) The provisions of subsection (a) of this section relating to the confidential treatment of data for particular individuals and establishments, shall not apply to the censuses of government . . .

The Bureau of the Census therefore must assure that each released microdata file serves useful statistical purposes not to the detriment of respondents and that the file cannot be used to identify data pertaining to individual respondents. Whereas the language of the law has remained fixed, its operative terms, "detriment" and "identify", are not precise statistically and must be reevaluated for each new data release in terms of factors such as new reidentification methods, the range of analyses which could be performed on the data, and the existence of alternative data sources which could be used in reidentification. We summarize the situation at the Bureau of the Census for release and disclosure-avoidance in microdata below. Cox, et. al (1985) provided an overview of the full set of confidentiality issues affecting the Bureau of the Census.

Beginning with the 1960 Decennial Censuses, the Bureau of the Census has released public use microdata files (so-called "Public Use Samples") from its decennial censuses of population and housing. Subject to the requirements of disclosure-avoidance and other statis-

tical standards regarding data release, these files contain the entire set of characteristics available from the decennial censuses: age, race, education, occupation, income, etc. for each person, plus housing characteristics for the household. The original sample released in 1963 contained little geographic information (census region designation and type of area code) and represented only a 0.1 % sample of all U. S. households derived from the 1-in-4 sample of all U. S. households receiving the census long form. A 0.01 % subsample was also publicly released. Purchasers of these files signed an agreement prohibiting unauthorized sharing of the file and requiring that the Bureau of the Census receive a copy of any publication containing data obtained from the file. It ist clear that these agreements were not upheld: by 1969, having sold 65 copies of the file, the Bureau of the Census was able to identify the existence of copies at over 200 institutions.

The 1970 census Public Use Samples represented a 1 % sample of all households derived from the 1-in-5 census long form sample and provided geographic information down to areas containing at least 250 000 persons (which was then the population threshold for Census microdata release). Six mutually exclusive 1 % Public Use Samples were released from the 1970 censuses, representing samples derived from two versions of the questionnaire and providing data by three different geographic schemes (metropolitan/nonmetropolitan, urban/rural and county groups). Their use was not restricted. These files were extremely popular with users with thousands of copies having been sold directly by the Bureau of the Census.

The 1980 Public Use Samples were also based on a 1 % sample of all U. S. households from the overall 1-in-5 sample (1-in-6 in most areas) of census long form questionnaires. As in 1970, three disjoint public use files were released, each according to a different geographic schema.

The Bureau of the Census releases annually for public-use the Annual Demographic File (ADF) derived from the March supplement to the Current Population Survey (CPS). Similar in content and structure to the Public Use Samples, these files provide a time series of data back to 1968. In the ADF, the population threshold must be monitored carefully as the CPS sampling design can overlap standard geographic classifications: data cannot be shown which would identify any area below the population threshold. So, for example, balanced half-sample codes and geographic designators such as urban/rural or central city/metropolitan may be suppressed in favor of State codes to ensure that the threshold is not crossed.

Microdata files are released from almost all other Census Bureau household surveys. These files usually contain the entire sample. In rough figures, these surveys have an average sampling rate of 1-in-1500. No restrictions on use are placed on Census Bureau census or survey public use microdata files.

A new, rich survey was begun at the Bureau of the Census in 1983, the Survey of Income and Program Participation (SIPP). This is a longitudinal survey of persons, families and households which collects detailed income, investments and other financial data. Public use microdata files containing cross-sectional data from the first two waves of the survey have been released, with release of files from waves 3, 4 and 5 expected soon. A microdata

file derived from the topical module (assets and liabilities) collected during wave 4 has also been released. Although the five wave files were processed individually (cross-sectionally), taken as a set they represent longitudinal microdata covering 15 — 20 months of data for individuals. Work is underway to create a truly longitudinal microdata file from SIPP. This survey represents a new challenge — disclosure-avoidance in a detailed longitudinal survey. These implications were discussed at a November 1985 seminar sponsored by the Social Science Research Council (proceedings forthcoming).

Because of the highly skewed distribution of business and manufacturing firms, the Bureau of the Census does not release microdata from its economic censuses and surveys. However, it has constructed a microdata file, the Longitudinal Establishment Data file (LED), derived from census of manufactures and Annual Survey of Manufactures data going back into the 1960's. There is considerable interest in this file by economists, and policies on its access and use are being developed. One option under consideration is to construct a file of (synthetic) microaggregated microdata by aggregating entire records for small subsets of similar establishments from the LED file. This file would be used to educate a large class of users in the structure and opportunities and characteristics for analysis of the LED. Those who sought more realistic analyses on the actual LED file would then be in a position to make specific proposals to the Bureau of the Census for Census-performed reimbursable work.

The Bureau of the Census performs a great deal of statistical data collection and processing for other government agencies and organizations. Such surveys, performed for the survey sponsor under contract on a cost-reimbursable basis, are usually performed under Title 13. Consequently, these files are subject to the same disclosure-avoidance policies and procedures as Census public use files. This can cause consternation for the survey sponsor who paid for the collection of the survey data, but may be entitled to receive it only in censored or summarized form.

The Bureau of the Census has an established review procedure for determining whether and in what format a microdata file should be released. These reviews, discussed in the next section, are applied uniformly to all microdata files prepared for release. The list of microdata files reviewed for release during 1985 follows.

Approved for release:

Revised Current Population Survey
1984 Survey of Scientists an Engineers
1980—81 Consumer Expenditure-Quarterly Interview Survey
1982—83 Consumer Expenditure Survey Diary
1982—83 Consumer Expenditure-Quarterly Interview Survey Wave 3 Topical Module for the 1984 Survey of Income and Program Participation (SIPP)
1982 Truck Inventory and Use Survey
Puerto Rico Public-Use Microdata Samples
Wave 4 Topical Module for SIPP
Additions to the Job Training Longitudinal Survey Terminee Control Card Administrative Data File

Disapproved for release:

1981 Survey of Work Experience of Young Men with State Codes Added
Census of Private Juvenile Detention, Correctional and Shelter Facilities

Pending:

Job Training Longitudinal Survey Interview Files Revised 1982—83 Diary Survey

The Microdata Review Panel

The decision whether to release a Bureau of the Census microdata file is made by the Microdata Review Panel. The Panel was created in 1981 for that purpose, and to ensure that reviews are mutually consistent. The Panel has the further aim of developing standards for microdata release and providing guidance to sponsors and their Census Bureau counterparts on the fundamentals of microdata disclosure-avoidance and the design of microdata files meeting minimum acceptability criteria. The list at the end of the preceding section represents the workload of the Panel during 1985.

The Panel consists of senior representatives from the Demographic, Economic and Statistical Standards & Methodology directorates, the Data User Services Division and the Program and Policy Development Office. Panel members rotate being Chairperson.

Continuing microdata releases, such as the ADF, are approved only once by the Panel unless there is a change in the file structure, the survey design or other important factors. Other files, particularly those from reimbursable surveys or files derived from the topical modules of SIPP, are reviewed as they are proposed. The Panel has been successful at educating and working with sponsors' Census Bureau representatives, so that beyond the early years of the Panel unfortunate "surprises" have become a rarity. The review is conducted based on information provided by the sponsor in response to a standard checklist of key identifiability vulnerabilities, required information on the level of geography to be shown on the file, and any special tabulations the Panel may require. The Panel is available to meet with those concerned before, during and after the review. In fact, the Panel encourages such informal contact at the earliest possible stage, preferably before agreements are signed and survey and product designs finalized. Collection or presentation of data which later might be sacrificed to preserve confidentiality is often thus avoided.

There are two hard and fast rules which the Panel must enforce — the 100000 population threshold (reduced from 250000 with the creation of the Panel in 1981) and the 0.5% income topcode, currently $ 100000 and revised periodically (revised from $ 75000 in 1985). All other judgments are based upon available information, the existence and structure of population registers which can be matched against the file, general policy considerations, established precedent, and experience with similar files. The principal tool in the review is the checklist, which appears here as Appendix.

The Panel does not simply approve or disapprove a microdata file for release, but, for a rejected file, typically suggests disclosure-avoidance methods which, if applied, would in the Panel's judgment reduce disclosure risk sufficiently to permit release. These methods include:

- data grouping
  - * aggregation, including combining (collapsing) definitional categories
  - * topcoding
  - * recoding or range coding
  - * reducing geographic detail
- suppressing certain data items entirely
- data perturbation
  - * error innoculation (quantitative variables)
  - * data swapping (categorical variables)
- data rounding.

The Panel does not have at its disposal the tools or staff necessary to perform a thorough disclosure risk analysis (generalized software for tabulation, matching, simulation and reanalysis), so its determinations, although based upon the information and expertise described above, remain subjective. What is being done to remedy this situation is to develop the knowledge and expertise needed to establish (1) a broad, statistically defensible set of guidelines for minimizing disclosure risk in microdata, based upon (2) a coherent statistical definition or standard for disclosure risk in Census microdata products and methods and tools for its measurement. Much of the knowledge required to do so will be new. A project was begun in late 1985 to investigate these problems and develop appropriate software. Much of this kind of work has been reported by Dr. Paaß and his colleagues at the Gesellschaft für Mathematik und Datenverarbeitung (GMD) and we hope to work closely with them in these investigations.

The project has just begun and detailed findings are not yet available. The scope of the effort is clear, however, and will include the areas of investigation listed below:

- state-of-the-art matching methods
- the existence and structure of population registers which could be used in reidentification
- development of a statistical definition or standard for disclosure risk in microdata and methods for its measurement
- modern or new disclosure-avoidance methods

— file splitting

— microaggregation and synthetic microdata

— the effects of alternative disclosure-avoidance methods on the quality, completeness and usefulness of the data file.

Concluding Comment

Historically, agency policies for preserving confidentiality in microdata developed in a reactive manner in the United States. Agencies responded to user demands and needs for richer data than could be provided in tabulations. With computing power and file matching methods being primitive in these early days (the 1960's), simple disclosure-avoidance procedures sufficed. From these early release procedures precedents were set and user appetites for microdata increased. This situation continued into the 1970's with the amount of microdata being released far outstripping the perceived and, at that time, real need for more comprehensive disclosure-avoidance policies and methods. With the dramatic increase in computing power experienced in the 1980's came more powerful file matching methods, and thus, reidentification methods. There was also a corresponding increase in the number of population registers on subgroups and enhanced accuracy and level of detail of the data they contain. A consequent increased focus on privacy issues within society developed. This situation is complicated within the decentralized U. S. statistical system because often the holders of these threatening population registers are other government agencies, some of which have both statistical and regulatory functions.

The first steps to deal responsibly with these problems were institutional, the Microdata Review Panel (MRP) being an early and effective example. The second step now is being taken: full assessment of disclosure risk for microdata, leading to informed policies, guidelines and analytical software for microdata release.

References

Cox, Lawrence, Johnson, Bruce, McDonald, Sarah-Kathryn, Nelson, Dawn, and Vazquez, Violeta (1985): "Confidentiality Issues at the Census Bureau", Bureau of the Census — Proceedings of the First Annual Research Conference, Reston, Va., 199—218.

Mugge, Robert (1984): "Issues in Protecting Confidentiality in National Health Statistics", Review of Public Data Use, 12, 289—294.

United States Department of Commerce (1978): Statistical Policy Working Paper 2: Report on Statistical Disclosure and Disclosure-Avoidance Techniques, U. S. Government Printing Office, Washington, D. C.

Appendix

Instructions for Submitting a Proposal Requesting Approval to Release a Microdata File
(The Checklist)

The project manager should submit eight copies of each of the materials listed below to the Chairperson of the Microdata Review Panel (MRP) at least 1 month prior to the time approval is needed. The chairperson will arrange a Panel meeting to discuss the proposal. The project manager usually will be requested to attend the meeting or send a representative who is knowledgeable about the proposal and the corresponding survey. The Panel's decision to approve or reject release of the file will be documented in a memorandum to the appropriate Division Chief.

Required Materials
(Eight copies of each):

1. Cover memorandum from the Division Chief that includes a brief description of the purpose and design of the survey and any other relevant information; e. g., the date by which approval is needed.

2. A tape record layout (sometimes called a data-base dictionary) that lists all of the variables and other information with the specific categories that will be on the file proposed for release. If time or costs prevent advance preparation of a tape record layout for the initial MRP review, submit the survey questionnaire marked-up to show the items proposed for the tape with a description of how write-in entries will be coded and a list of any other information that will be on the tape (e. g., sample weights and geographic information).

3. A completed "Checklist" providing information needed by the Panel to evaluate the disclosure potential of the file. The Checklist asks about items that you are proposing to delete or change for confidentiality reasons and about items where you are not sure whether such treatment is necessary. One copy of the Checklist is attached; additional copies should be reproduced, as needed, by using this copy.

4. A table documenting that the population of every geographic area identified on the file proposed for release has at least 100 000 inhabitants. This requirement was established in the Criteria for Disclosing Public Use Microdata issued by the Bureau in February 1981. These criteria do not allow selective exemptions from the minimum population requirement; however, the Panel may determine that a higher population threshold is necessary to compensate for file content with greater-than-normal disclosure potential.

   The Panel considers the minimum population requirement met if each area to be identified has 100000 inhabitants in the areas subject to sampling (e. g., a Primary Sampling Unit (PSU)) as of the most recent census. Population estimates nearer the survey date may be used, if desired, unless the intended geography includes urban/rural, size of

place, or other categories summarized only in a census. Use of population data from another source, for example a prior census, must be approved in advance by the Panel.

The table should show the total population in sampled areas (PSUs) cross-tabulated by every geographic identifier to be shown on the file (see the example in Attachment). Geographic information, for this purpose, does not generally include information provided by the respondent; for example, farm status. The source of the population figures used in the table must be specified. If this file was previously released (or will be released again) with different geographic identifiers, the table must show both sets of geography in combination. Each cell in the table for an identified area or the remainder should show a population of 100 000 or more; if any cell falls below 100 000, the geographic specifications must be revised to meet the requirement and reflected in the table submitted to the Panel. If the sample was a PSU-based design, the total of all cells in the table should be the total population in sampled PSUs. If the sample was not selected from within designated sample areas (e. g., a sample from a list of addresses not in PSUs), the total of all cells should be the total population of all areas subject to sampling (often the entire U.S.). Please contact a Panel representative for more information on how to prepare the table when the sample is not a PSU-based design.

Attachment

Preparation of a Population Table

Example: For a file identifying urban/rural, central city/noncentral city/nonmetro, region, and selected SMSAs, the table could be presented as follows:

Table 1

	METROPOLITAN			NONMETROPOLITAN	
	Central City	Other Urban	Rural	Urban	Rural
Northeast:					
Identified SMSA a					
Identified SMSA b					
Remainder					
North Central:					
Identified SMSA c					
Identified SMSA d					
Identified SMSA e (pt.)					
Remainder					
South:					
Identified SMSA e (pt.)					
Identified SMSA f					
West:					
Total:					

Grand total = (total population in PSUs or other sample areas).

Source: 1980 census (posted from PC80-1-A reports)

Checklist on Disclosure Potential of a Microdata File

SURVEY TITLE:_____ DATE_____
Project Mgr.: Name_____ Div._____ Br._____ Phone_____
Sponsoring Agency:_____

(NOTE: If you need more space for an answer, please attached a continuation sheet. Be sure to identify the section and number of the question.)

To reduce your reporting burden, it is not necessary to make a separate application for every issuance of a repetitive survey, e. g., CPS or AHS, as long as geographic information is not changed, no new subject matter is introduced, and the disclosure measures approved for the first file are implemented on the subsequent files. Check the applicable category below:

[ ] This application is for a single file.

[ ] This application is for a series of files with substantially the same content. Specify the interval at which future files will be released _____ .

[ ] This application is for the re-release of an approved file, with the addition of supplemental or previously unreleased data. Give the date the original file was submitted to the MRP _____ .
(Only those checklist questions for which the answers are now different need be completed.)

Section A. Additional Geographic Information on the File

In addition to explicit geographic identifiers on the file, the data items, record identifiers, or file structure may provide additional geographic information by inference. Therefore, steps must be taken to avoid inadvertently identifying geographic areas that do not meet the 100 000 minimum population criteria. Potential problem areas are discussed below. For each area, please indicate the actions that have been or will be taken before the proposed file is released.

1. Primary sampling unit (PSU) or other geographic information usually is embedded in control numbers designed for internal use.

   o How will this problem be avoided on the released file:

   _____ Control numbers deleted or do not contain geographic information.
   _____ Control numbers scrambled; describe _____
   _____ Other; describe _____
   _____
   _____

2. Records in many data bases are sequenced so that the first cases are in the lowest numbered PSU or county that is first in alphabetic order.

   o Briefly, describe how the records on this file will be sequenced to avoid such geographic inferences. _____

3. Data items that imply specific geography of residence may reveal more than the explicit identifiers displayed on the population table prepared for the Panel. Examples: inclusion of Spanish surname (coded only in five southwestern states) when the explicit identification of that group of states will not be on the file; a migration code specifying movement from a metro area to a nonmetro area when metro-nonmetro will not be included as part of the geographic identifiers; residence within X miles of a nuclear reactor or an airport when there is only one in an identified geographic area; telephone area code; or latitude and longitude coordinates.

   o List all items that will be deleted for this reason: _____

   o List all other items that you think might have geographic significance, but could not decide if they should be deleted: _____

4. Sampling information also may provide some geographic indicators. For example, certain weights may distinguish between self-representing and nonself-representing PSUs or identify types of areas intentionally oversampled. Also, codes for "Durbin type", "Hit number", etc., may be related to geography.

   o List all sampling information that will be deleted for confidentiality reasons or subsampling plans to make weights less identifying: _____

   o List all other sampling information that you think might have geographic significance, but could not decide if it should be deleted: _____

Section B. File Contents Presenting an Unusual Risk of Individual Disclosure

The disclosure criteria for public use microdata require a review of each file to determine if any of the proposed contents present an unusual risk of individual disclosure. The MRP has identified several measures that can be taken to reduce the possibility of identifying an individual through the characteristics available on a file. The measures are discussed below and

relevant information pertaining to the proposed file is requested to assist the Panel in its review.

1. Names, addresses, and other unique numeric identifiers such as Social Security, Medicare or Medicaid numbers must be removed from the file.

2. High income is a visible characteristic of individuals or households and is considered to be a sensitive item of information. Therefore, each income figure on the file, whether for households, persons, or families, including total income and its individual components should be topcoded at a maximum of $ 75 000 (now $ 100 000). Exceptions to this rule are possible under certain circumstances; for example, if there is very little geographic detail. Variances from the $ 75 000 topcode should be discussed with the Panel well in advance of the final submission for approval to release a file.

   o Is the topcode

   _____ $ 75 000 or less

   _____ More than $ 75 000. Specify amount and briefly summarize discussions with the Panel _____
   _____
   _____

3. In addition to income, certain other characteristics may make an individual more visible than others; for example, very high age, value or purchase price of own property, rent, mortgage amount. Depending on the geographic detail shown on the file, consideration should be given to topcoding these items when they are represented as interval or ordinal variables. The Panel suggests that these topcode categories include at least ½ of 1 % of all persons or all households (depending on the type of characteristic) represented on the file. For example, topcodes for the 1980 census were:

   Age — 90 years old and over.

   Value of property — $ 200 000 or more.

   Gross Rent (including utilities) — $ 1 000 or more.

   o List all items that will be topcoded and the corresponding topcode: _____
   _____
   _____

   o List all other items about which you have questions regarding the need to topcode: _____
   _____
   _____

4. There are other characteristics that may make a person highly visible, depending upon the geography, that are represented as nonordinal variables, and therefore cannot be top-coded; for example, codes indicating Foreign or Indian Tribal language spoken; detailed racial identification such as Eskimo, Aleut, Guamian, or Samoan; codes for place of prior residence, etc. In these cases, the amount of detail on the file may have to be collapsed into larger categories.

   o List all items that will be collapsed (or deleted) for confidentiality reasons: _____

   o List any other items about which you have questions regarding the need to collapse the detail: _____

Section C. Disclosure Risks Associated with the Ability to Match to External Files

Efforts must be made to reduce the potential for matching microdata on this file to data on external files, because external files usually contain names and addresses, and thus can be used to identify survey respondents. Such matching may be possible if the survey contains highly specific characteristics that are also found on "population registers" or administrative lists maintained by other agencies or organizations. For example, the inclusion of vehicle make, model, and year in conjunction with specific geographic identifiers is unacceptable because these items can be matched to automobile registration lists that contain name and address. These items probably could be left on the file if they were recoded into broad categories. Other lists or registers include credit bureau files; a manufacturer's list of purchasers of particular major durable goods (for example, airplanes); voter registration lists in some states; Federal, state, or local tax records; criminal justice system records; state hunting and fishing license registers; and membership rosters of certain trade associations.

Matching is also highly possible if the sampling frame for a survey comes from a source outside the Census Bureau. The agency that provided the sampling frame may be able to match survey records to its original records, particularly if the survey records include data from the originating agency's files; e. g., amount of program benefit received, date of entry into program.

Section C.1

1. Are you aware of a population register or other administrative list that contains data also included in this proposed file?

   _____ Yes — Identify the list(s) _____
   _____ No

2. Based on readily available information, will any data item on the file identify residence in a particular type of institution of which there may be only one in an identified area; or for which a population register could be obtained?

_____ Yes — Identify the type of institution _____
_____ No

3. Were any of the sample cases contained in the proposed file selected from a list provided by a source outside the Census Bureau?

_____ Yes — Identify the source and describe how and by whom sample cases were selected from the list _____
_____
_____ No

Section C.2

When an external file exists, several steps may be taken to reduce the possibility of matching survey data to this file; for example, selected items may be deleted or recoded, or "noise" (i. e., small amounts of random variation) may be introduced into these items. The Panel cannot specify in advance exactly which steps must be taken to reduce sufficiently the potential for matching. However, it does consider several factors in determining the risk associated with releasing a file when the possibility of matching to an external data base exists: 1) the number of variables available for matching purposes,2) the resources needed to perform the match, 3) the age of the data, 4) the accessibility, reliability, and completeness of the external file, and 5) the sensitivity or uniqueness of the data. Some factors that make matching easier are listed below and information is requested on steps that will be taken before the file is released to reduce the matching potential. (NOTE: This information is necessary even if you are not aware of any external files that could be used in matching.)

Matching is easier —

a) . . . if any data item or combination of items isolates any small and readily identifiable population. The inclusion of codes that identify very small population segments should be avoided; for example, Indian tribes or condominium status in combination with highly specific geography. Normally one does not have to consider more than one variable at a time unless that group of variables is likely to appear together on an population register. For example, age and sex are likely to appear together on external files but not country of birth and occupation; thus, it should not be necessary to protect against rare occurrences like Russian-born architects.

o List all data item(s) proposed for inclusion on the file that isolate a small, readily identifiable population: _____
_____
_____

o List all data item(s) that will be altered (i. e., deleted, recoded, noise added) for this reason: _____

_____

_____

b) . . . if the file includes substantially every member of a population. Examples: large employers, high-income individuals, doctors, scientists of a specified type, or inmates of certain types of institutions. Subsampling frequently is required within certain strata prior to data release.

o Identify these populations, if any are on the file, and how they will be subsampled. _____

_____

c) . . . if the file contains any information obtained from records or other sources where that information could serve as a link to an external file that has individual identifiers or detailed geographic information. Examples include fuel consumption or cost records from a utility company; neighborhood, tract, or ED summary characteristics from a decennial census; welfare or social security data from a private or government agency; arrest records from a police department.

o List all data item(s) proposed for the file that were not obtained from an interview with the respondent: _____

_____

o List all data item(s) altered or deleted for this reason: _____

_____

d) . . . if the file includes data items frequently used for matching, such as exact date of birth, sex, and race, or if it includes other items that should be identical on both files such as an exact income amount, real estate taxes or other taxes, or date of entry or termination from a government-sponsored program.

o List these data items, if any: _____

_____

o List all data item(s) altered or deleted for this reason: _____

_____

e) . . . if longitudinal data are being collected; i. e., if the data for the same respondents/units will be collected for several different reference periods. Primary concern relates to time series of data items potentially matchable to outside records; e. g., income tax or employment records. If data are collected from the same respondents more than once, indicate the frequency of interview, length of time any one unit may be in sample, and factors affecting the likelihood of matching a sample unit from one time period to the next. _____

_____

_____

f) . . . if highly specific geography is included on the file; for example, states, SMSAs, etc. (This geography should be presented in the Population Table.)

g) Describe any considerations not previously mentioned that reduce the ability to match this file to an external register; e. g., unreliability or natural noise in the data. _____

_____

_____

Section D. Other Issues

1. Files that include every sample case are more likely to lead to disclosure than files containing only a subsample of cases. For example, if it were known that a certain individual participated in a particular survey, one could infer that the person's record could be found in the corresponding microdata file, assuming all sample cases were available on that file.

   o Does this file contain

   _____ Every case
   _____ A subsample of cases

2. Project managers should be aware that confidentiality problems may arise if special tabulations are made from an internal version of a file, which includes detail omitted from the public use file. For example, the tabulations might provide specific geography not included on the public use file, cross-tabulated by multiple data items on the file. The Panel has prepared guidelines outlining procedures for reviewing these tabulations. Please refer to these guidelines ("Disclosure Potential of Survey Tabulations Given the Availability of Public Use Microdata") and consult with the Panel if you are planning to release tabulations that make use of detail not available on the public-use file.

Chris J. Denham
Office of Population Censuses and Surveys, London

Census Microdata in Great Britain: The Possibilities *)

Introduction

The possibility of releasing microdata from the British Census of population was discussed throughout the 1970s and has continued to be an issue in the 1980s. Much of the demand has come from researchers familiar with such tapes as well-established means of dissemination from the censuses of the United States and Canada. The earlier demand was reviewed by the Office of Population Censuses and Surveys (OPCS), (Hakim 1978) and the continuing demand has been reviewed from a user's point of view (Norris 1983). More recently, the accessibility of census microdata to researchers in Italy has prompted some fresh demands (Openshaw, 1986).

Microdata from official surveys are made available by the OPCS, but no census microdata have been produced as yet in Britain. However, in 1981 the Government expressed its willingness to consider the production of microdata in a White Paper (an official publication of government policy) in the following terms:

> "The Registrars General [1]) invite . . . any . . . interested body or person to make proposals on the form of a public use tape that would serve a wide range of users whilst effectively protecting confidentiality" (Her Majesty's Stationery Office [HMSO] 1981A).

This statement, however, was hedged around with important qualifications relating to value and cost, legality, and confidentiality. The production of census microdata would have to meet three main criteria:

— the uses of the tape would have to be sufficient to justify its costs,

— statutory (legal) authority would be required, and

— the data would have to be designed to protect confidentiality.

It has also been made clear that the statutory position of the Census Offices makes it necessary that any users who wish to have microdata will have to order, specify and pay for the data. The Census Offices cannot initiate the production of microdata as a 'speculative' venture without there being a customer willing to meet the cost, although the Offices would decide whether a particular proposal met the criteria.

*) The views expressed in this paper are those of the author and are not necessarily those of the Office of Population Censuses and Surveys.

1) The Registrars General are the directors of the two Census Offices in Great Britain: OPCS which has responsibility for England and Wales, and the General Register Office which has responsibility for Scotland.

At the date of writing only one group of users has effectively taken up the Government's invitation on census microdata. Indeed it is probably fair to say that census microdata have not been the subject of major debate in the community of census users in Britain, nor, perhaps as a consequence, have microdata been a matter of any public or political controversy. The final section of the paper looks at some of the ways that census microdata and related means of dissemination might develop in future.

1 Census microdata in Britain

The potential importance of census microdata

1.1 Although there is other official statistical material in Britain that might, and in some cases does, provide microdata on the social and economic condition of the population, considerable interest has focussed on the census of population for the following reasons:

- the example set by census microdata in the United States and Canada;
- the fact that there is no other source in Britain for social science research comparable to the census in comprehensiveness;
- other official data collected for administrative purposes cover fewer attributes, for example, the returns on the unemployed, and/or are not made available in a detailed statistical form, for example, taxation returns; and
- there is no general registration of people, households or housing for statistical purposes in Britain, nor any general linkage of official data for statistical purposes that might provide an alternative to the census;
- Census microdata appear to offer large rewards to the social scientists who can overcome the difficulties of meeting the Government's criteria.

1.2 Historic census microdata are available in Britain (Norris 1983). Census forms with names and addresses are preserved as public records and become open for public inspection and use when 100 years has elapsed from a census – so historians and genealogists may now extract data from the 1881 census forms. Indeed, some machine readable historical microdata have been produced by researchers. The machine readable records from recent censuses are similarly preserved, although the data are encoded and names and addresses are not included. However, the issues relating to these historical microdata are mainly distinct from those relating to current data. Few precedents are provided that relate to current microdata and the subject is not taken further in this paper.

The concept of census microdata

1.3 The terms used to describe ". . . individual data for a sample of persons in anonymized and machine readable form . . ." (HMSO 1981A) in Britain vary interchangeably and without real difference of meaning between: 'public use sample', 'public use tape' and 'microdata'. There tends, however, to be a more restricted notion of microdata among those responsible

for official statistics than among potential users – a not altogether unexpected state of affairs.

1.4 The official view of census microdata is on the lines that

— such data would relate to small samples of persons or households – one in a hundred or less – taken from the coded and verified basic census files containing all processed attributes;

— more than one sample might be drawn to provide different combinations of attributes for different uses;

— the risk of the identification of individuals through combinations of characteristics and subsequent risk of disclosure of additional characteristics (which are two related but distinct risks, with disclosure normally of more serious consequence) would be reduced by one or more of the following:

— removal of all forms of personal identifiers, addresses and most local area codes,

— coding data into a limited number of categories, and

— corruption of the data;

— on balance, restriction of access to microdata to selected users would not be helpful to public trust in confidentiality measures and microdata should be open to all; and

— the record would be in machine readable form.

1.5 Potential users of census microdata have suggested that the concept should be less restrictive (Norris 1983). They have stressed, in particular, that microdata should have 'hierarchical' links between persons within households. There have been requests that there should be some type of microdata for all individuals in a sample of small areas, or with detailed area codes, so that geographical patterns could be studied. Growth of computing power would also enable users to analyse large sets of microdata. Only in such ways, it is claimed, can the limitations of census crosstabulations on the analysis of complex social and geographical patterns be avoided.

2 Microdata from Social Surveys in Britain

2.1 OPCS is responsible for conducting social surveys for the Government. The primary recipients for the results of these surveys are the Government departments that commission the surveys, although reports of nearly all surveys are published. However, OPCS has recognised that there is often further potential for academic research in the survey data and anonymized data tapes from the surveys are released for research purposes. The practice

has been established for five or so years and the material released includes data from the regular General Household Surveys (GHS), Family Expenditure Surveys (FES) and Labour Force Surveys (LFS) plus data from certain 'ad hoc' surveys of a more specialized nature. There is relatively little literature describing the availability and use of the survey data tapes, and the most comprehensive guide was published some four years ago (Hakim 1982).

2.2 The release of the data tapes is seen as marginal to the main task of conducting and reporting surveys. It is a response to users' interest in being able to re-analyse the data, and is done on the basis of minimal additional cost to OPCS. However, OPCS feels that the release of data helps the dialogue with the research community and that new methods of data analysis may develop from wider use. No special legal, administrative or financial measures have been introduced in connection with the release of data tapes.

2.3 At present, the Data Archive of the Economic and Social Research Council (ESRC) at the University of Essex is the sole repository for the survey data tapes (the ESRC is an officially supported organisation funding research, and the Data Archive is the major national repository for social science data in Britain). All enquiries about the data are currently channelled to the Archive, but OPCS would be prepared to supply copies of the tapes to other research institutions if requested. Data from the regular surveys are passed to the Archive without charge, except for the recovery of administrative expenses, and the Archive in turn is not permitted to charge for the data. Charges, however, are made for data from 'ad hoc' surveys. In fact the data tapes are not modified for use as microdata before release and the user is faced with some expense in handling the data. The tape structures are relatively complicated and, for example, the user has to re-create derived variables used in the survey reports. Nor is the documentation designed for external consumption.

2.4 Measures taken to anonymize the data before release of the tapes are straightforward:

— names, addresses and other personal identifiers are removed,

— all data are encoded, that is placed into categories, and

— data are geographically coded to standard region level only (10 areas in Britain).

Otherwise there is no corruption of the data or re-coding into broad categories. Two other aspects of the data contribute significantly to ensuring anonymity and to ensuring that there is very little risk of disclosure. First, the names and addresses of the people and households selected for survey samples are not known outside OPCS, and the chance of any particular individual being in a sample are many hundreds to one against. Second, the unknown date of actual survey of any individual in a 'continuous' survey and the lapse of time before access to the data increase the difficulty of identification – data tapes are released one year after initial production, so data are 18 months to two years out of date. Finally, of course, any individual has the option of not providing data, as all surveys are voluntary.

2.5 There are no particular sanctions against misuse of the data tapes, although OPCS could opt to cease supply at any time. So far, the release of data has not attracted any adverse publicity or discernable public reaction during surveys.

2.6 In comparison to the census, the survey microdata include much wider ranges of attributes, but sampling fractions are relatively small and geographic coding is coarse. It seems that survey microdata do not offer a complete alternative to census microdata and that the two would be complementary.

3 The need for census microdata in Britain

3.1 The first of the three criteria put forward for the production of census microdata is that uses should justify costs. This depends in part on what is put into the census, and on what statistics are already available and whether any relevant precedents have been set for microdata. (Background to the 1981 Census can be found in a number of retrospective studies, for example: Thatcher 1984, and Denham 1985.)

The input

3.2 In brief, the 1981 Census was planned to be as straightforward as possible so that it would not be a burden on the form-fillers and so that results could be produced quickly. The most important consequence for microdata was that the Census had a limited number of questions by international standards, and hence a limited number of variables for relational analysis. But the resultant data were 'worked hard' to provide a relatively large range of aggregated data. The census questions are shown in figure 1, page 50.

Aggregated data

3.3 A description of the products available from the 1981 Census has been published by the Census Offices as a User Guide (OPCS 1985), but, in summary, before considering microdata the census user has the choice of two types of source of aggregated data:

1) printed reports, sold by government bookshops or directly from the Census Offices (the reports comprise either series of local reports or reports on major topics at national level), or

2) statistical abstracts obtainable, for a charge, from the Census Offices either as

a) standard abstracts giving comprehensive data for small areas, or

b) 'customized' abstracts designed to a user's requirements.

All such output is in the form of univariate counts or cross-tabulations of counts with two or more dimensions up to a normal maximum of five or six; derivced statistics such as ratios are rarely used except in summary reports.

3.4 The cost of collecting and processing the census to the stage of fully validated data files is met by the Government. The cost of preparing the reports is also met by the Government, and only the costs of printing are recovered from purchasers. Customers pay the full marginal additional costs of the abstracts which can be high in the case of standard customized abstracts, although costs may be shared in the case of abstracts.

Small Area Statistics

3.5 At the core of the 1981 Census output lie the Small Area Statistics (SAS) – standard abstracts designed through detailed consultations with users and from which a whole family of census products has been developed (OPCS 1983). The SAS are a standard set of cross-tabulations for areas throughout Britain, covering every topic in the census and, for each area, giving over 4000 separate statistical counts. The SAS are available for each of the 130000 enumeration districts (small zones averaging between 150 and 200 households each) in Britain and a wide range of larger areas built from the enumeration districts. SAS are available on tapes of various standard formats, and in microform and printed versions.

3.6 The SAS tables became the basis of local census reports and extracts formed summary reports. Many of the SAS tables, in expanded form also feature output at all geographical levels, as well as in specialized and summary reports, and have the benefit of widely comparable data.

3.7 The SAS have been purchased and used by virtually every local and health authority in Britain, and are available for use throughout the academic and research community. An important aid to this widespread use of SAS was the establishment of an independent consortium, before the census, to commission the production of a software package – SASPAC – to handle the SAS on every type of computer possessed by the members of the consortium.

3.8 In summary, the comprehensive development of the SAS and their very widespread availability may have drawn use of the census towards area based and mainly local studies using simpler standard statistics rather than towards national or regional studies requiring more complex statistics, perhaps provided by analysis of microdata.

Confidentiality measures in existing sources

3.9 In designing all forms of output, the Census Offices aim to ensure that categories and counts will not become too exclusive in relation to the size of typical populations covered. The statistics in reports are not modified or corrupted in any way. (In fact, rounding counts to the nearest 0 or 5 was dropped after the 1971 Census with no adverse consequences.) In the SAS, counts drawn from the 100% processing of persons and households (see figure 1, page 50) are modified for confidentiality reasons by the addition of 0, +1 or −1 to each cell (except for basic counts of persons and households), and the SAS are supressed entirely if there are less than 25 persons or 8 households in an area.

The limitations of existing analyses

3.10 Despite census users having access to sources that exhaustively cross-analyse census topics and provide a vast amount of local detail, particularly in relation to the limited range of questions asked, many would still agree with a conclusion reached before the 1981 Census that ". . . the non-availability of microdata constitutes a significant limitation on the type of census analysis which can be carried out in Britain" (Hakim 1978).

3.11 The continuing demand for microdata arises largely from academic and research institutions concerned with the social sciences, also to some extent from users in the private sector undertaking 'market segmentation', and, to a limited extent, from Government departments and a few of the larger local authorities. Their requirements may be summarized as:

- needs to relate variables or to study populations and sub-groups not otherwise covered in standard census sources and/or to avoid the risk of spurious correlations between counts presented in separate cross-tabulations, that is to avoid the risk of 'ecological fallacy';

- the possibility of developing new derived variables or classifications from a sample of microdata;

- a wish to examine underlying distributions in data otherwise presented in generalized classes, including fixed geographical divisions; and

- a wish to avoid the delay and cost of obtaining customized tables.

3.12 The Census Offices would also see some likely advantages in being asked to produce microdata:

- a net reduction in the resources required for producing customized output (although the resources required to produce microdata are significant);

- an increase in the benefit from the Government's investment in the census through the production and use of analyses which could not otherwise have been produced within the Census Offices' limited programme; and

- the possibility of the development of improved methodologies for the future analysis of results.

3.13 In summary, there are applications for census microdata in Britain. But this does not necessarily show that the criterion of good value would be met. The discussion returns to this question later, in section 7, after the points of legality and confidentiality, which may have an important bearing on value, have been considered.

Figure 1: 1981 Census Questions

The 1981 Census questions for all people, whether they were in households or in communal establishments like hospitals or hotels, asked about:

age (date of birth)
sex
marital status
relationship to head of household*
whereabouts on census night
usual address
usual address one year ago (migration)
country of birth

And for all those aged 16 or over:

economic activity in preceding week
employment status (self employed, employee, etc.)
industry of employment (name and business of employer)*
occupation*
address of work-place*
means of daily journey to work*
higher qualifications*

* analysed for a sample of the population – see below.

In Wales there was an extra question on Welsh language and in Scotland an extra question on Gaelic.

In addition, the form-filler in each household was asked about:

number of rooms
tenure
amenities
number of cars and vans

If households were absent, or accommodation was vacant, the census enumerators recorded basic information. For communal establishments, the type of establishment was also recorded as was the position or status of the people in the establishments, for example, whether they were staff, residents or visitors.

Sample processing

Answers to some questions (for example, date of birth) were straightforward and could be put easily and quickly into the computer for processing. Answers to other questions (for example, occupation) were more complex and for most of these – marked by an asterisk in the list above – the coding of the answers was done for a carefully selected sample of 10% households (and a 10% sample of people in establishments). So these required more time and separate processing, which is why census statistics from the '10%' topics are often presented separately.

4 The statutory basis for census microdata

The 1920 Census Act

4.1 The second of the three criteria for the production of microdata is that they should have statutory (legal) authority. This is because the British census is conducted by statute (law) – the 1920 Census Act. No later statute impinges on the census in a major way, but some recent statements of official intentions are relevant and are mentioned below. The Registrars General (Directors of the Census Offices) have authority to do what is laid

down in the Census Act, but they hove no authority to do what is not covered by the Act. So any production of census microdata must be judged to come within the authority of the Act.

4.2 One brief section of the Act sets out the authority for disseminating data from the census. After a clause instructing the Registrar General to prepare reports for Parliament (these are the main printed reports of the census sold in Government bookshops), a second clause states that

> "The Registrar General may, if he thinks so fit, at the request and cost of any local authority or person, cause abstracts to be prepared containing any such statistical information, being information which is not contained in the reports made by him under this section and which in his opinion it is reasonable for the authority or person to require . . ." (Census Act, 1920 [10 & 11 GEO. 5. CH 41.] section 4. (2)).

Microdata could not be presented as a printed report, so, if they are to be produced, it must be as "statistical abstracts". In this case it follows that the microdata would have to be ordered, specified and paid for by the user. But it must also be established that the issue of microdata is within the authority given by the Act, and that no other provisions rule out microdata.

4.3 The Act does not make any explicit provision for any confidentiality measures to be applied to statistical abstracts which might affect microdata, although a clause (8.[2]) provides quite severe penalties for publication or communication of information from the census without "lawful authority". It has been held, however, that these penalties may only appy to people engaged in collecting the census information in the field.

4.4 Recent data protection legislation in Great Britain would appear not to impinge on anonymized microdata since data on individuals must contain names and addresses or other personal identification to fall within the scope of the legislation. But further advice would be sought from the data protection authorities before the release of any census microdata.

Do census microdata have statutory authority – a logical proposition

4.5 The question of legality seems, then, to rest on the meaning placed on the phrase "abstracts . . . containing . . . statistical information". Microdata may be no different from a cross-tabulation. A count in the cell of a census cross-tabulation with several dimensions where there is no corruption or modification for confidentiality reasons may be seen as a short set of microdata about an anonymous individual, or individuals where there is a count of two or more. This may be illustrated from a table for the SAS for an enumeration district. Figure 2 (see page 53) shows an extract of SAS table 50 for one enumeration district (No. BNBA 14) in the Longsight Ward of the City of Manchester. The enumeration district had 544 people resident in 177 households in 1981, and its boundaries and the addresses within it are shown on large scale maps available to census users.

4.6 The extract of table 50 shows, for example, that there is one person who is (1) a resident, (2) aged 16 or over, (3) female, (4) married, (5) economically active (EA) but not in employment and (6) in socio-economic group (SEG) 9, a skilled manual worker. Thus the table

provides a short set of sample microdata with six broad-banded elements; there is also a detailed area code and copious 'environmental' data from the other SAS tables. The one in ten sample is felt to provide a reasonable safeguard against identification and disclosure of additional information. (The data are provided in such detail in the SAS to permit aggregation of the sample data within tables and between areas.)

4.7 Such a table, the proposition runs, has legality by precedent and practice, so the addition of further attributes, such as birthplace or the tenure of the accomodation of the person, to extend the set of microdata does not represent any logical difference in this position (except to increase the probability of the uniqueness of the sampled individual in most cases). Thus microdata are statistical abstracts because they are merely extensions of cross-tabulations, and thus microdata have legality.

4.8 However, the proposition of equivalence between microdata and cross-tabulations has not been tested in a court and it might well be that legal advisors or a court would place a limit of 'reasonableness' on such a 'logical' argument. It could be accepted that a six dimensional cross-tabulation was established statistical practice, but it could be shown that, say, tables of 20 dimensions had not been constructed, so any longer records of anonymized census data might be held to be materially distinct from a statistical abstract.

4.9 The Government has acknowledged that ". . . new legislation might be needed (for census microdata) because there is doubt whether the Census Act 1920 provides statutory authority" (HMSO 1981 A). No decision has been made to clarify the position by new legislation, and there must be considerable doubt whether such specialized legislation could be achieved in the near future unless it was part of wider legislation relating to official statistics. The legal uncertainty remains for the present.

Other official statements on microdata

4.10 Although statutory considerations must be paramount, there has been some official encouragement of the idea of statistical microdata. In 1981, the review of the Statistical Services by Sir Derek Rayner for the Government, which sought more cost-effective operations, recommended, as a general principle

> "Less costly (to Government) and more flexible means of enabling the public . . . to have access to figures held in government should be exploited. I have in mind . . . public use tapes . . . The costs of providing such facilities should be covered by appropriate charges . . ." HMSO 1981 B).

Also, in 1984, the Government published a code on practice on the handling of data obtained from statistical inquiries which stated

> "Where it is not forbidden by law and where no commitments have been entered into the contrary, a (government) department may transfer anonymous information about statistical units to another (government) department or to organisations and bona fide researchers outside government departments" (HMSO 1984).

So some significant official encouragement has been given in Britain to the principle of producing census and other microdata from official statistical material.

Figure 2: Example of a Small Area Statistics table*)

Residents, economically active or retired

Socio-Economic Groups	Economic position			
	All residents economically active or retired	Economically active not in employment		
		Males	Females	
			Single, widowed, divorced	Married
1	2	0	0	0
...				
9	8	0	1	1[1])
...				
17	3	0	0	0
TOTAL	30	3	1	1

The complete table gives counts for each 18 Socio-Economic Groups (SEGs) and also gives counts for retired males, the economically active and economically active migrants.

*) Enumeration District No. BNBA 14, Longsight Ward, City of Manchester. – Extract from table 50. The figures in this table are a 10% sample of the Census. The full table has 171 cells (9 cols. × 19 rows).

1) See text of paper.

5 Confidentiality

5.1 The third of the three main criteria for the production of census microdata is that confidentiality would have to be respected. Although the Census Act does not lay down confidentiality procedures to apply to statistical information from the census, the standard of confidentiality is set out in undertakings given to Parliament and on the census form. The undertaking given to every person completing a census form in 1981 was

> "Your replies will be treated in strict confidence. They will be used to produce statistics but your name and address will not be fed into the computer"

And the Government stated before the Census

> "As in the past the Census Offices will not pass information about identified persons or households to other government departments or to anyone else outside the census organisations" (HMSO 1978).

These statements do not rule out microdata, since the data would be issued only for statistical purposes, but they do set a stricter limit on the release of census data than in some European countries. No individual data, for example, are passed to municipal or local authorities for cross-checks of local registers or for other purposes, nor are individual data passed

to any other part of central government. The standard reflects a well-developed British concern with privacy and confidentiality (the impact on social research and the census in particular has been described by Cope [1979], whilst Hakim [1979] has described the evolution of census confidentiality procedures).

5.2 The Order and Regulations made by Parliament before every British Census make it compulsory by law for every householder and person in Britain to complete a census form. But, despite this power, the census can only be successfully taken on the basis of co-operation from public, and from their elected representatives and the media "acting on the public behalf".

5.3 Any actual or perceived breach of confidentiality by the Census Offices would be very likely to undermine the census, even where legal authority is not compromised and where the data breaching confidentiality were being passed on with the best intentions of serving effective dissemination. It is a risk with a potential loss to all census users likely to be far greater than any benefit to the users of census microdata.

5.4 The Census Offices must, therefore, when considering any proposal for microdata, ensure that there are adequate measures to meet confidentiality undertakings. It seems likely that measures to give anonymity and prevent disclosure would follow the straightforward and open approach adopted generally for the 1981 Census. Measures would be robust, visible and within the understanding of the general public. They would not seek the near impossible by aiming or claiming to give full security against disclosure, but would be designed to prevent any systematic disclosure and reduce the chances of disclosure to an acceptably low level of random occurences.

5.5 The measures that seem most likely to be adopted in the current climate, with other, more extreme measures like data corruption held in 'reserve', are:

- generalized area codes (the SAS are felt to represent the maximum acceptable combination of geographical and statistical detail);
- data coded to broad-banded classes, with non-essential data omitted; and
- sampling intervals of no greater frequency than 1 in 100.

A general principle would be that increase of detail in one of the main dimensions – geographical, statistical, or coverage - would be matched by reduction of detail in other dimensions. For example, microdata containing 'hierarchies' of persons within households would be stripped of any non-essential data and would have only the most generalized area codes. However, 'anonymous' environmental data could be added to individual microdata records, for example, area type (such as multi-variate classifications) or local indicators.

5.6 Steps would be taken to help avoid suggestions of convert breaches of confidentiality, or special treatment for selected users, by giving publicity to the availability of microdata and by selling the microdata to any person or organisation on the same basis as other custom-

ized abstracts, subject to the Registrar General's very rarely used statutory power of refusal. Public confidence might also be helped by having extracts of the microdata freely open for inspection.

6 The value of census microdata

6.1 If census microdata can be produced with the statutory authority of the 1920 Census Act, the Act puts the onus on the user to order, specify and pay for the data. Thus the census user has to weight the costs, and any statistical limitations of microdata, against the expected benefits.

6.2 The cost of a single general-purpose microdata file from the 1981 Census might be around £ 50000, and a complete package, say of two samples plus a handling system might cost a user around £ 100000. This arises in part because the Census Offices do not have existing files that they could convert readily to microdata. Sample selection, extract runs, modification for confidentiality measures and verification would all be required, as well as the production of a package back-up documentation.

6.3 The potential user of 1981 Census microdata must also face a number of statistical problems:

- microdata from the 1981 Census would, for all practical purposes, be limited to the codes (including those for derived variables) already in the census records, that is they would be limited by the preconceptions of the designers of the existing tabular output, since resources are most likely to be available for re-coding on a scale of a 1 % sample, even if users could accept the delay;
- the limitations of existing codes may also restrict the selection of any samples of population sub-groups;
- the delay in moving towards a practical trial of microdata would result in any 1981 Census microdata being five or more years out-of-date;
- a limitation of microdata to small, geographically well-spread samples would severely diminish the opportunity to use the microdata to examine geographical relationships;
- the small sample size and the limitations on geographical detail would limit the scope of analysis of workplace, journey-to-work and migration data;
- the estimation of population values from sample microdata would be complicated by the clustering of individuals in sampled households; and
- there would be no opportunity for the user of microdata to link non-census data at individual level, any links of this type have to be done within the confidentiality confines of the Census Offices.

6.4 Overall, potential customers for 1981 Census microdata have to consider whether the data would represent value for money in terms of new analyses that could be produced in a situation where the number of variables in the census is limited and where there is little chance of introducing fresh coding for the microdata, where the data have been exhaustively cross-tabulated, where the results are widely and inexpensively available, and where customized tables can be produced for much less than the potential cost of a set of microdata.

A proposal for 1981 Census microdata

6.5 In Britain, one body – the Economic and Social Research Council (ESRC) – has expressed interest in purchasing 1981 Census microdata. It has sounded out interest among researchers and has a draft formulation of proposals (Openshaw 1985). (The ESRC is an officially funded organisation making grants for research, and the main channel for the supply of census statistics to universities and research institutes.)

6.6 The proposal states that 1981 Census microdata are still relevant to research and will set a precedent for future censuses. A conflict is recognised between what is safest and simplest for the Census Offices and what is best for research, so a number of overlapping proposals have been formulated in an attempt to test the as yet undefined limits of feasibility. These have also been given priorities on the basis of user interest; third and fourth ranking priorities are considered unlikely to offer enough value.

6.7 The options seen are:
1. Microdata without any geographical referencing (fourth priority).
2. Microdata with standard region level codes (10 areas in Britain),(fourth priority).
3. Microdata with county level codes (66 areas), (second priority).
4. Microdata with local authority district level codes (456 areas), (second priority).
5. Microdata with partly blurred district level codes (third priority).
6. Microdata with various small area classifications, plus coarse area codes (first priority).
7. Dichotomously coded microdata at fine area level (third priority).
8. A very small sample (0.1 %) of microdata codes to fine area level (third priority).

The proposal suggests that hierarchial linkages between persons and household characteristics should be retained where possible and that broad-banding of existing codes would be undesirable, but perhaps necessary for some 'sensitive' variables. In general, a 2% sample would be required, that is 2 out of every 10 households in the fully coded '10 %' census file.

6.8 The proposal is most developed in the geographical dimension. Less attention has been paid so far, for example, to problems of designing microdata for minority populations that are geographically spread, such as one parent families or 'ethnic' minorities. These would each amount to no more than a few % of a straightforward microdata sample, yet feature prominently as examples of populations said to be understudied because of the lack of microdata.

6.9 At the time of writing, a formal submission of the proposal is awaited by the Census Offices.

7 Beyond the 1981 Census

7.1 The Census Offices are planning for the next census in Britain on the assumption that it will be held in 1991 on conventional lines with much in common with the 1981 Census. The conditions for census microdata may not change. That is, there would be no change in the Census Act, there would be no change in the basic methods of disseminating and charging for census results, and there would not be a significant increase in the number of census questions and variables. Census users are also likely to become increasingly proficient at anticipating their needs when they are consulted about cross-tabulations in the standard abstracts, thus diminishing the number of otherwise un-met demands. Assumptions about the future climate of confidentiality are less easy. The climate may become more hostile, or more receptive to a widening use of microdata in research for reasons not necessarily connected with the data source or the research.

7.2 Possibilities for the future in Britain are:

a) Interest in census microdata declines as a result of continuing uncertainties and absence of key variables such as income or ethnicity, coupled with increased use of survey microdata or other sources.

b) There is further evolution of the current concept of census microdata either by:

- development of microdata as a statistical abstract ordered specified and paid for by users, particularly if 1981 Census microdata create a useful precedent; or by

- production of future census reports based on small samples that would provide a ready input for microdata and reduce costs; there was, for example, no early, sample-based national report of 1981 Census results and, should such a report be re-instated in a 1991 Census, microdata could be released early to meet more of the demand for customized abstracts; and

- in both cases, the design of coding and classification systems is done not only with tables for reports and abstracts in mind but also with microdata analysis in mind.

c) Microdata are enriched by linkage of data for samples of individuals from other sources. Such links would have to be made within a 'confidentiality perimeter' before release of anonymous microdata. Within OPCS a 1 % sample of 1971 and 1981 Census individual records are linked to various vital events also registered at an individual level at the Office; this 'Longitudinal Study' already provides a service of customized tables for research purposes, but, as yet, no microdata (Brown and Fox 1984).

d) The need for flexible analysis of census data is met by developments to provide access to the entire set of census records:

- by the establishment of a disclosure-proof system of on-line access by users to the basic census data files to permit users to produce customized counts, multi-variate counts, indicators and cross-tabulations for populations from small area level upwards; and

- by the introduction of finer basic areal building brick – 1.6 million postcode unit zones rather than 130000 enumeration districts in Britain – from which areas could be built with great flexibility for customized products produced on-line.

The latter development is being actively planned at the time of writing; the former will be studied during 1986. Should there be a system of on-line access, it will be important that the coding and classification of the basic census record is designed to permit maximum benefit from this method of access. It may be more difficult to encourage users to anticipate their needs if they are not actually working with draft tables.

7.3 It should be stressed that, although there are firm plans for developments that could impinge on the future of microdata, the possibility of census microdata has been considered so far very largely in the context of the 1981 Census. Nevertheless, should an affordable, on-line access system with disclosure-proof access to basic census data at very fine area level be developed, the production of any separate sets of microdata may become unneccessary.

8 Conclusions

Census microdata – 10 years as a possibility

8.1 It is more than 10 years since the first suggestions were made in Britain for census microdata, and almost five years since the Government accepted that there was a case for microdata, yet none have been produced. The underlying doubts about the legality of microdata and caution by the Census Offices over protection of confidentiality have played a part, and the need to complete high priority output from the 1981 Census would in any case have delayed microdata. But a significant factor in the lack of progress seems to be that users find that the onus for initiating, designing and paying for microdata is on them, while the Census Offices retain the power to decide whether or not to produce the data once asked.

8.2 Although there is no hard-and-fast evidence about the amount and value of analysis of the census that cannot be done because of the lack of cross-tabulations, there is no overwhelming evidence of un-met needs. Perhaps researchers have already gone elsewhere for their data, but the combination of a limited number of questions in the 1981 Census, codes and classes designed for pre-planned tables, a wealth of small area statistics and a dominance of geographically precise area-based uses must all weigh against the likelihood that

there is sufficient of value in the existing basic files of the 1981 Census to be exploited as microdata.

8.3 The Census Offices have maintained an open-minded and co-operative approach over the possibility of microdata, and production from the 1981 Census remains a definite prospect – perhaps, first, as a trial to measure the value of the data and to test reactions for future censuses. There are also ways of improving prospects for microdata from the 1991 Census. But there are also some indications that the potential of microdata may be subsumed before the next census by a move to on-line access that would offer a combination of customized output and geography beyond the scope of conventional microdata.

Further reading

1. All developments in the 1981 Census have been reported, from February 1978 onwards, in the OPCS Monitor 'CEN' series – a newsletter issued several times annually, without charge by the Information Branch of OPCS (address above) to census users in Britain and to those interested outside Britain. The Monitor is now beginning to cover developments for the 1991 Census.

2. Further reading, generally of a more specialised nature, is covered by the references given below. For a general introduction to the availability and use of microdata from official sources in the context of social research, the following book by Dr. Catherine Hakim, Principal Research Officer in the Department of Employment, and formerly a Senior Research Officer in OPCS, is recommended

   Hakim, C., (1982) Secondary Analysis in Social Research: a guide to data sources and methods with examples. London: Methuen.

   The book is a comprehensive review of the secondary analysis potential of major British datasets, making the distinction throughout between microdata and aggregated statistical data, and describing both types of sources. The book is particularly helpful in describing the microdata available from the major Government sample surveys – the General Household Survey (GHS), the Family Expenditure Survey (FES) and the Labour Force Survey (LFS) – and the uses to which the data are put, a subject beyond the scope of this present paper on census microdata and one not systematically described elsewhere in official or other literature.

References

Brown, A., and Fox, J. (1984): "OPCS Longitudinal Study: ten years on." Population Trends 37: pp. 20-22.

Cope, D.R., (1979): "Census-taking and the debate on privacy: a sociological view." In Censuses, Surveys and Privacy (M. Bulmer, ed.). London: Methuen, pp. 184-198.

Denham, C., (1985): "The 1981 Census in retrospect." Journal of economic and social measurement. Vol 13., No. 1.

Her Majesty's Stationery Office – HMSO (1978): 1981 Census of Population. Cmnd 7146. London: Her Majesty's Stationery Office.

Her Majesty's Stationery Office – HMSO (1981 A): 1981 Census of Population: Confidentiality and Computing. Cmnd 8201. London: Her Majesty's Stationery Office.

Her Majesty's Stationery Office – HMSO (1981 B): Government Statistical Services. Cmnd. 8236. London: Her Majesty's Stationery Office.

Her Majesty's Stationery Office – HMSO (1984): The Government Statistical Service Code of Practice on the Handling of Data Obtained from Statistical Inquiries. Cmnd. 9270. London: Her Majesty's Stationery Office.

Hakim, C., (1978): Census confidentiality, microdata, and census analysis. Occasional Paper, 3. London: OPCS.

Hakim, C., (1979): "Census confidentiality in Britain." In Censuses, Surveys and Privacy (M. Bulmer, ed.). London: Methuen, pp. 132-157.

Hakim, C., (1982): Secondary Analysis in Social Research: a guide to data sources and methods with examples. London: Methuen.

Norris, P., (1983): "Microdata from the British Census." In A Census User's Handbook (D. Rhind, ed.). London: Methuen, pp. 301-319.

Office of Population Censuses and Surveys – OPCS (1983): Small Area Statistics (SAS): Background notes and SAS tables for Great Britain. OPCS User Guide 88. Titchfield: OPCS.

Office of Population Censuses and Surveys – OPCS (1985): Guide to Census Sources. OPCS User Guide 199. Titchfield: OPCS.

Openshaw, S., (1985): A proposal for the purchase of a sample of micro-census data from the 1981 census of population. Unpublished paper: University of Newcastle-upon-Tyne.

Openshaw, S., Sforzi, F., Wymer, C., (1986): "A national classification of individual and areal census data: methodology, comparisons and geographical significance." Papers and proceedings of the Regional Science Association (forthcoming).

Thatcher, A. R., (1984): "The Census of Population in England and Wales." Population Trends 36: pp. 5-9.

Prof. Dr. Walter Müller/Prof. Dr. Richard Hauser
Universität Mannheim/Universität Frankfurt

Der Bedarf der Wissenschaft an anonymisierten Einzelangaben

In dem von Karl Martin Bolte und Stefan Hradil 1984 veröffentlichten Band „Soziale Ungleichheit in der Bundesrepublik Deutschland" finden sich im Kapitel „Armut" die folgenden Tabellen 1 und 2. Sie zeigen, daß in den noch stark durch wirtschaftliches Wachstum geprägten sechziger Jahren der Anteil der Bevölkerung, der in Armut lebte, in der Bundesrepublik kleiner geworden ist. Aber was soll man mit diesem historischen Befund im Jahre 1986? Bolte und Hradil schreiben als Abschluß ihres Armutskapitels: „Allerdings ist zu vermuten, daß sie (die Armut) infolge der 1974 verstärkt einsetzenden Arbeitslosigkeit dann wieder zunahm." Ist es nicht ein Armutszeugnis für die Autoren, daß sie in einer so zentralen und politisch so bedeutsamen Frage so unpräzise sind? So leicht kann man den Autoren diesen Vorwurf jedoch nicht machen. Es gibt zur Zeit keine präziseren Aussagen dazu, und dieser Mangel ist ein Beispiel für die Folgen des Umstandes, daß Daten, die vorhanden sind, der Wissenschaft nicht in einer Weise zur Verfügung stehen, daß sie die entsprechenden Untersuchungen durchführen könnte. Würden die Daten der Einkommens- und Verbrauchsstichproben (EVS) für die Jahre 1978 und 1983 bereits der Wissenschaft in Form von Einzelangaben zur Verfügung stehen, so hätte man diese Zeitreihen bis zu einem gegenwartsnahen Jahr fortführen können.

Ohne die Möglichkeit des Rückgriffs auf Informationen über einzelne Individuen ist dies jedoch nicht möglich, denn Berechnungen von Armutsquoten auf der Basis des Einkommens involvieren komplexe Operationen. Sämtliche Einkommensarten und Belastungen von Haushalten müssen in einer detaillierten und verläßlichen Weise erfaßt sein und dann mit einer oder mehreren Armutsdefinitionen verglichen werden. Es sind große Stichproben erforderlich, die die gesamte Bevölkerung repräsentativ abbilden und die präzise Aussagen auch über kleine Bevölkerungsgruppen ermöglichen. Unter Umständen müssen Datensätze gebildet werden, die Informationen aus verschiedenen Erhebungen integrieren, wie dieses in den integrierten Mikrodatenfiles (MDF) aus EVS und Mikrozensus des SPES-Projektes und des Sonderforschungsbereiches (Sfb 3) geschehen ist.[1]) Soll das Vorliegen einer Armutssituation anstatt auf Basis des verfügbaren Einkommens gar unter Berücksichtigung der realen Lebensbedingungen erfaßt werden (z. B. Ernährungszustand, Wohnungssituation, Gesundheitszustand, Ausbildungsstand, Ausmaß der Integration), so muß eine größere Zahl von sozialen Indikatoren herangezogen werden, die für jede Person noch viel größere Informationsanforderungen stellen.

Das Armutsbeispiel ist kein Einzelfall. Vor kurzem wurde das umfassende System von Sozialindikatoren, das Wolfgang Zapf als SPES-Indikatorensystem für die Bundesrepublik

[1]) Vgl. beispielsweise Kortmann (1982)

zum ersten Mal veröffentlicht hat[2]), fortgeschrieben, um zu sehen, welche Veränderungen in einem breit verstandenen Konzept gesellschaftlicher Wohlfahrt sich im Jahrzehnt nach der Ölkrise und bei den verstärkt einsetzenden technologischen Veränderungen vollzogen haben.[3]) Bei dieser Fortschreibung fehlen eine Reihe wichtiger Indikatoren. Sie konnten vor 10 Jahren noch berechnet werden, weil damals die Wissenschaft noch Zugang zu aktuellen Mikrodaten der EVS und des Mikrozensus hatte. In der Zwischenzeit war dieses nicht mehr der Fall. Die angeführten Beispiele — und viele weitere könnten genannt werden[4]) — verdeutlichen, welche Kosten mit einer Datenpolitik verbunden sind, die die Wissenschaft von einer ihren Erfordernissen gerecht werdenden Nutzung von Daten, die mit öffentlichen Mitteln gesammelt werden, ausschließt. Wenn wir es richtig verstehen, ist es ein Ziel dieser Veranstaltung, Wege zu suchen, die eine Entwicklung zum Besseren bringen können.

Tabelle 1: Absolute Armut und Armutsquote im Zeitvergleich*)

Gegenstand der Nachweisung	1963	1969	1973
(1) Anzahl der Personen in deutschen Haushalten, die in verdeckter Armut leben			
in 1000	1647	567	832
in % der Bevölkerung	2,8	1,1	1,4
(2) Anzahl der Personen, die laufende Hilfe zum Lebensunterhalt außerhalb von Anstalten empfingen			
in 1000	798	707	861
in % der Bevölkerung	1,4	1,2	1,4
(3) Summe (1) + (2) in 1000	2445	1294	1693
(4) Ausschöpfungsgrad (2) in % von (3)	33	55	51

*) Originaltabelle: Berechnungen der Arbeitsgruppe Armutsforschung auf der Basis der amtlichen Sozialhilfestatistik und auf der Grundlage der EVS 1962/63, 1969 und 1973 (R. Hauser u. a. (1981), S. 73).
Quelle: Bolte/Hradil, Soziale Ungleichheit in der Bundesrepublik Deutschland (1984, S. 142).

Tabelle 2: Relative Armut und Armutsquoten im Zeitvergleich*)

Gegenstand der Nachweisung	40%-Grenze			60%-Grenze		
	1963	1969	1973	1963	1969	1973
Personen in relativer Armut						
in 1000	3018	1582	1507	15331	11516	11135
in % der Bevölkerung	5,3	2,8	2,6	26,9	20,6	19,5
Haushalte in relativer Armut						
in 1000	860	515	480	4364	3520	3492
in % aller privaten Haushalte	4,3	2,5	2,3	22,1	17,1	16,5

*) Originaltabelle: Berechnungen der Arbeitsgruppe Armutsforschung auf der Grundlage der EVS 1962/63, 1969 und 1973 (R. Häuser (1981), S. 118).
Quelle: Bolte/Hradil, Soziale Ungleichheit in der Bundesrepublik Deutschland (1984, S. 143).

[2]) Vgl. Zapf (1977).
[3]) Vgl. Diewald (1984).
[4]) Beispielsweise konnten differenzierte Einkommens- und Umverteilungsanalysen für die Bundesrepublik Deutschland bisher nicht über das Jahr 1969 hinaus fortgeführt werden. Vgl. die Beiträge in Krupp und Glatzer (1978) und Stolz (1983). Auch die differenzierte Analyse der personellen Vermögensverteilung bricht mit dem Jahr 1973 ab; vgl. Mierheim und Wicke (1978).

Wir haben die Aufgabe übernommen, den Datenbedarf der Wissenschaft insbesondere im Hinblick auf den Zugang zu anonymisierten Einzelangaben darzustellen. Im Sinne des Themas geht es hier um die Wissenschaft, die außerhalb von Statistischen Ämtern und sonstigen datenproduzierenden staatlichen Stellen tätig ist. Für die in Statistischen Ämtern tätigen Wissenschaftler stellen sich die Probleme des Datenzugangs in anderer Form, die hier nicht zu behandeln ist. Wir wollen im Rahmen des gestellten Themas zwei Hauptfragen behandeln:

— Braucht die Wissenschaft überhaupt Daten der amtlichen Statistik oder sollte sie sich nicht besser auf die Analyse von Daten begrenzen, deren Generierung sie selbst kontrolliert? Und wenn sie Daten der amtlichen Statistik benötigt, weshalb benötigt sie diese zunehmend als anonymisierte Einzelangaben?

— Wie jeder weiß, ist das Hauptproblem bei der Weitergabe von Einzeldaten die Festlegung von Kriterien der Anonymisierung, die sicherstellen, daß anonymisierte Daten gegen das Risiko der Deanonymisierung hinreichend geschützt sind. Wir diskutieren einige Grundsätze, die u. E. bei der Lösung der zum Teil konfligierenden Interessen von Wissenschaft und Datenschutz zu berücksichtigen sind und entlang deren eine beide Gesichtspunkte befriedigende Lösung auch möglich erscheint.[5])

1 Benötigt die Wissenschaft (weiterhin) Daten der amtlichen Statistik und weshalb zunehmend als Einzelangaben?

Der erste Teil dieser Frage ist fast eine rhetorische Frage. Man kann ohne Einschränkung sagen: Die Daten der amtlichen Statistik sind eine unverzichtbare Ressource für die Forschung in den gesellschaftswissenschaftlichen Disziplinen, von der Ökonomie über die Soziologie, die Demographie, die Politikwissenschaft, die Sozialgeographie bis hin zur Geschichtswissenschaft und eingeschlossen eine Reihe interdisziplinär ausgerichteter Forschungsrichtungen, wie etwa die Regional- und Raumforschung oder die Bildungsforschung. Diese Disziplinen haben im einzelnen Vorstellungen darüber, in welcher Weise angesichts der Entwicklung der Fragestellungen und der Forschungsmethoden sowie der gesellschaftlichen Wandlungsprozesse auch die amtliche Erhebung und Bereitstellung von Daten Veränderungen erfahren müßte.[6]) Darauf wollen wir hier nicht im einzelnen eingehen. Jede dieser Disziplinen wird die Schwerpunkte auch etwas anders setzen, aber für alle gilt, daß Daten der amtlichen Statistik sowohl für die in diesen Disziplinen zentralen analytischen Fragestellungen als auch für ihre Beiträge, die sie aus ihren Erkenntnissen für planvolles Handeln und rationale politische Entscheidungen liefern können, nach wie vor eine wesentliche Basis bilden. Da dieses kaum umstritten ist, können wir uns auf einige wenige Illustrationen für jene Disziplinen begrenzen, denen wir selbst angehören.

[5]) Zu dieser Diskussion vgl. Kaase u. a. (1980), Flaherty (1979), Grohmann u. a. (1980).

[6]) Als Beispiel für die spezifischen Datenwünsche der Bevölkerungswissenschaft vgl. Heilig (1985); für die Soziologie vgl. Zapf (1974, 1985), Müller (1982); für die Ökonomie Krupp (1975, 1982). Für viele Fragestellungen der sozialwissenschaftlichen Forschung ist es ein besonderer Mangel, daß einige Großstichproben, insbesondere solche auf freiwilliger Basis, bestimmte schwer erfaßbare Gruppen, wie Ausländer, Anstaltsbevölkerung, Personen mit hohem Einkommen, Nicht-Seßhafte und Obdachlose, nicht repräsentativ oder überhaupt nicht erfassen, so daß häufig verallgemeinerte Informationen in Umlauf kommen, die sich aber in Wirklichkeit nur auf Teilgruppen beziehen.

Dabei wollen wir vier Arten von amtlichen statistischen Informationen unterscheiden:[7])

(1) Statistische Daten, wie sie in aggregierter Form in Tabellen laufend veröffentlicht werden.

(2) Statistische Daten, die außerhalb der Standardaufbereitungsprogramme des Statistischen Bundesamtes auf Anforderung in eigens berechneten Sondertabellen in aggregierter Form zur Verfügung gestellt werden.

(3) Anonymisierte Einzelangaben über Personen, Haushalte, Betriebe, Unternehmen, Organisationen und öffentliche Körperschaften.

(4) Nicht-anonymisierte Einzelangaben über die erwähnten Einheiten.

Über den zweifelsohne bestehenden Bedarf an routinemäßig veröffentlichten Daten ist es nicht nötig, hier zu sprechen. Ebensowenig möchten wir aus unserer Sicht etwas zum Bedarf an nicht-anonymisierten Einzelangaben aus dem Bereich der amtlichen Statistik sagen. Sie spielen zumindest in den Sozial- und Wirtschaftswissenschaften, in denen regelmäßig nur Hypothesen und Aussagen über Gruppen von Bezugseinheiten, allerdings auf der Basis von flexibel aggregierten Einzelangaben, angestrebt werden, keine wesentliche Rolle.

Wir werden deshalb vor allem auf den Bedarf an anonymisierten Einzelangaben und auf den Bedarf an speziell aggregierten Daten in Tabellenform außerhalb der amtlichen Standardauswertungen eingehen:

— In den vielfältigen Fragen der Entwicklung der Sozialstruktur und der Analyse des sozialen Wandels muß immer wieder auf Daten in Form von Einzelangaben, die dann zweckentsprechend zu aggregieren sind, zurückgegriffen werden. Das bereits zitierte SPES-Indikatorentableau bildet sozialen Wandel auf der Basis von 200 Indikatoren ab, von denen fast alle auf Daten der amtlichen Statistik basieren und die eine große Palette gesellschaftlicher Lebensbereiche repräsentieren: zur Entwicklung der Bevölkerung und Familienstruktur, zur Situation auf dem Arbeitsmarkt, zu den Veränderungen in den Arbeitsbedingungen, zum Wandel der Berufsstruktur und des Bildungssystems, zu den Einkommensverhältnissen, zur Vermögensverteilung, zu sozialer Ungleichheit und sozialer Mobilität, zur Versorgung der Haushalte mit Konsumgütern und zu ihren Wohnungsverhältnissen, zur Entwicklung der Gesundheit der Bevölkerung und anderes mehr. Der Bundesminister für Forschung und Technologie hat — um einen anderen Bereich zu nennen — im Rahmen eines großen Forschungsprogramms zur Technikfolgenabschätzung gerade in jüngster Zeit Projekte ausgeschrieben, in denen die noch nicht ausgeschöpften Möglichkeiten der Daten der amtlichen Statistik hierzu besser genutzt werden sollen.

  Viele dieser Fragestellungen erfordern die gleichzeitige Berücksichtigung vieler Einzelinformationen über einzelne Individuen oder über mehrere Individuen, die untereinander in einer sozialen Beziehung stehen, z. B. einen gemeinsamen Haushalt bilden. Dies ist immer der Fall, wenn komplexe Indikatoren gebildet werden müssen. Die Armutsdefi-

[7]) Vgl. dazu Bürgin und Reimann (1982).

nition ist ein Beispiel dafür. Viele andere lassen sich leicht anführen. In der Ungleichheitsforschung z. B. spielt das Konzept der Statusinkonsistenz eine große Rolle. Um zu klären, ob ein Individuum in einer Situation von Statusinkonsistenz lebt, müssen mehrere Einzelindikatoren, wie Ausbildung, Beruf, Stellung im Beruf, Einkommen und möglicherweise auch noch die gleichen Indikatoren für seinen Partner, in einer komplexen Formel miteinander verrechnet werden. Ähnliches gilt für das Konzept der relativen Deprivation in der Zufriedenheitsforschung oder für das Konzept des Lebensstils in der Freizeitforschung. Für die Bildung aussagefähiger Haushaltstypologien oder Familientypologien wird die Notwendigkeit des Rückgriffs auf eine Mehrzahl von Einzelangaben von Individuen um so dringender, je vielfältiger die empirischen Erscheinungsformen des sozialen Zusammenlebens werden und je weniger Aussagekraft und Verbindlichkeit Begriffe wie Familienvorstand oder Haushaltsvorstand haben.

Ein herausragendes Beispiel für den großen Ertrag, den die Weitergabe von Einzelangaben an die Wissenschaft erbringen kann, sind die Ergebnisse, die aus der Zusatzerhebung zum Mikrozensus 1971 „Berufliche und soziale Umschichtung der Bevölkerung" erzielt wurden. Dieser Datensatz wurde 1975 in anonymisierter Form dem SPES-Projekt für eigene Sekundäranalysen zur Verfügung gestellt. Seither sind auf der Basis dieser Daten eine Vielzahl von Untersuchungen erschienen, die das Wissen zu verschiedenen Fragen der Sozialstrukturforschung wesentlich bereichert haben, z. B. zur Entwicklung der Klassenlagen, zur Frage der Konsequenzen der Bildungsexpansion für die Egalisierung sozialer Chancen, zur Struktur sozialer Mobilität, zur beruflichen Ungleichheit zwischen Männern und Frauen, zur Integration der Vertriebenen und Flüchtlinge oder zu regionalen Disparitäten in der Bundesrepublik.[8]) Alle diese Studien wären ohne den problemlosen Zugang zu den Einzelangaben undenkbar gewesen.

— Ein großer Bedarf liegt weiter in den Forschungsbereichen vor, in denen es um die Quantifizierung von Kosten, Folgen und Nebenfolgen wirtschafts- und sozialpolitischer Maßnahmen geht. Zur Analyse derartiger Probleme wurden in den siebziger Jahren sehr differenzierte Mikrosimulationsmodelle entwickelt, die — ausgehend von einer Stichprobe von Bezugseinheiten — diese im Zeitverlauf unter Berücksichtigung demographischer und gesamtwirtschaftlicher Entwicklungen und unter Einbeziehung der relevanten institutionellen Regelungen im Steuer- und Transferbereich fortschreiben.[9]) Einen besonderen Schwerpunkt bilden dabei Untersuchungen über alternative sozialpolitische und steuerliche Regelungen, etwa ein geändertes Rentensystem, eine andere BAFÖG- oder Kindergeldregelung oder über einen geänderten Steuertarif[10]), die nur mit

[8]) Im Ergebnis sind aus diesem Datensatz bislang sicher über 50 wissenschaftliche Arbeiten entstanden, die hier nicht alle aufgeführt werden können, vgl. dazu u. a. Handl, Mayer und Müller (1977), Müller und Mayer (1976), Mayer (1979), Müller, Willms und Handl (1983), Handl (1984), Lüttinger (1986), Haller (1983), König und Müller (1986), Mammey und Schwartz (1982).

[9]) Vgl. für einen Überblick über diese Forschungsrichtung in mehreren Ländern Orcutt, Merz und Quinke (1986).

[10]) Eine weitreichende Anwendung von Mikrosimulationsmodellen im Zusammenhang mit Reformvorschlägen für eine Rentenreform in der Bundesrepublik zeigt die Untersuchung von Krupp, Galler, Grohmann, Hauser und Wagner (1981). Eine Anwendung eines Mikrosimulationsmodells auf demographische Fragen findet sich beispielsweise in Steger (1980). Beispiele für die Anwendung von Mikrosimulationsmodellen zur Analyse von Steuertarifen, Regelungen zur Ausbildungsförderung und zur Gewährung von Wohngeld bieten die Beiträge von Dick und Bungers/Quinke in Orcutt, Merz und Quinke (1986).

diesem Instrument in ihren differenzierten Auswirkungen auf einzelne Bevölkerungsgruppen analysierbar werden. Auch die Analyse der Umverteilungswirkungen des Steuer- und Transfersystems erfordert eine mikroanalytische Vorgehensweise, weil bei Gruppenanalysen eine Vielzahl von Effekten sich gegenseitig kompensieren und damit verschwinden.[11])

Wenngleich diese Mikrosimulationsmodelle bisher vor allem auf den Haushaltssektor ausgerichtet sind, so gibt es doch bereits Versuche, solche Modelle auf den Unternehmenssektor auszuweiten und auch zur Repräsentation diese Sektors für eine Stichprobe von Unternehmen die Konsequenzen ihrer jeweiligen Entscheidungen auf den Arbeits-, Finanz-, Beschaffungs- und Absatzmärkten fortzuschreiben.[12]) Auch hier sind für sämtliche in die Mikrosimulation einbezogenen Unternehmen Einzelangaben zu allen relevanten abhängigen und unabhängigen Variablen erforderlich.

— Schließlich gewinnen Longitudinalanalysen — seien es individuelle Verläufe von Erwerbs- und Mobilitätskarrieren oder von Haushalts- und Familienbiographien, Verläufe von Anspruchsakkumulationen im Bereich der Altersvorsorge oder auch das schrittweise Absinken in Armut, Gesundheitsverläufe, Unternehmensentwicklungen u. ä. — in neuerer Zeit immer größeres Gewicht, weil nur auf diese Weise eine Ursachenanalyse für Verlaufstypen möglich wird.[13]) Für derartige Untersuchungen sind Einzelangaben über im Zeitverlauf identische Bezugseinheiten für möglichst viele Zeitpunkte unabdingbar, wobei wiederum anonymisierte Angaben genügen, sofern sie sich nur für identische Individuen als Zeitreihe verknüpfen lassen.

Daten der amtlichen Statistik sind dann besonders wichtig, wenn Aussagen über differenzierte Bevölkerungsgruppen gefordert sind: die Entwicklung einzelner Berufe zum Beispiel, die in den mit vergleichsweise kleinen Stichproben arbeitenden Umfragen der Sozialwissenschaften nicht in einem Umfang erfaßt sein können, daß sich über solche Gruppen verläßliche Aussagen ableiten lassen. Das gleiche gilt für regional gegliederte Untersuchungen oder für Analysen, die sich auf spezifische Problemgruppen beziehen.

Daten der amtlichen Statistik sind schließlich unverzichtbar, um Ergebnisse, die die Sozialwissenschaften aus eigenen, kleineren Stichproben gewinnen, auf die Gesamtbevölkerung hochrechnen zu können oder um überhaupt erst sinnvolle Stichprobenpläne konstruieren zu können oder um Stichproben, die, aus welchen Gründen auch immer, verzerrt sind, adäquat gewichten zu können.

Fast keine dieser Forschungsfragen läßt sich aus den Standardveröffentlichungen der amtlichen Statistik beantworten. Nur in wenigen Fällen, wie etwa bei der Ermittlung von Gewichtungsfaktoren, ist die Bestellung von Sondertabellen, die in ihrer Endform keine Einzelanga-

[11]) Vgl. Stolz (1983) sowie Hauser/Heldmann (1981).
[12]) Ein derartiges Mikrosimulationsmodell wird zur Zeit in Schweden entwickelt. Vgl. Eliasson (1986).
[13]) Vgl. Wagner (1986), Abschn. 2.1.2 und 2.1.5; Göbel (1983); sowie Schmähl (1983). Herausragende Projekte, in denen sozialwissenschaftliche Längsschnittdaten erhoben und analysiert werden, sind das Projekt „Lebensverläufe und Wohlfahrtsentwicklung" von Karl Ulrich Mayer (vgl. u. a. Mayer und Papastefanou (1983), Mayer (1985), Blossfeld (1985)) und das sozio-ökonomische Panel (vgl. Sonderforschungsbereich 3, 1985).

ben enthalten, ein gangbarer Weg. Das Problem dabei sind nicht nur die erheblichen Kosten, sondern der oft recht langwierige Prozeß, insbesondere dann, wenn sich bei fortschreitender Untersuchung zeigt, daß Klassifizierungen geändert und Neuberechnungen angefordert werden müssen. Verfügt der Wissenschaftler dagegen über die anonymisierten Einzelangaben, kann er die erforderlichen Tabellen und Revisionen selbst schnell erstellen.[14])

Die meisten der angeführten Forschungsfragen setzen aber den direkten Zugang zu anonymisierten Einzelangaben voraus. Dieser Bedarf, insbesondere auch aus dem Bereich der amtlichen Statistik, läßt sich noch durch einige weitere Argumente verdeutlichen.

- Einzelangaben über Individuen und Haushalte können häufig für komplexe wissenschaftliche Analysen nicht in der Form verwendet werden, wie sie als Endprodukte der amtlichen Statistik zur Verfügung stehen. Erstens kann es erforderlich sein, Dateien, die sich nur auf Teile der Bevölkerung beziehen, durch andere Dateien in vielfältig modifizierter Form zu ergänzen, um repräsentative Aussagen für die gesamte Bevölkerung ableiten zu können. Zweitens kann es erforderlich sein, aus den vorhandenen Daten komplexe Zusatzvariablen zu kreieren oder mit Hilfe differenzierter Bereinigungsverfahren Inkonsistenzen in den auf eine Bezugseinheit bezogenen Daten aufzudecken und zu berichtigen. Drittens kann die Notwendigkeit bestehen, durch statistische Datenverknüpfung synthetische Datensätze mit ausgeweitetem Variablensatz zu schaffen.

- Zunehmend gewinnen auch internationale Vergleiche auf der Basis von Einzeldatensätzen an Gewicht. Typischerweise stimmen dabei die nationalen Datensätze in den Abgrenzungen der Bezugseinheiten (z. B. Haushaltsbegriff), der Variablen u. ä. nicht überein, so daß zur Sicherstellung der Vergleichbarkeit von Ergebnissen gegebenenfalls Umrechnungen und Reklassifizierungen erforderlich werden, die sich wiederum nur auf der Basis von Einzelangaben durchführen lassen.[15])

- Die Methoden der Datenanalyse haben sich in den letzten Jahrzehnten eindeutig von der Aggregatdatenanalyse weg hin zur Individualdatenanalyse entwickelt. Konnte man für eine Weile den Eindruck haben, daß über Verfahren der log-linearen Analyse die Analyse von Aggregatdaten eine Wiederbelebung erfahren würde, so setzen die neueren Entwicklungen gerade auch der log-linearen Analyse wiederum Individualdaten voraus. In den neueren Verfahren der Analyse zeitbezogener Daten orientiert sich gegenwärtig selbst die Demographie, die mit ihren Bevölkerungs- und Sterbetafeln eine der Domänen der Aggregatdatenanalyse bildete, sehr stark in Richtung Individualdatenanalyse um. Daß in den Simulationsverfahren die Individualsimulation neben der Gruppensimulation ein zunehmend stärkeres Gewicht bekommt, ist ebenso offensichtlich. Multivariate Analysen auf der Basis von Einzelfällen werden schlicht mehr und mehr zum unverzichtbaren Instrument.

[14]) Gerade das hierin liegende Zeitelement für die Planung vieler wissenschaftlicher Vorhaben sollte keinesfalls unterschätzt werden. Es sei nur an die durch die Drittmittelgeber festgelegte Projektlaufzeit und an die Problematik der befristeten Verträge für die meisten wissenschaftlichen Mitarbeiter erinnert.

[15]) Vgl. die noch nicht veröffentlichten Arbeiten des LIS-Projekts (Luxemburg Income Study). Einige Ergebnisse sind vorveröffentlicht in: Hauser und Fischer (1985).

— Je mehr die Forschung in komplexe multivariate Modellbildung eintritt, um so mehr vollzieht sie sich in einem interaktiven Prozeß zwischen forschungsleitender Theorie und empirischer Beobachtung. Der Forscher prüft in der Regel eine Mehrzahl alternativer Modelle auf den Grad ihrer Übereinstimmung mit der empirischen Wirklichkeit, und er kann nicht im vorhinein angeben, wie die Daten für das letztendlich (vorläufig) akzeptierte Modell strukturiert sein müssen. Wenn er auch im Endergebnis ein Modell veröffentlicht, das vielleicht nur wenige Parameter enthält, ist im Forschungsprozeß immer wieder der Rückgriff auf die Einzelangaben der individuellen Fälle erforderlich.

Dieser Stand der Forschung macht Vorgehensweisen unpraktikabel, in denen der Wissenschaftler am Schreibtisch Analysen konzipiert, die er dann einer anderen Institution, beispielsweise einem Statistischen Amt, zur Ausführung überträgt. Auch der Rückgriff auf Datenbanken hilft nicht weiter, wenn in diesen Datenbanken die Informationen nicht als Einzelangaben, sondern in einer bereits aggregierten Form gespeichert sind.

Für die Zwecke der wissenschaftlichen Analyse in unseren Disziplinen ist dabei aber die Identifikation des Individuums in der Regel nicht erforderlich. Es interessiert nur als nicht bekannter Merkmalsträger.

Der Forscher muß also selbst Zugang zu Mikrodaten — auch aus der amtlichen Statistik — haben, bei denen kein Personenbezug erforderlich ist und die entsprechend anonymisiert sein können. Dieser Sachverhalt ist, soweit wir sehen, heute nicht mehr bestritten. Das Gutachten der Gesellschaft für Mathematik und Datenverarbeitung (GMD), Sankt Augustin, geht davon aus. Im Statistikgesetz ist die Weitergabe von anonymisierten Einzelangaben vorgesehen. Das Urteil des Bundesverfassungsgerichtes zur Volkszählung, das in seinen Grundsätzen dem Schutz individueller Persönlichkeitsrechte einen außerordentlich hohen Rang einräumt, hält explizit fest, daß diese Grundsätze durch die Weitergabe anonymisierter Einzelangaben an die Wissenschaft nicht beeinträchtigt sind.

2 Anonymisierung von Einzelangaben

Wenn über das Grundsätzliche Einigkeit besteht, dann geht es vor allem darum, Lösungen für die praktische Frage der Anonymisierung zu finden. Eine gegen alle Anschläge absolut sichere Anonymisierung gibt es praktisch nur für Daten, deren Informationsgehalt sehr gering ist. Deshalb geht der Entwurf des neuen Bundesstatistikgesetzes (BStatG) erfreulicherweise auch davon aus, daß Einezlangaben für wissenschaftliche Zwecke dann übermittelt werden dürfen, „wenn sie nur mit einem unverhältnismäßig hohen Aufwand an Zeit, Kosten und Arbeitskraft (Auskunftpflichtigen oder Betroffenen) zugeordnet werden können".[16] Im Vergleich zu früheren Formulierungen erleichtert diese Vorschrift einer faktischen Anonymisierung eine Weitergabe, weil sie das Vorhandensein eines Restrisikos zuläßt. Sie etabliert ein Nutzungsprivileg für die wissenschaftliche Forschung, indem sie für diese Nutzergruppe das Prinzip einer „ausreichenden" Anonymisierung[17] bestätigt. Ausreichende Anonymisie-

[16] Vgl. Entwurf des Bundesstatistikgesetzes 1986, § 16 (4).
[17] Vgl. Tuner (1982)

rung meint einen im Hinblick auf das Kosten-Nutzen-Verhältnis eines Datenangriffs genügenden Schutz vor einem Reidentifikationsversuch.

In § 16 (6) (BStatG) wird jedoch eine restriktive Regelung der Nutzungsberechtigung vorgeschrieben. „Bei den Empfängern muß durch organisatorische Maßnahmen sichergestellt sein, daß nur Amtsträger oder für den öffentlichen Dienst besonders Verpflichtete Kenntnis von den Einzelangaben erhalten." Wenn damit gemeint ist, daß Doktoranden oder Habilitanden, deren wissenschaftliche Arbeiten beispielsweise durch Stipendien finanziert werden, von der Nutzung ausgeschlossen sein sollten, wäre dies eine harte Restriktion; denn die wissenschaftliche Forschung lebt in hohem Maße gerade auch von Dissertationen und Habilitationsschriften. Der Forschungsprozeß würde auch wesentlich behindert, wenn studentische Hilfskräfte von der Beteiligung an Analysearbeiten ausgeschlossen würden.

In der Annahme, daß dieser Entwurf in seinen wesentlichen Formulierungen auch Gesetz wird, möchten wir deshalb im folgenden eher grundsätzlich das Verhältnis von Datenbedarf und Anonymisierungsstrategien diskutieren.

— Der Datenbedarf der Wissenschaft ist prinzipiell umfassend, d. h. die Wissenschaft muß im Rahmen der Grenzen forschungsethischer Maßstäbe, für die sie allerdings ebenfalls weitgehend selbst verantwortlich ist, die Möglichkeit zur freien Forschung in allen Bereichen haben, die ihr zum Erkenntnis- und Wissensgewinn wichtig erscheinen. Wenn es dabei zu Zielkonflikten mit dem Datenschutz[18]) kommt, so müssen die Güter der Freiheit der wissenschaftlichen Forschung und des Schutzes der Privatsphäre der Persönlichkeit gegeneinander in einer Weise abgewogen werden, in der auch der Gesichtspunkt der Verhältnismäßigkeit zur Geltung kommt. Der Forschung können nicht Restriktionen auferlegt werden, die aus einer unverhältnismäßigen Überschätzung der durch sie entstehenden Risiken resultieren.[19])

  In diesem Zusammenhang ist es wichtig, darauf hinzuweisen, daß von den im AIMIPH-Gutachten[20]) geprüften Szenarien im Grunde genommen nur das Adressenverlags-Szenario Annahmen über das Zusatzwissen enthält, das für den wissenschaftlichen Bereich leicht zugänglich ist. Da unter diesen Voraussetzungen die Reidentifikationsmöglichkeiten praktisch gleich Null sind, unterstützen die Ergebnisse des Gutachtens grundsätzlich die Datenweitergabe an die Wissenschaft. Die Befunde eines höheren Reidentifikationsrisikos bei den anderen Szenarien bedeuten nur, daß in der Datenweitergabe getrennt werden muß zwischen der Weitergabe an die Wissenschaft und der

[18]) Vgl. Bull und Dammann (1982).

[19]) Es ist bekannt, daß aufgrund der spezifischen Interessenlage der Wissenschaft bei ihr die Gefahren von Reidentifikationsversuchen äußerst gering sind und bislang auf der ganzen Welt praktisch keine Mißbräuche durch die Wissenschaft vorgekommen sind, die zu einer Beeinträchtigung von Betroffenen geführt hätten. Andererseits leben wir in einer Welt, in der wir überall Restrisiken akzeptieren müssen, denn sonst dürfte es keine Raumfahrt, keine Atomkraftwerke, keine Strahlenforschung, keine Genforschung, ja wohl die meisten Arbeitsplätze nicht geben. Dieser Hinweis soll nicht die Anliegen des Datenschutzes bagatellisieren. Güterabwägung und Befolgung des Grundsatzes der Verhältnismäßigkeit bedeuten aber, sowohl die Reidentifikationswahrscheinlichkeiten von anonymisierten Daten in der wissenschaftlichen Forschung als auch die daraus im schlimmsten aller Fälle eintretenden tatsächlichen Gefährdungen der persönlichen Integrität der Betroffenen richtig einzuschätzen und einen angemessenen Ausgleich herzustellen.

[20]) Vgl. Paaß und Wauschkuhn (1985).

Weitergabe an Institutionen, bei denen mehr Zusatzwissen und höheres Reidentifikationsinteresse angenommen werden müssen. Sie bedeuten auch, daß Sicherungen dafür erforderlich sind, daß Daten, die an die Wissenschaft weitergegeben wurden, den wissenschaftlichen Bereich nicht verlassen. Die Ergebnisse des Gutachtens und die Formulierung des Gesetzestextes lassen die Einrichtung von Public Use Samples, zu denen jeder Interessierte Zugang hat, für die Bundesrepublik als kaum realisierbar erscheinen. Sie unterstützen aber übereinstimmend die Einrichtung von Scientific Use Samples, also von „Wissenschaftsstichproben", d. h. die Weitergabe von Stichproben anonymisierter Einzeldatensätze an die Wissenschaft.

— Wer versucht, im Hinblick auf die Anonymisierungsproblematik die Anforderungen der Wissenschaft und des Datenschutzes miteinander in Einklang zu bringen, kommt schnell zu der Erkenntnis, daß der Datenbedarf der Wissenschaft zu unterschiedlich ist, als daß einheitliche Anonymisierungsregeln eine befriedigende Lösung für die verschiedenen Nutzungserfordernisse bringen könnten. Für unterschiedliche Datenbedarfe müssen unterschiedliche Verfahren des Schutzes vor Reidentifikation gefunden werden. Neben Anonymisierungsvorkehrungen auf der Datenseite müssen dabei auch organisatorische Schutzvorkehrungen zur Verringerung des Reidentifikationsrisikos bei den datenhaltenden wissenschaftlichen Institutionen genutzt werden.

Allerdings wäre es wenig sinnvoll, eine uferlose Sonderbehandlung jedes Einzelfalls anzustreben. Dies würde der Willkür Tür und Tor öffnen und wegen Unübersichtlichkeit sowohl die Datensicherheit als auch die Chancengleichheit im Datenzugang verringern. Unseres Erachtens ist dies auch nicht erforderlich. Denn im wesentlichen läßt sich die Notwendigkeit unterschiedlicher Anonymisierungsstrategien und Schutzvorkehrungen auf einen Faktor zurückführen.

Die Reidentifikationschance ist ceteris paribus um so größer, je regional eingegrenzter die Population ist, der ein Individuum zugehört. Eine Gefährdung stellen vor allem kleinräumige Regionalangaben dar. Es scheint uns deshalb sinnvoll, Anonymisierungsstrategien und organisatorische Schutzmaßnahmen vor allem in Abhängigkeit von der regionalen Tiefengliederung zu diskutieren, in der Einzelangaben erforderlich sind.

In den amerikanischen Public Use Samples hat sich als Kriterium für Public Use Samples ohne besondere organisatorische Schutzvorkehrungen etabliert, daß in solchen Stichproben nur Gebietseinheiten mit einer Bevölkerungszahl von mindestens 100 000 Einwohnern als Merkmalsausprägungen ausgewiesen sind. Eine solche Grenze, unterhalb derer man ohne besondere Schutzvorkehrungen nicht auskommen wird, scheint mir auch auf deutsche Verhältnisse übertragbar.

Besondere Schutzvorkehrungen für regional tiefgegliederte Daten kann aber nicht heißen, daß in solchen Fällen die Weitergabe von Einzelangaben ausgeschlossen ist. Die Regionalforschung, die Stadtsoziologie oder die epidemiologische Forschung arbeiten mit den prinzipiell gleichen methodischen Instrumenten wie die übrige sozialwissenschaftliche Forschung. Wissenschaftler, die in diesen Bereichen arbeiten,

können deshalb auch bei Daten, die sich auf kleine Regionaleinheiten beziehen, nicht auf anonymisierte Einzelangaben verzichten. Für diesen Typ des Datenbedarfs ist es jedoch erforderlich, gesonderte Schutzvorkehrungen zu treffen. Das Ziehen von Substichproben ist oft nicht möglich, weil bei kleinräumigen Analysen die ursprüngliche Stichprobe selbst oft nur noch wenige Fälle enthält. Je nach dem Einzelfall wird man unterschiedliche Lösungen finden müssen, bei denen mehrere Strategien der Sicherheitserhöhung zu verbinden sind. Bevor jedoch Maßnahmen ergriffen werden, die die Analysemöglichkeiten einschränken, wie z. B. die Aggregierung von Untersuchungseinheiten, Aggregierungen von Merkmalsausprägungen oder die Bildung von Merkmalsscheiben mit nur wenigen Variablen, sollten organisatorische Maßnahmen zur Erhöhung des Datenschutzes ausgeschöpft werden.

Ähnliche Gesichtspunkte einer Sonderbehandlung wird man für Dateien zu berücksichtigen haben, die aus Einzelfällen bestehen und die aus anderen Gründen als denen der kleinräumigen Lokalisierung ein höheres Reidentifikationsrisiko aufweisen, sei es weil z. B. ihre Fallzahl in der Grundgesamtheit klein ist (z. B. bei einer Stichprobe von Großunternehmen) oder weil die Einzelfälle eine besondere Auffälligkeit zeigen (wie etwa bei einer Stichprobe von Obdachlosen).

Wir wollen diesen Typ von Dateien hier nicht weiter diskutieren, sondern uns auf die „Wissenschaftsstichproben" konzentrieren, bei denen solche Sonderrisiken nicht angenommen werden müssen. Für die Weitergabe dieser Daten scheint uns die Berücksichtigung der folgenden Gesichtspunkte wichtig.

— Im Vergleich zu den in der empirischen Sozialforschung üblichen Datensätzen zeichnen sich die Datensätze der amtlichen Statistik dadurch aus, daß sie thematisch sehr eingeschränkt sind und zu den einzelnen Themen nur kleine Variablenkataloge enthalten. Die bereits bei der Datenerhebung sehr eingegrenzte Wirklichkeitserfassung darf nicht dadurch noch verkürzt werden, daß die Datenweitergabe nur selektiv erfolgt. Insbesondere muß sichergestellt sein, daß der in den Datensätzen der Bevölkerungsstatistik wie dem Mikrozensus, der Volkszählung oder der EVS enthaltene Haushalts- oder Familienbezug in den übermittelten Daten erhalten bleibt. Mit anderen Worten: Die Wissenschaft muß die Möglichkeit des Zugangs zu den vollständigen Merkmalssätzen, d. h. zur Gesamtheit der Variablen eines Datensatzes, haben.

— Die Wissenschaft kann nicht mit Daten minderer Qualität arbeiten. Anonymisierungsverfahren, die dazu führen, daß das Analysepotential eingeschränkt wird, oder bei denen sogar mit der Möglichkeit falscher Ergebnisse gerechnet werden muß, können nicht in Betracht kommen. Es wäre eine absurde Vorstellung, daß der Wissenschaft Daten übermittelt werden, die aus Anonymisierungsgründen in einer Weise transformiert sind, daß daraus Falschaussagen resultieren können. Es genügt der Nachweis eines einzigen Fehlurteils, um ein Datentransformationsverfahren zu diskreditieren. Wir können nicht aufwendige Forschung betreiben, um unsere analytischen Instrumente zu verbessern, und dann in Kauf nehmen, daß durch Datenverfälschung unkontrollierbare Irrtümer in die Basis unserer Forschung Einzug halten. Auch beim Abwägen zwischen unpräzisen oder falschen wissenschaftlichen Aussagen und einer Verringerung des Reidentifikationsrisikos muß der Grundsatz der Verhältnismäßigkeit gelten.

— Unter den Strategien, das Reidentifikationsrisiko zu senken, ist für die Wissenschaft in der Regel das Ziehen von zufälligen Substichproben mit den geringsten Einschränkungen des Analysepotentials verbunden. Ausnahmen bilden Untersuchungen, die sich auf kleine Regionen oder sehr spezielle Subgruppen beziehen. Das Ziehen von zufälligen Substichproben ist eine vergleichsweise billige Technik. Bei den großen Stichproben der amtlichen Statistik wird die Aussagekraft nicht wesentlich eingeschränkt. Die Gefahr von systematischen Verzerrungen ist als minimal einzustufen. Die Vergrößerung des Zufallsfehlers kann über die Wahrscheinlichkeitstheorie präzise bestimmt werden. Gleichzeitig kann nach den Ergebnissen des AIMIPH-Gutachtens der GMD die Schutzwirkung als erheblich eingestuft werden. Das Ziehen von zufallsgesteuerten Stichproben versetzt den potentiellen Angreifer systematisch in Unkenntnis darüber, ob eine Person in der Stichprobe enthalten ist oder nicht. In allen Szenarien des AIMIPH-Gutachtens verringert Nichtwissen über die Stichprobenzugehörigkeit die Wahrscheinlichkeit einer korrekten Reidentifikation erheblich. Vor anderen, schädlicheren Anonymisierungsstrategien sollte deshalb u. E. zuerst die Möglichkeit des Ziehens von Substichproben genau geprüft werden. Allerdings wird dies abhängig sein von der Größe der Gesamtstichprobe, und die Analysemöglichkeiten dürfen natürlich auch nicht durch zu kleine Substichproben erheblich eingeschränkt werden.

— Der Nutzen von Daten der amtlichen Statistik für die Wissenschaft ist wesentlich bestimmt durch den Zeitpunkt, zu dem sie verfügbar werden. Zur Sicherung gleicher Chancen für die unabhängige wissenschaftliche Forschung ist es unabdingbar, daß die Daten zum frühestmöglichen Zeitpunkt zugänglich gemacht werden. In aller Regel dürfte dies der Zeitpunkt sein, zu dem die Aufbereitung der Daten soweit abgeschlossen ist, daß mit den Analysearbeiten begonnen werden kann.

— Die Wissenschaft muß Zugang zu möglichst allen Mikrodaten der amtlichen Statistik bekommen. Wären einzelne Daten für die wissenschaftliche Forschung nicht mehr relevant, wäre dies wahrscheinlich ein guter Indikator dafür, daß man auf ihre Erhebung verzichten könnte. Dieses Postulat reicht über die durch Erhebungen der statistischen Ämter gewonnenen Daten hinaus. Es schließt auch anonymisierte prozeßproduzierte Daten ein, wie beispielsweise Sozialversicherungsdaten oder Steuerdaten. Zumindest insoweit, als Ministerien im Rahmen von Forschungsaufträgen eigene Datenerhebungen durchführen, wird dieser Vorstellung bereits dadurch Rechnung getragen, daß diese Daten über eine Gemeinschaftseinrichtung der Wissenschaft, das Zentralarchiv für empirische Sozialforschung in Köln, den Forschern zur Verfügung stehen.

In diesem Zusammenhang muß auch ein in der Datenschutzdiskussion heißes Eisen angesprochen werden: Im Rahmen des durch die gesetzlichen Bestimmungen Möglichen müssen für Forschungszwecke auch Datenverknüpfungen praktizierbar werden. Man darf Datenverknüpfungen sicher nicht zur Routine werden lassen, und jeder Einzelfall muß geprüft werden. Am naheliegendsten sind Verknüpfungen dann, wenn die Daten schon nach der gesetzlichen Grundlage als Wiederholungsbefragung, wie beim Mikrozensus, angelegt sind. Ein großes Analysepotential geht der Forschung dadurch verloren, daß u. W. die Daten der verschiedenen Mikrozensuswellen nicht systematisch ver-

knüpft[21]) und damit als Paneldaten auswertbar werden. Als Beispiel für eine erfolgreiche Datenverknüpfung in der Wissenschaft kann auf die Lebenslageumfrage verwiesen werden, in der nach Zustimmung der Befragten die bei den Sozialversicherungsträgern über den einzelnen Befragten gespeicherten Daten mit seinen Interviewdaten sowie mit weiteren Daten aus einer Nachbefragung (Transferumfrage 1981 des Sfb3) verknüpft werden konnten.[22])

Die Wissenschaft muß auch Zugang zu den Einzelangaben solcher verknüpfter Datensätze haben. Dabei können die Verknüpfungen entweder bei den datenproduzierenden Institutionen oder mit Hilfe von Verknüpfungscodes durch den Wissenschaftler selbst vorgenommen werden, wobei natürlich eine Personenidentifizierung ausgeschlossen sein muß.

— Der Wissenschaftsrat hat in einer vor kurzem abgegebenen Stellungnahme[23]) empfohlen, innerhalb der zu gründenden „Gesellschaft sozialwissenschaftlicher Infrastruktureinrichtungen e.V." (GESIS) ein Zentrum für Mikrodaten einzurichten. Diese Empfehlung ist in der hohen Bedeutung begründet, die der Wissenschaftsrat in der Nutzung vorhandener Mikrodaten durch die Wissenschaft sieht. Sie resultiert aber auch aus der Erkenntnis, daß die adäquate Nutzung einen außerordentlich hohen Aufwand an EDV-Kapazität, Datenanalysekompetenz und Datenschutzvorkehrungen erfordert. Der Wissenschaftsrat geht deshalb von der Vorstellung aus, daß die Nutzung von Mikrodaten gefördert werden kann, wenn ein solches Zentrum neben eigenständiger inhaltlicher und methodischer Forschung auch Dienstleistungen für Wissenschaftler erbringt, die an ihren Institutionen nicht über die fachliche und technische Infrastruktur und die organisatorischen Möglichkeiten für einen angemessenen Datenschutz verfügen, um effizient mit Mikrodaten zu arbeiten. Ein solches Zentrum könnte als Gemeinschaftseinrichtung der Wissenschaft auch eine Treuhandfunktion übernehmen, insofern es unter besonders strenger Prüfung seiner Datenschutzvorkehrungen auch über Daten verfügen könnte, deren weitere Verbreitung als Wissenschaftsstichprobe nicht verantwortet werden kann. Um Untersuchungen über Entwicklungen in der Zeit zu erleichtern oder überhaupt erst zu ermöglichen, muß eine solche Gemeinschaftseinrichtung der Wissenschaft auch die Möglichkeit haben, die ihr übermittelten Daten über längere Zeit zu archivieren.

Es ist sehr zu begrüßen, daß das Statistische Bundesamt in jüngster Zeit seine Zurückhaltung in der Weitergabe anonymisierter Einzelangaben etwas gelockert hat.[24]) Daß aber die derzeitige Praxis in der Bundesrepublik dem begründeten Bedarf der Wissenschaft

[21]) Hervorhebenswerte Ausnahmen stellen die Untersuchungen von Mayer (1983) zu Längsschnittanalysen des Mikrozensus und die entsprechenden Untersuchungen von Gerhard und Stärk-Rötters (1985) aus der Bildungsstatistik dar.
[22]) Vgl. Stubig und Berntsen (1985).
[23]) Vgl. Wissenschaftsrat (1986).
[24]) Vgl. dazu den Beitrag von Manfred Euler in diesem Band (siehe S. 157ff.). Ein wegweisendes Beispiel für die verstärkte Kooperation zwischen amtlicher Statistik und universitärer Forschung ist die gemeinsame Publikation des Datenreportes 1985 durch das Statistische Bundesamt (1985) und den Sonderforschungsbereich 3 der Universitäten Frankfurt und Mannheim.

nicht gerecht werden kann, geht auch aus dem Vergleich mit anderen Ländern hervor. Die von der GMD in Auftrag gegebenen Untersuchungen machen dies eindrucksvoll deutlich.[25]) Auch die in diesem Band abgedruckten Vorträge über die Praxis in den USA und Großbritannien haben gezeigt, daß die Beeinträchtigungen, die die wissenschaftliche Forschung in der Bundesrepublik hinnehmen muß, ganz erheblich sind.

Bei bestimmten Untersuchungen ist es viel leichter, mit Daten aus dem Ausland zu arbeiten als mit Daten aus der Bundesrepublik. Wir verfügen in Mannheim über Mikrozensen oder Mikrozensus-ähnliche Datensätze aus den USA, England, Frankreich, Österreich, ja sogar aus Ungarn und Polen. Nur die entsprechenden Daten aus der Bundesrepublik fehlen.

Diese Situation erscheint um so paradoxer, als es gerade unter dem Gesichtspunkt der Akzeptanz von Datenerhebungen durch die Öffentlichkeit gewichtige Argumente dafür gibt, Daten in einer den heutigen Erfordernissen entsprechenden Weise an die Wissenschaft zu übermitteln. Die Bevölkerung ist um so eher zur Mitwirkung bei Erhebungen bereit, je mehr durch vielfältige Auswertungen deren Nutzen demonstriert werden kann. Die Wissenschaft wird sich anderen Daten zuwenden, wenn der Zugang zu amtlichen Mikrodaten nicht sicher absehbar ist und eine Weitergabe nicht kontinuierlich und zu erträglichen Kosten erfolgt.

Für die Wissenschaft ergibt sich um so stärker die Notwendigkeit, eigenständig Daten zu sammeln, je begrenzter die amtliche Datenproduktion ist und je weniger diese der Entwicklung in den Wissenschaften Rechnung trägt. Die Notwendigkeit eigenständiger wissenschaftlicher Datenproduktion wird aber auch dann dringlicher, wenn die durch die Wirtschaft oder die öffentliche Verwaltung gesammelten Daten der Wissenschaft nicht in einer Weise zur Verfügung stehen, wie sie diese für ihre Zwecke benötigt. Eine dem Bedarf der Wissenschaft gerecht werdende Regelung der Datenweitergabe dient damit nicht nur der Kostenersparnis. Wir meinen, sie liegt auch im Interesse der amtlichen Statistik, weil sie zur Vermeidung der Verdoppelung von Datenerhebungen beiträgt und damit den Bürger entlastet.

Eine verbesserte Datenweitergabe unterstützt auch die „Gewaltenteilung", insofern mehrere Instanzen an der Analyse der gleichen Daten arbeiten. Schließlich kann die Kooperation mit der Wissenschaft auch ein Schutz vor Pressionsversuchen durch die Politik sein. Und letztlich führen diese Argumente zu den wissenschaftstheoretischen Grundlagen der empirischen Wissenschaften und des Forschungsprozesses: Ein Spezifikum jeder Wissenschaft ist die gegenseitige Kontrolle einzelner Wissenschaftler untereinander, die nur stattfinden kann, wenn empirische Ergebnisse intersubjektiv nachvollziehbar präsentiert und Erstanalysen durch weitere Analysen des gleichen empirischen Datenmaterials überprüft werden. Es gibt viele Beispiele, daß derartige Folgeanalysen Fehler aufgedeckt haben, die — mit fatalen Folgen — unentdeckt geblieben wären, wenn mangels Verfügbarkeit der empirischen Ausgangsdaten solche Folgeanalysen unterblieben wären.[26])

[25]) Vgl. Majka/Wertheimer (1983) und Dalenius (1981).
[26]) Vgl. Wagner (1986), Abschnitt 4.3.

Es ist das Verdienst des Datenschutzes, daß das Bewußtsein für Gefährdungen, die aus den Fortschritten der Informationstechnologie resultieren, geschärft wurde. Gewiß gab es auch an manchen wissenschaftlichen Forschungseinrichtungen in dieser Beziehung einiges zu verbessern.[27]) Aber die Sozialwissenschaftler wissen seit langem, daß die Gewißheit der Bevölkerung, daß ihre Daten vertraulich behandelt werden und daß der Datenschutz in jeder Hinsicht gewahrt bleibt, eine conditio sine qua non der Forschung ist. Deshalb ist es selbstverständlich, daß aus Veröffentlichungen keinerlei Rückschlüsse auf einzelne Personen möglich sein dürfen, auch wenn der Forscher seine Analysen auf der Basis von Einzelangaben erstellt. Was die Bedeutsamkeit des Datenschutzes betrifft, stimmen wir dem Datenschutzbeauftragten des Bundes, Dr. Baumann, voll und ganz zu, wenn er in seinen Thesen zum 1. Wiesbadener Gespräch schreibt:

> These 2: „Das Bundesverfassungsgericht hat im Volkszählungsurteil . . . betont, daß statistische Erhebungen nur dann erfolgreich durchgeführt werden können, wenn der auskunftspflichtige Bürger das notwendige Vertrauen in die Wahrung seines informationellen Selbstbestimmungsrechtes hat."
>
> These 3: „Die strikte Einhaltung des Statistikgeheimnisses ist wesentliche Voraussetzung für die Bildung des notwendigen Vertrauens in der Bevölkerung. Daher muß jeder Anlaß vermieden werden, der auch nur den Verdacht aufkommen lassen könnte, daß das Statistikgeheimnis nicht gewahrt bleibt."[28])

Die Wissenschaft sitzt mit der amtlichen Statistik im gleichen Boot. Sie ist nicht weniger als die Statistik an diesem Vertrauen interessiert. Es gibt aber erfreulicherweise auch keine Anhaltspunkte dafür, daß durch die Weitergabe anonymisierter Daten an die Wissenschaft dieses Vertrauen beeinträchtigt würde.

Literaturhinweise

Baumann, Reinhold (1985): Datenschutz und Statistik. In: Statistisches Bundesamt (Hrsg.), Datennotstand und Datenschutz. Ergebnisse des 1. Wiesbadener Gesprächs. 30./31. Oktober 1984. Stuttgart: Kohlhammer.

Blossfeld, Peter (1985): Berufseinstieg und Segregationsprozeß. Eine Kohortenanalyse über die Herausbildung von geschlechtsspezifischen Strukturen im Bildungs- und Berufsverlauf. Arbeitspapier Nr. 171 des Sonderforschungsbereichs 3, Frankfurt/Mannheim.

Bolte, Karl Martin und Hradil, Stefan (1984): Soziale Ungleichheit in der Bundesrepublik Deutschland. Opladen: Leske + Budrich.

Bungers, Dieter und Quinke, Hermann (1986): A Microsimulation Model for the German Federal Training Assistance Act — Principles, Problems and Experiences, in: Orcutt, Guy H., Merz, Joachim und Quinke, Hermann (Hrsg.), Microanalytic Simulation Models to Support Social and Financial Policy. Amsterdam: North Holland.

Bürgin, Gerhard, und Reimann, Brigitte (1982): Empirische Sozialwissenschaft und amtliche Statistik aus der Sicht der amtlichen Statistik. Sonderdruck der Referate zur 29. Tagung des Statistischen Beirates. Beilage zu Wirtschaft und Statistik.

Bull, Hans Peter und Dammann, Ulrich (1982): Wissenschaftliche Forschung und Datenschutz. Die Öffentliche Verwaltung 35 (6), S. 213—223.

[27]) Als Beispiel für Datenschutzvorkehrungen für Mikrodaten im universitären Bereich vgl. Wagner und Lutterbeck (1985).
[28]) Vgl. Baumann (1985).

Bundesverfassungsgericht (1983): Das Volkszählungsgesetz-Urteil des Bundesverfassungsgerichts vom 15. Dezember 1983 - 1BvR 209/83 u. a. Datenschutz und Datensicherung 4/84, S. 258–281.

Bundesstatistikgesetz (1986): Entwurf eines Gesetzes über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG). Bundesrat Drucksache 19/86.

Dalenius, Tore E. (1981): Release of Microdata — the International Picture. Bonn: Gesellschaft für Mathematik und Datenverarbeitung.

Dick, Eugen (1986): Use of Simulation Techniques in Developing a Housing Allowance Allocation System (with comment by Volker Lietmeyer). In: Orcutt, Guy H., Merz, Joachim und Quinke, Hermann (Hrsg.), Microanalytic Simulation Models to Support Social and Financial Policy. Amsterdam: North Holland.

Diewald, Martin (1984): Das „SPES-Indikatoren-Tableau 1976" — Fortschreibung bis zum Jahr 1982. Arbeitspapier Nr. 150 des Sonderforschungsbereichs 3, Frankfurt/Mannheim.

Eliasson, Gunnar (1986): The Swedish Micro-to-Macro Model: Idea, Design and Application, in: Orcutt, Guy H., Merz, Joachim, und Quinke, Hermann (Hrsg.), Microanalytic Simulation Models to Support Social and Financial Policy. Amsterdam: North Holland.

Flaherty, David H. (1979): Privacy and Government Data Banks. An International Perspective. London: Mansell.

Gerhardt, Herbert, und Stärk-Rötters, Doris (1985): Zur statistischen Darstellung von Studienverläufen. Wirtschaft und Statistik, H. 8, S. 657—666.

Göbel, Dieter (1983): Lebenseinkommen und Erwerbsbiographie, Eine Längsschnittuntersuchung mit Daten der gesetzlichen Rentenversicherung. Frankfurt: Campus.

Grohmann, Heinz, Bürgin, Gerhard, Krupp, Hans-Jürgen, und Simitis, Spiros (1980): Vielseitige Nutzung statistischer Einzelangaben und Datenschutz. Allgemeines Statistisches Archiv 64 (1), S. 39—75.

Haller, Max (1983): Klassenstruktur und Beschäftigtensystem in Frankreich und der Bundesrepublik Deutschland. Eine makro-soziologische Analyse der Beziehungen zwischen Qualifikationen, Technik und Arbeitsorganisation. In: Haller, Max, und Müller, Walter (Hrsg.), Beschäftigtensystem im gesellschaftlichen Wandel. Frankfurt/New York: Campus, S. 287—370.

Handl, Johann, Mayer, Karl Ulrich, und Müller, Walter (1977): Klassenlage und Sozialstruktur. Frankfurt: Campus.

Handl, Johann (1984): Educational Chances and Occupational Opportunities of Women: A Social-historical Analysis. Journal of Social History 17, S. 463—487.

Hauser, Richard, Cremer-Schäfer, Helga, Nouverté, Udo (1981): Armut, Niedrigeinkommen und Unterversorgung in der Bundesrepublik Deutschland. Frankfurt: Campus.

Hauser, Richard, und Heldmann, Elanie (1981): Die Verteilung impliziter Transfers zugunsten von Eigennutzerhaushalten im Jahr 1969 — Eine mikroökonomische Analyse auf Basis von Individualdaten der EVS; Arbeitspapier Nr. 53 des Sonderforschungsbereichs 3, Frankfurt/Mannheim.

Hauser, Richard, und Fischer, Ingo (1985): The Relative Economic Status of One-Parent-Families in Six Major Industrialized Countries, Arbeitspapier Nr. 187 des Sonderforschungsbereichs 3, Frankfurt/Mannheim.

Heilig, Gerhard (1985): Probleme der Datenbeschaffung in der (universitären) Bevölkerungsforschung. In: Bundesinstitut für Bevölkerungsforschung (Hrsg.), Materialien zur Bevölkerungswissenschaft. Wiesbaden.

Kaase, Max, Krupp, Hans-Jürgen, Pflanz, Manfred, Scheuch, Erwin K., und Simitis, Spiros (Hrsg.) (1980): Datenzugang und Datenschutz. Königstein/Ts.: Athenäum.

König, Wolfgang, und Müller, Walter (1986): Worklife Mobility of Men in France and West Germany 1965—1970. European Sociological Review. Im Druck.

Kortmann, Klaus (1982): Verknüpfung und Ableitung personen- und haushaltsbezogener Mikrodaten. Frankfurt: Campus.

Krupp, Hans-Jürgen (1975): Möglichkeiten der Verbesserung der Einkommens- und Verbraucherstatistik. Göttingen: Schwartz und Co.

Krupp, Hans-Jürgen, und Glatzer, Wolfgang (Hrsg.) (1978): Umverteilung im Sozialstaat. Empirische Einkommensanalysen für die Bundesrepublik Deutschland. Frankfurt: Campus.

Krupp, Hans-Jürgen, Galler, Heinz Peter, Grohmann, Heinz, Hauser, Richard, Wagner, Gerg (Hrsg.) (1981): Alternativen der Rentenreform '84. Frankfurt: Campus.

Krupp, Hans-Jürgen (1982): Empirische Sozialwissenschaft und amtliche Statistik aus der Sicht der sozialwissenschaftlichen Politikberatung. Sonderdruck der Referate zur 29. Tagung des Statistischen Beirates. Beilage zu Wirtschaft und Statistik.

Lüttinger, Paul (1986): Der Mythos der schnellen Integration. Eine empirische Untersuchung zur Integration der Vertriebenen und Flüchtlinge in der Bundesrepublik Deutschland. Zeitschrift für Soziologie 15 (1), S. 20—36.

Majka, Maribeth, und Wertheimer, Richard (1983): Public Use Samples in the United States: Perspectives of Producers and Users. Bonn: Gesellschaft für Mathematik und Datenverarbeitung.

Mayer, Hans-Ludwig (1983): Erwerbstätigkeit — Umschichtung der Erwerbsbevölkerung: Bestands- und Längsschnittergebnisse des Mikrozensus. Wirtschaft und Statistik, H. 10, S. 782—791.

Mayer, Karl Ulrich (1979): Class Formation and Social Reproduction. Current Comparative Research on Social Mobility. In: Geyer, R. F., (Hrsg.): Cross-National and Cross-Cultural Comparative Research in the Social Science. Oxford: Pergamon-Press, S. 37—56.

Mayer, Karl Ulrich (1985): The process of leaving home: A comparison of three cohorts in West Germany. Arbeitspapier Nr. 168 des Sonderforschungsbereiches 3, Frankfurt/Mannheim.

Mayer, Karl Ulrich, und Papastefanou, Georg (1983): Arbeitseinkommen im Lebenslauf. In: Schmähl, Winfried (Hrsg.), Ansätze der Lebenseinkommensanalyse. Tübingen.

Mammey, Ulrich, und Schwartz, Wolfgang (1982): Chancen des sozialen Aufstiegs in den Teilräumen der Bundesrepublik Deutschland. Schriftenreihe Raumordnung des Bundesministeriums für Raumordnung, Bauwesen und Städtebau, Bonn.

Mierheim, Horst, und Wicke, Lutz (1978): Die personelle Vermögensverteilung. Tübingen: Mohr.

Müller, Walter, und Mayer, Karl Ulrich (1976): Chancengleichheit durch Bildung? Deutscher Bildungsrat, Gutachten und Studien der Bildungskommission. Stuttgart: Klett.

Müller, Walter (1982): Empirische Sozialwissenschaft und amtliche Statistik aus der Sicht der empirisch orientierten Forschung. Sonderdruck der Referate zur 29. Tagung des Statistischen Beirates. Beilage zu Wirtschaft und Statistik.

Müller, Walter, Willms, Angelika, und Handl, Johann (1983): Strukturwandel der Frauenerwerbstätigkeit 1880—1980. Frankfurt/New York: Campus.

Orcutt, Guy H., Merz, Joachim, und Quinke, Hermann (1986): Microanalytic Simulation Models to Support Social and Financial Policy. Amsterdam: North Holland.

Paaß, Gerhard, und Wauschkuhn, Udo (1985): Datenzugang, Datenschutz und Anonymisierung. Analysepotential und Identifizierbarkeit von anonymisierten Individualdaten. München und Wien: Oldenbourg.

Schmähl, Winfried (Hrsg.) (1983): Ansätze der Lebenseinkommensanalyse. Tübingen: Mohr.

Sonderforschungsbereich 3 (1985): Das sozio-ökonomische Panel. Bericht über die Forschungstätigkeit 1983—1985. Antrag auf Förderung der Forschungsphase 1986—1988. Universität Frankfurt und Universität Mannheim, Deutsches Institut für Wirtschaftsforschung, Berlin.

Statistisches Bundesamt (Hrsg. 1985): Datenreport 1985. Zahlen und Fakten über die Bundesrepublik Deutschland. Bonn: Schriftenreihe der Bundeszentrale für politische Bildung.

Steger, Almut (1980): Haushalte und Familien bis zum Jahr 2000. Frankfurt: Campus.

Stolz, Irene (1983): Einkommensumverteilung in der Bundesrepublik Deutschland. Frankfurt: Campus.

Stubig, Hans-Jürgen, und Berntsen, Roland (1985): Datenhandbuch zur Lebenslage-Studie. Frankfurt und Mannheim: Sonderforschungsbereich 3, Mikroanalytische Grundlagen der Gesellschaftspolitik.

Tuner, Lotte (1982): Probleme der Übermittlung von statistischen Einzelangaben zu wissenschaftlichen Zwecken. DSWR, S. 61—67.

Wagner, Gerd (1986): Analysepotentiale und -grenzen der gegenwärtigen amtlichen und nichtamtlichen Datenproduktion für einen „Problemlösungsoperator Sozialwissenschaft". In: N. Müller und H. Stachowiak (Hrsg.), Problemlösungsoperator Sozialwissenschaft. Stuttgart, im Druck.

Wagner, Gerhard, und Lutterbeck, Michael (1985): Ergänzungen zur amtlichen Bevölkerungsstatistik: Strategien zur Datenbeschaffung und für den Datenschutz im Sonderforschungsbereich 3. In: Bundesinstitut für Bevölkerungsforschung (Hrsg.), Materialien zur Bevölkerungswissenschaft. Wiesbaden.

Wissenschaftsrat (1986): Stellungnahme zur Gründung einer „Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen e.V.". (GESIS). Berlin: Wissenschaftsrat Drs. 7174/86.

Zapf, Wolfgang (1974): Sozialberichterstattung und amtliche Statistik. Referate zum Thema „Messung der Lebensqualität und amtliche Statistik" anläßlich der 21. Tagung des Statistischen Beirats am 16. Mai 1974. Beilage zu Wirtschaft und Statistik, Heft 8, S. 3—8.

Zapf, Wolfgang (Hrsg.) (1977): Lebensbedingungen in der Bundesrepublik. Sozialer Wandel und Wohlfahrtsentwicklung. Frankfurt: Campus.

Zapf, Wolfgang (1985): Der Zugang der Wissenschaft zur statistischen Information – Forderung und Realität. In: Statistisches Bundesamt (Hrsg.), Datennotstand und Datenschutz. Ergebnisse des 1. Wiesbadener Gesprächs. 30./31. Oktober 1984. Stuttgart: Kohlhammer.

Prof. Dr. Franz U. Pappi
Universität Kiel

Allgemeine Bevölkerungsumfragen für die Sozialwissenschaften
Konzeption – Umsetzung im ALLBUS – Nutzung

Einleitung

Allgemeine Bevölkerungsumfragen für die Sozialwissenschaften (ALLBUS) werden derzeit in mehreren Ländern nach einer Konzeption durchgeführt, die vor etwa 15 Jahren im Rahmen der Sozialindikatorenbewegung entwickelt worden ist. Die Grundidee ist, gesellschaftlichen Wandel mit den Mitteln der Umfrageforschung zu erfassen. Subjektive Sozialindikatoren wie Zufriedenheit, Erwartungen etc. werden ebenfalls mit Hilfe von Umfragen gemessen. Nach Zapf bedarf diese Art der Wohlfahrtsmessung, die auf normativ vorgegebene Wohlfahrtsziele ausgerichtet ist, der Ergänzung durch eine Dauerbeobachtung des sozialen Wandels (1977a: 234–236). Es ist eine Aufgabe der Sozialwissenschaft, unabhängig von normativen Vorgaben, relevante Dimensionen des sozialen Wandels herauszuarbeiten. Dies kann beim gegenwärtigen Wissensstand nicht theoretisch nach den wissenschaftlichen Regeln der Deduktion, sondern muß auf induktivem Wege erfolgen durch Beobachtung von Indikatoren, die sich, wenn vielleicht auch nur vorläufig, bisher in der Forschungspraxis bewährt haben. Für diese wesentlich auf Duncan (1969) zurückgehende Konzeption einer Dauerbeobachtung des sozialen Wandels spielen replikative Surveys eine entscheidende Rolle (vgl. auch Zapf 1977b: 220).

Ein replikativer Survey ist eine Wiederholungsbefragung im Sinne einer Trendstudie. Aus derselben Zielpopulation werden in regelmäßigen Abständen unabhängige Stichproben gezogen, wobei man mit 1500 bis 3000 realisierten Interviews rechnet. Die einzelnen Befragungen müssen zum großen Teil äquivalente Fragen enthalten, so daß Veränderungen in den Angaben über Einstellungen und Verhaltensweisen tatsächlich auf Wandel zurückgeführt werden können. Veränderungen von Frageformulierungen, aber auch von Interviewerstrategien können zu Forschungsartefakten führen, die einen Wandel in den gemessenen Merkmalen nur vortäuschen.

Neben dem wissenschaftlichen Ziel der Untersuchung des sozialen Wandels wird ein zweites Ziel verfolgt. Die erhobenen Daten werden zur Verbesserung der Infrastruktur in den Sozialwissenschaften interessierten Forschern und Studenten sofort nach der Datenaufbereitung für Sekundäranalysen zur Verfügung gestellt. Dieses Dienstleistungsangebot ist für die empirischen Sozialwissenschaften von großer Bedeutung. Im Unterschied zu den Naturwissenschaften und zur Psychologie muß die Makro-Soziologie und die politische Wissenschaft auf Laboratorien verzichten. Theoretisch relevante Daten lassen sich nicht nach Belieben in Laboruntersuchungen erzeugen. Allgemeine Bevölkerungsumfragen für die Sozialwissenschaften, die regelmäßig durchgeführt werden, schaffen hier einen gewissen Ersatz. Will man dabei über die Arbeitsweise der Moralstatistik

des letzten Jahrhunderts hinausgehen, müssen für Sekundäranalysen die erhobenen Individualdaten zur Verfügung gestellt werden. Tabellen und Randverteilungen reichen für die normalen Fragestellungen nicht aus. Auf diesen Aspekt wird im dritten Teil einzugehen sein, wenn die Nutzung der ALLBUS-Daten besprochen wird.

1 Allgemeine Konzeption

Für die Untersuchung des sozialen Wandels mit Mitteln der Umfrageforschung sind verschiedene Strategien denkbar. Den Anregungen von Duncan (1969) folgend, sind auch in der Bundesrepublik inzwischen zwei Bezugsstudien vollständig repliziert worden (vgl. Noelle-Neumann/Piel 1983; Allerbeck/Hoag 1985). Man kann natürlich auch versuchen, Zeitreihen ohne eigene Sekundäranalyse allein aus dem Archivmaterial von Umfragedatenarchiven zusammenzustellen. Dazu eignen sich in erster Linie die Umfragen der politischen Meinungsforschung, die von kommerziellen Instituten mit gewisser Regelmäßigkeit für Regierungen, Parteien oder Massenmedien durchgeführt werden. Wenn genügend Meßzeitpunkte zur Verfügung stehen wie z. B. für die Popularität der Regierung, kann man sich bei dieser Art Daten manchmal mit den Randverteilungen zufrieden geben und braucht nicht auf die primären Individualdaten zurückzugreifen.

Allgemeine Bevölkerungsumfragen für die Sozialwissenschaften sind anders konzipiert. Die wichtigsten Gesichtspunkte lassen sich wie folgt zusammenfassen:

1. Omnibus-Replikation

Die Umfragen sind als Mehrthemenbefragungen geplant, wobei die zu replizierenden Fragen aus verschiedenen Bezugsstudien stammen. Ist zu den früheren Zeitpunkten bereits dieselbe Zielpopulation befragt worden, kann sofort ein Zeitvergleich durchgeführt werden. Die Alternative ist der Start einer neuen Meßreihe mit neu entwickelten Einstellungsskalen und Verhaltensfragen, die erst nach mehreren Meßwiederholungen unter dem Gesichtspunkt des sozialen Wandels ausgewertet werden können.

2. Theoriebezogenheit der Indikatoren

In Ermangelung einer allgemeinen Theorie des sozialen Wandels muß die Indikatorenauswahl zwar induktiv erfolgen, dies schließt aber nicht aus, wenigstens von den Einzelindikatoren Theoriebezogenheit zu verlangen. Die Theoriebildung erfolgt kumulativ; wer sich an der heutigen Theoriediskussion orientiert, wird deshalb auch künftig eher brauchbare Indikatoren auswählen als derjenige, der sich ausschließlich von Aktualitätsgesichtspunkten der öffentlichen Meinung leiten läßt.

3. Auswertungsgesichtspunkte für Querschnitte

Viele Sekundäranalysen werden als Querschnittsanalysen geplant. Dafür ist die konventionelle Auswertungsstrategie in den Sozialwissenschaften zu berücksichtigen. Einstellungen und Verhaltensweisen werden in der Regel sozialstrukturell erklärt. Man erwartet

Unterschiede nach Berufszugehörigkeit, Konfession, Stellung im Lebenszyklus, Haushaltszusammensetzung usw. Daraus folgt, daß die sozialstrukturellen Hintergrundsmerkmale des Befragten ausführlich zu erfassen sind. Unabhängig von der Verwendung der Hintergrundsmerkmale als unabhängige Variablen ist der deskriptive Wert dieser Information zu berücksichtigen.

4. Periodizität des sozialen Wandels

Wandel kann kurz- und langfristig sein, er kann erfolgen als Reaktion auf äußere Ereignisse oder als Auswirkung tiefer liegender struktureller Veränderungen. Aus dem vermuteten Prozess des sozialen Wandels für die untersuchten Indikatoren kann man ableiten, in welchen Zeitintervallen die Umfragen durchzuführen sind. Soziologen interessieren sich eher für langfristigen Wandel, der sozialstrukturell erklärt werden kann. Im einfachsten Fall handelt es sich dabei um Kompositionshypothesen, die von einer Veränderung der Gesamtbevölkerung nach sozialen Gruppen oder Generationen ausgehen. Jährliche oder sogar zweijährige Datenerhebung mit einer genügend großen Stichprobe ist für solche Prozesse langfristigen Wandels voll ausreichend.

2 Der ALLBUS als deutsche Version einer allgemeinen Bevölkerungsumfrage der Sozialwissenschaften

Der erste replikative Survey, der die vorgetragene Konzeption in die Praxis umgesetzt hat, ist der amerikanische General Social Survey (GSS), (vgl. den Beitrag von Duane F. Alwin in dieser Veröffentlichung, S. 12 ff.). Er ist anläßlich der Veröffentlichung des kumulativen Codebuchs der ersten sechs Umfragen von 1972 bis 1977 von Fachkollegen sehr gut rezensiert worden (vgl. Glenn 1978; Converse 1978; Cutler 1978; Hyman 1978). An diesem Vorbild hat sich der ALLBUS orientiert. Dabei darf nicht vergessen werden, daß es in den USA dank der starken Stellung der akademischen Umfrageforschung einen sehr viel größeren Fundus an replikationsfähigen Umfragen gibt, als vom GSS ausgeschöpft werden. So werden politische Indikatoren schwerpunktmäßig in den Wahlstudien des Center for Political Studies in Ann Arbor erhoben (vgl. zur Studie 1976 Converse und Markus 1979 und zur Studie 1980 Markus 1982), die seit 1948 in mindestens vierjährigem Turnus durchgeführt werden. Das Center for Political Studies ist aus dem Survey Research Center hervorgegangen; beide Zentren sind Teile des Instituts for Social Research (ISR) in Ann Arbor, die auf eine inzwischen fast 40jährige Tradition der nationalen Umfrageforschung in den USA zurückblicken können. Ihr gemeinsamer Fundus an Zeitreihen für soziale Einstellungen, die über den engeren politischen Bereich hinausgehen, ist in dem „American Social Attitude Data Source Book 1947–1978" (Converse et. al. 1978) dokumentiert.

Da es in der Bundesrepublik keine akademischen Umfrageforschungsinstitute gibt, die mit dem Institute for Social Research (ISR) oder dem National Opinion Research Center (NORC) auch nur annähernd vergleichbar wären, überrascht es nicht, daß der deutsche ALLBUS schlechtere Ausgangsbedingungen vorfand als der GSS. Trotzdem waren zwei günstige infrastrukturelle Voraussetzungen gegeben, die auch genutzt wurden. Die Antragstellergruppe, die die notwendigen Mittel bei der Deutschen Forschungsgemeinschaft

(DFG) beantragte, wurde von Anfang an bei der Konzeption und Datenerhebung vom Zentrum für Umfragen, Methoden und Analysen e.V. (ZUMA), Mannheim, und bei der Datenaufbereitung und -distribution vom Zentralarchiv für empirische Sozialforschung der Universität Köln unterstützt. Nur dank dieser institutionellen Unterstützung konnte dieses Dienstleistungsangebot an die Profession bisher als Serie von Einzelprojekten abgewickelt werden.

Vier Erhebungen sind bisher von der DFG bewilligt worden, die erste für 1980 (vgl. Lepsius, Scheuch, Ziegler 1982), die zweite für 1982 (vgl. Lepsius, Scheuch, Ziegler 1984), die dritte für 1984 (vgl. Müller, Pappi, Scheuch, Ziegler o. J.) und eine vierte Erhebung, die von Müller, Mayer, Pappi, Scheuch und Ziegler beantragt wurde, wird im Frühjahr 1986 abgeschlossen. Damit liegt für die Bundesrepublik bereits jetzt eine Datenfülle für sozialwissenschaftliche Sekundäranalysen vor, wie für sonst kein anderes europäisches Land. Immerhin ist für Großbritannien 1983 ein vergleichbares Projekt gestartet worden (vgl. Jowell und Airey 1984) und andere Länder werden vielleicht folgen.

Wenn im Herbst dieses Jahres vier ALLBUS-Erhebungen für Sekundäranalysen zur Verfügung stehen werden, heißt das nicht, daß die Konzeption für derartige allgemeine Bevölkerungsumfragen geradlinig in die Praxis umgesetzt worden wäre. Für den ersten ALLBUS sind Kriterien für die Aufnahme von Fragen entwickelt worden, die direkt aus der allgemeinen Zielvorstellung abgeleitet waren (vgl. Porst 1985: 47). Die wichtigste Forderung war die nach Fragenkontinuität. Es sollten mit Vorrang Fragen aufgenommen werden, „die bereits in früheren nationalen Erhebungen gestellt worden waren, sich methodisch bewährt haben und wissenschaftlich diskutiert waren" (Porst 1985: 47). Das Angebot an Einstellungsskalen und Einzelfragen, die dieser Forderung genügt hätten, erwies sich in der deutschen Umfrageforschung aber als äußerst schmal (vgl. Mayer 1984: 17). Dies gilt vor allem für Einstellungsskalen. Etwas günstiger war die Situation im Hinblick auf Einzelfragen, aber auch hier gab es häufig Interpretationsprobleme, weil der Indikatorwert der Frage unklar blieb. Bei anderen Fragen war der Problembezug inzwischen überholt.

Damit sind allgemeine Probleme der Erfassung von Einstellungswandel über längere Zeiträume angesprochen, die zu den besonderen deutschen Schwierigkeiten der zu schmalen Datenbasis der akademischen Umfrageforschung hinzukommen. Es ist bekannt, daß kleine Änderungen der Frageformulierung bereits zu anderen Ergebnissen führen können (vgl. Schuman und Presser 1981). Dies führt zur Forderung der strikten Replikation einer Frage ohne Veränderung des Wortlauts und der Antwortkategorien. Andererseits ändert sich natürlich auch die Sprache, so daß Formulierungen, die in den fünfziger Jahren normal klangen, inzwischen merkwürdig anmuten. Die Kontroverse um die angeblich sinkende Arbeitszufriedenheit der Deutschen vermittelt einen aktuellen Anschauungsunterricht zu diesem Problem (Noelle-Neumann, Strümpel 1984; Reuband 1985).

Der Beitrag des ALLBUS zur Lösung dieser Probleme besteht in den folgenden Schwerpunktsetzungen, die unter Beibehaltung der ursprünglichen Zielsetzung, sozialen Wandel mit den Mitteln der Umfrageforschung zu erfassen, angesichts der praktischen Erfahrungen mit dem ersten ALLBUS festgelegt wurden.

1. Der heutige ALLBUS als Baseline-Studie für künftige ALLBUS-Erhebungen

Unter Berücksichtigung dieser Forderung wurden bereits für die erste Erhebung neue Fragen und Skalen entwickelt. Als Beispiel sei die Skala „Einstellung gegenüber Gastarbeitern" erwähnt, die diskriminierende Einstellungen gegenüber dieser Bevölkerungsgruppe messen soll (vgl. ZUMA 1983: E 01). 1984 wurde diese Skala repliziert mit dem eindeutigen Ergebnis, daß die negativen Einstellungen gegenüber Gastarbeitern abgenommen haben. Dieses Ergebnis kann als gut gesichert angesehen werden, weil es nicht nur mit einer einzelnen Frage belegt wird, sondern mit einer Skala, deren Einzelitems alle dieselben Antworttendenzen zeigen. Das Ergebnis wird aber noch durch einen zweiten Grund erhärtet. Als mögliche Ursache der Einstellung sind Kontakte zu Gastarbeitern erhoben worden. Da die freiwilligen Kontakte zu Gastarbeitern von 1980 bis 1984 ebenfalls zugenommen haben, kann man die Abnahme der negativen Einstellungen nicht nur konstatieren, sondern auch erklären (Gehring und Boeltken 1985).

Dieses Beispiel verdeutlicht die ALLBUS-Strategie: Erfassung von Einstellungen möglichst mit Skalen und Aufnahme möglicher Ursachen der Einstellungen in dieselbe Erhebung. Die Konsequenz dieser Strategie ist, daß inhaltliche Schwerpunkte für die einzelnen Erhebungen festgelegt werden müssen, um die notwendigen Fragen in genügender Breite auch stellen zu können.

Ab der zweiten ALLBUS-Erhebung sind die Antragsteller diesen Weg konsequent gegangen. Die folgenden inhaltlichen Schwerpunkte wurden bisher untersucht: Religiöse Einstellungen (1982), Einstellungen zum Wohlfahrtsstaat und zur sozialen Ungleichheit (1984), Bildung und Kulturfertigkeiten (1986), soziale Unterstützung in persönlichen Netzwerken (1986).

Für eine Dienstleistungseinrichtung wie den ALLBUS ist mit der Festlegung solcher inhaltlicher Schwerpunkte natürlich ein erhöhter Rechtfertigungszwang gegenüber der Profession verbunden. Warum soll Primärforschung auf diesen Gebieten im Rahmen des ALLBUS und nicht in separat zu finanzierenden Einzelprojekten erfolgen? Diese Rechtfertigung hängt unmittelbar mit der zweiten Schwerpunktsetzung zusammen.

2. Der internationale Vergleich als Ersatz für Längsschnittuntersuchungen

Internationale Vergleichsstudien sind schwer zu organisieren und können von einem eingespielten Forschungsapparat wie dem des ALLBUS profitieren. Aber warum sollte man sie überhaupt durchführen? Die Rechtfertigung muß im Zusammenhang mit der allgemeinen Zielsetzung gesehen werden.

Bestimmte strukturelle Wandlungsprozesse erfassen alle westlichen Industriegesellschaften in gleicher Weise; allerdings gibt es Vorreiter und Nachzügler dieser Entwicklung. Der internationale Vergleich kann deshalb als Querschnittsersatz für eine noch fehlende lange Zeitreihe dienen, indem Länder unterschiedlicher Entwicklungsstufe verglichen werden.

Andere Entwicklungen sind abhängig von institutionellen Besonderheiten der einzelnen Gesellschaften wie dem System der sozialen Sicherung, dem Parteiensystem oder der Organisationsstruktur der Kirchen (Anstaltskirchen versus Denominationalismus). Ein Verständnis der Zusammenhänge zwischen Institutionen und Einstellungen bzw. Verhaltensweisen der Bevölkerung ist eine wichtige Voraussetzung für eine empirische Theorie des sozialen Wandels.

Von den bisherigen Schwerpunkten wurden die religiösen Einstellungen, die Einstellungen zum Wohlfahrtsstaat und die soziale Unterstützung in persönlichen Netzwerken als Teil internationaler Vergleichsstudien geplant. Diese Daten eignen sich in erster Linie zur Untersuchung der Abhängigkeit individueller Verhaltensweisen und Einstellungen von den institutionellen Besonderheiten der einzelnen Länder. Im Falle der Religion ist die Vergleichsgesellschaft die der Niederlande. Fragen zum Wohlfahrtsstaat sind teilweise im amerikanischen GSS repliziert worden und der Fragenteil über die soziale Unterstützung in persönlichen Netzwerken wird in diesem Jahr in den USA, in Australien, in Großbritannien und in der Bundesrepublik in nationale Erhebungen eingeschaltet.

Primärforschung wird also im ALLBUS-Projekt für internationale Vergleichsstudien betrieben. Daneben ist sie auch für den dritten Schwerpunkt erforderlich, der in besonderer Beziehung zur Aufgabenstellung von ZUMA steht.

3. Methodenforschung zur Verbesserung der Zuverlässigkeit und Gültigkeit von Umfrageergebnissen

Die Schwierigkeiten der Interpretation von Trenddaten sind häufig methodischer Natur. Sind die Zielpopulationen vergleichbar? Wie verzerrt die Vorgehensweise bei der Auswahl der Zielpersonen oder wie verzerren Ausfälle die Ergebnisse? Welcher Einfluß geht von der Zusammensetzung der Interviewerstäbe aus? Welche Fragebogeneffekte (z. B. Stellung der Frage im Fragebogen) und typischen Verständnisprobleme von seiten der Befragten gibt es? (Vgl. zu diesen Problemen z. B. Martin 1983).

Bisher ist mit jeder Allbus-Erhebung ein Methodenforschungsprojekt verbunden worden. Dabei standen Stichprobenprobleme bereits zweimal im Mittelpunkt, zum einen 1980 (Kontaktverlauf Interviewer – Zielperson; zur Stichprobe allgemein vgl. Kirschner 1984) und 1986 (Non-Response-Studie). 1980 wurden auch Interviewereinflüsse untersucht (vgl. Esser 1984; Schanz und Schmidt 1984). 1982 wurde in einer getrennten Erhebung die internationale Vergleichbarkeit von Einstellungsskalen untersucht und 1984 wurde die erste deutsche Test-Restest-Studie zur Überprüfung der Zuverlässigkeit der Meßinstrumente unter den normalen Bedingungen der Umfrageforschung durchgeführt.

Die allgemeine Konzeption für derartige Bevölkerungsumfragen mußte also den besonderen Bedingungen der deutschen akademischen Umfrageforschung angepaßt werden. Dies war ohne Aufgabe des ursprünglichen Ziels möglich. Das bisherige Schwerpunktprogramm setzt aber einen eindeutigen Akzent zu Gunsten der mittel- und

langfristigen Wandlungsprozesse und gegen kurzfristigen Wandel als Reaktion auf äußere Ereignisse. Die Fragen des ALLBUS-internen Replikationsteils müssen nur in größeren Abständen wiederholt werden.

3 Nutzung

Das zweite Hauptziel des ALLBUS betrifft die Datengenerierung für Sekundäranalysen. Den empirischen Sozialforschern wird „ein kontinuierliches Angebot an thematisch interessanten und methodisch hochwertigen Daten aus sozialwissenschaftlichen Repräsentativbefragungen der bundesdeutschen Bevölkerung" gemacht (Mayer et. al. 1985: 10). Dieses Ziel wird durch die tatsächliche Nutzung der ALLBUS-Daten durch die Profession realisiert.

Für die Distribution gelten die folgenden Gesichtspunkte. Die Daten können sofort nach Aufbereitung und Bereinigung vom Zentralarchiv für empirische Sozialforschung ohne inhaltliche Beschränkung für die Auswertung erworben werden. Es werden die anonymisierten Individualdaten zur Verfügung gestellt, wobei eine Identifikation der Stimmbezirke als der Primary Sampling Units nicht möglich ist. Die kleinste regionale Einheit ist der Regierungsbezirk. Für die Weitergabe der Daten gilt also die folgende Datenschutzregel: Umfragedaten sind anonymisiert, wenn Name und Adresse des Befragten gelöscht sind und die primäre Stichprobeneinheit nicht namentlich bekannt ist.

Dem Institut, das die Feldarbeit für die Umfrage durchgeführt hat, ist die Identität der Primäreinheiten der Stichprobe bekannt. Diese Information darf auch nicht vernichtet werden, weil über diese Nummer Kontextmerkmale zu den Individualdaten hinzugelesen werden können. Solche Kontextmerkmale können z. B. die Wahlergebnisse für die Stimmbezirke oder die soziale Zusammensetzung der Bevölkerung in der Befragungsgemeinde sein. Ziel der Kontextanalyse ist es, Milieueinflüsse auf individuelle Einstellungen und Verhaltensweisen zu untersuchen. Es handelt sich hier um einen Nebenaspekt der Anonymisierungsproblematik, der für normale Sekundäranalysen keine Rolle spielt.

Was die tatsächliche Nutzung der ALLBUS-Daten betrifft, kann ein voller Erfolg vermeldet werden. Nach dem Stand von Mitte 1985 sind insgesamt 94 wissenschaftliche Arbeiten (Veröffentlichungen und nicht veröffentlichte Diplom- bzw. Magisterarbeiten, Dissertationen und Habilitationsschriften) gezählt worden, in denen ALLBUS-Daten verwendet wurden (vgl. Porst 1985b). Die Ergebnisse wurden fast ausschließlich durch Auswertung der Individualdatensätze erzielt.

Knapp die Hälfte der Arbeiten beschäftigt sich mit methodischen Fragen, wobei die Spanne von der illustrativen Verwendung der Daten zur Klärung bestimmter Methoden wie der Faktorenanalyse (Arminger 1984) oder der log-linearen Analyse (vgl. Küchler 1980) bis zu den Berichten über die verschiedenen Methodenforschungsprojekte des ALLBUS selbst reicht. Über das Bild hinaus, das sich aus den Veröffentlichungen ergibt, ist bekannt, daß der ALLBUS häufig in universitären Forschungsübungen ausgewertet wird. Der Schluß erscheint gerechtfertigt, daß sich der ALLBUS als Laborersatz bei der Ausbildung von Sozialwissenschaftlern bereits bestens bewährt hat.

Nach der Einschätzung von Porst liegt der Schwerpunkt der ALLBUS-Nutzung allerdings auf „Analysen zu inhaltlichen Fragestellungen, und hier vor allem auf solchen Arbeiten, die man im weitesten Sinne als Einstellungsanalysen bezeichnen könnte" (Porst 1985a: 137). Wie in der Anfangsphase nicht anders zu erwarten, stehen einfache Querschnittsauswertungen im Vordergrund. Analysen des sozialen Wandels und internationale Vergleiche sind noch kaum durchgeführt worden.

Die Verbesserung der Infrastruktur für die sozialwissenschaftliche Forschung bedeutet nicht automatisch, daß die besseren Möglichkeiten auch genutzt werden. Die intensivere Nutzung dieser Möglichkeiten ist auch aktiv zu fördern und diese Förderung kann sich nicht auf den Aspekt der Public Relations beschränken. Zwei Erweiterungen des Datenangebotes sind aus meiner Sicht vorrangig.

1. Die Vergleichsstudien für Trendaussagen müssen den modernen Anforderungen entsprechend aufbereitet werden.

Das Zentralarchiv für empirische Sozialforschung hat eine ganze Reihe von Vergleichsstudien bereits so aufbereitet, daß sie ohne große Mühe ausgewertet werden können. Außerdem wurde aus den bisherigen drei ALLBUS-Erhebungen ein kumulierter Datensatz gebildet, der Trendauswertungen wenigstens für den kurzen Zeitraum von 1980 bis 1984 außerordentlich erleichtert. Gerade wenn man an sozialem Wandel interessiert ist, muß man aber darauf dringen, daß die nationalen Umfragen aus den fünfziger und sechziger Jahren systematisch für Auswertungszwecke aufbereitet werden. Nur wenn Individualdatensätze in bereinigter Form zur Verfügung stehen, werden die für eine Untersuchung des sozialen Wandels dringend nötigen Längsschnittanalysen auch durchgeführt werden.

2. Der Mikrozensus muß für Sekundäranalysen nach dem Vorbild des ALLBUS zugänglich gemacht werden.

Dadurch kann nicht zuletzt die Attraktivität des ALLBUS für sozialstrukturelle Analysen gesteigert werden.

Bisher ist der ALLBUS kaum für Sozialstrukturanalysen verwendet worden (vgl. aber Diekmann 1984 und Porst 1984). Dies hängt nicht zuletzt auch mit einer gewissen Unsicherheit über die systematischen Fehler von allgemeinen Bevölkerungsumfragen zusammen. Um hier eine vernünftige Beurteilungsbasis zu bekommen, müssen die Ergebnisse des ALLBUS mit denen des Mikrozensus für die Variablen verglichen werden, die in beiden Erhebungen identisch abgefragt werden. Leider sind solche Parametervergleiche auf der Basis der veröffentlichten Ergebnisse der amtlichen Statistik meist nicht möglich, weil die Zielpopulationen abweichen. Der ALLBUS erfaßt die deutsche erwachsene Wohnbevölkerung in Privathaushalten, im Mikrozensus ist die Zielpopulation umfassender definiert. Technisch wäre dieses Problem leicht zu lösen. Der interessierte Sozialwissenschaftler erstellt sich die benötigten Tabellen für die Parametervergleiche selbst aus einem allgemein zugänglichen Individualdatensatz des Mikrozensus. Mit derartigen Ver-

gleichen könnte der Aussagewert der einzelnen Erhebung besser abgeschätzt werden, was sich zum Vorteil der akademischen Umfrageforschung und der amtlichen Statistik auswirken würde. Im Interesse der Forschung sind für die Zukunft hier praktikable Lösungen anzustreben.

Literaturhinweise

Allerbeck, Klaus, und Hoag, Wendy J. (1985): Jugend ohne Zukunft? Einstellungen, Umwelt, Lebensperspektiven. München: Piper.

Arminger, Gerhard (1984): „Neuere Entwicklungen der explorativen Faktorenanalyse." Allgemeines Statistisches Archiv 68 (Heft 1): S. 118–139.

Converse, Philip E. (1978): "Toward a more cumulative inquiry." Contemporary Sociology 7 (September): S. 535–541.

Converse, Philip E., und Markus, Gregory B. (1979): "Plus ca change . . .: The new CPS Election Study Panel." American Political Science Review 73 (März): S. 32–49.

Converse, Philip E., Dotson, Jean D., Hoag, Wendy J., und McGee III, William H. (1980): American Social Attitudes Data Sourcebook 1947–1978. Cambridge, Mass., und London: Harvard University Press.

Cutler, Stephen (1978): "Instructional uses of the General Social Surveys." Contemporary Sociology 7 (September): S. 541–545.

Diekmann, Andreas (1984): „Einkommensdiskriminierung von Frauen – Messungen, Analyseverfahren und empirische Anwendungen auf Angestelltenein kommen in der Bundesrepublik." S. 315–351 in: Mayer, Karl Ulrich, und Schmidt, Peter (Hrsg.), Allgemeine Bevölkerungsumfragen der Sozialwissenschaften. Frankfurt, New York: Campus.

Duncan, Otis Dudley (1969): Toward Social Reporting: Next Steps. New York: Russell Sage.

Esser, Hartmut (1984): „Determinanten des Interviewer- und Befragtenverhaltens: Probleme der theoretischen Erklärung und empirischen Untersuchung von Interviewereffekten." S. 26–71 in: Mayer, Karl Ulrich, und Schmidt, Peter (Hrsg.), Allgemeine Bevölkerungsumfragen der Sozialwissenschaften. Frankfurt, New York: Campus.

Gehring, Anne-Katrin, und Böltken, Ferdinand (1985): „Einstellungen zu Gastarbeitern 1980 und 1984: Ein Vergleich." ZA-Information 17 (November): S. 23–33.

Glenn, Norval (1978): "The General Social Surveys: Editorial introduction to a symposium." Contemporary Sociology 7 (September): S. 532–535.

Hyman, Herbert H. (1978): "A banquet for secondary analysis." Contemporary Sociology 7 (September): S. 545–549.

Jowell, Roger, und Airey, Colin (Hrsg. 1984): British Social Attitudes in the 1984 report. Aldershot: Gower Publishing Company. Chapter 1: Introducing the Survey by Roger Jowell.

Kirschner, Hans-Peter (1984): „ALLBUS 1980: Stichprobenplan und Gewichtung." S. 114–182, in: Mayer, Karl Ulrich, und Schmidt, Peter (Hrsg.), Allgemeine Bevölkerungsumfragen der Sozialwissenschaften, Frankfurt, New York: Campus.

Küchler, Manfred, und Schwedler, Erhard (1980): „Die Analyse von kreuztabellierten Massendaten: Eine Diskussion neuerer Verfahren." Allgemeines Statistisches Archiv 64 (Heft 4), S. 360–389.

Lepsius, M. Rainer, Scheuch, Erwin K., Ziegler, Rolf (1982): Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS 1980. Codebuch mit Methodenbericht und Vergleichsdaten ZA-Nr. 1000. Köln und Mannheim: Zentralarchiv für empirische Sozialforschung und ZUMA.

Lepsius, M. Rainer, Scheuch, Erwin K., Ziegler, Rolf (1984): Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS 1982. Codebuch mit Methodenbericht und Vergleichsdaten ZA-Nr. 1160. Köln und Mannheim: Zentralarchiv für empirische Sozialforschung und ZUMA.

Markus, Gregory B. (1982): "Political attitudes during an election year: A report on the 1980 NES Panel Study." American Political Science Review 76 (September): S. 538–560.

Martin, Elisabeth (1983): "Surveys as social indicators: Problems in monitoring trends." S. 677–743 in: Rossi, Peter H., Wright, James D., Anderson, Andy S. (Hrsg.), Handbook of Survey Research. New York u. a.: Academic Press.

Mayer, Karl Ulrich, Müller, Walter, Pappi, Franz U., Scheuch, Erwin K., Ziegler, Rolf (1985): Antrag auf Gewährung einer Sachbeihilfe zum Thema Allgemeine Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS) 1986. Universität Mannheim.

Müller, Walter, Pappi, Franz U., Scheuch, Erwin K., Ziegler, Rolf, (o. J.): Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS 1984, Köln und Mannheim: Zentralarchiv für empirische Sozialforschung und ZUMA.

Noelle-Neumann, Elisabeth, und Piel, Edgar (Hrsg. 1983): Eine Generation später. Bundesrepublik Deutschland 1953–1979. München u. a.: Saur.

Noelle-Neumann, Elisabeth, und Stümpel, Burkhard (1984): Macht Arbeit krank? Macht Arbeit glücklich? Eine aktuelle Kontroverse. München: Piper.

Porst, Rolf (1984): „Haushalte und Familien 1982." Zeitschrift für Soziologie 13 (April): S. 165–175.

Porst, Rolf (1985a): Praxis der Umfrageforschung. Erhebung und Auswertung sozialwissenschaftlicher Umfragedaten. Stuttgart: Teubner.

Porst, Rolf (1985b): ALLBUS-Bibliographie. (4. Fassung, Stand: 30.6.1985). Mannheim: ZUMA.

Reuband, Karl-Heinz (1985): „Arbeit und Wertewandel – mehr Mythos als Realität?" Kölner Zeitschrift für Soziologie und Sozialpsychologie 37 (Dezember): S. 723–746.

Schanz, Volker, und Schmidt, Peter (1984): „Interviewsituation, Interviewermerkmale und Reaktion von Befragten im Interview: eine multivariate Analyse", S. 72–113, in: Mayer, Karl Ulrich, und Schmidt, Peter (Hrsg.), Allgemeine Bevölkerungsumfragen der Sozialwissenschaften. Frankfurt, New York: Campus.

Schumann, Howard, und Presser, Stanley (1981): Questions in Attitude Surveys. New York u. a.: Academic Press.

Zapf, Wolfgang (1977a): „Soziale Indikatoren – Eine Zwischenbilanz." S. 231–246 in: Krupp, Hans-Jürgen, und Zapf, Wolfgang, Sozialpolitik und Sozialberichterstattung. Frankfurt, New York: Campus.

Zapf, Wolfgang (1977b): „Gesellschaftliche Dauerbeobachtung und aktive Politik." S. 210–230 in: Krupp, Hans-Jürgen, und Zapf, Wolfgang, Sozialpolitik und Sozialberichterstattung. Frankfurt, New York: Campus.

Zentrum für Umfragen, Methoden und Analysen e. V. – ZUMA, Informationszentrum Sozialwissenschaften (Hrsg.), 1983: ZUMA-Handbuch sozialwissenschaftlicher Skalen. Mannheim und Bonn.

Dr. Gerhard Paaß
Gesellschaft für Mathematik und Datenverarbeitung, Sankt Augustin

Re-Identifikationsrisiko von Einzelangaben

Einleitung

Seit Mitte der siebziger Jahre arbeitet die Gesellschaft für Mathematik und Datenverarbeitung (GMD) an der Entwicklung von mikroanalytischen Simulationsmodellen. Hauptsächliches Ziel war hierbei die Bereitstellung eines Instrumentariums zur Untersuchung der Konsequenzen verschiedener Transfer-, Leistungs- und Steuergesetze. Ausgangspunkt mikroanalytischer Simulationsmodelle ist eine repräsentative Stichprobe von Personen bzw. Haushalten für ein gegebenes Basisjahr. Für jeden dieser Haushalte wird sowohl die innere Entwicklung des Haushalts (Geburt, Tod, Heirat, Berufskarrieren, usw.) als auch die Einkommenssituation (Einkommen, Konsum, Vermögen etc.) für zukünftige Jahre simuliert. Hierbei läßt sich auf Einzelfallebene die Auswirkung unterschiedlicher Ausgestaltungen der fraglichen Transfergesetze auf die Einkommenssituation der Personen untersuchen. Durch Hochrechnung kann man Mittelwerte für interessierende Bevölkerungsgruppen (z. B. alleinstehende Frauen mit Kindern) berechnen oder auch den gesamten zu erwartenden Finanzbedarf einer Regelung bestimmen.[1])

Da bei Simulationen dieser Art eine Kenntnis der Identität der einzelnen Personen nicht erforderlich ist, werden direkte Identifikationsmerkmale, wie Name und Adresse, aus den Datensätzen der Haushalte entfernt. Trotzdem kann bei derartigen formal anonymisierten Daten unter gewissen Umständen der Personenbezug eines Datensatzes durch Abgleich der erhobenen Datensatzmerkmale mit den aus anderen Quellen bekannten Merkmalen einer Person rekonstruiert werden. Daher dürfen nach dem § 11 Abs. 5 des Bundesstatistikgesetzes (BStatG) [2]) statistische Einzelangaben – d. h. Daten einzelner Personen, Haushalte, Firmen etc. – nur dann freigegeben werden, wenn der Personenbezug nicht mehr gegeben ist:

> „Einzelangaben, die so anonymisiert werden, daß sie Auskunftspflichtigen nicht mehr zuzuordnen sind, dürfen vom Statistischen Bundesamt oder von den Statistischen Landesämtern übermittelt werden."

Allerdings waren bisher keine operationalisierbaren Kriterien verfügbar, an Hand derer die Anonymität der Daten konkret festgestellt werden kann. Diese Unsicherheit führte zu einer restriktiven Praxis der Weitergabe von statistischen Einzelangaben. Daher wurde im Bereich der Wissenschaft mehrfach die Forderung erhoben, die datenschutzrechtlichen Anforderungen für eine Weitergabe von Datenbeständen mit statistischen Einzelangaben (Mikrodatenfiles (MDF)) zu klären und zu konkretisieren.[3])

---

[1]) Ein Überblick über den Entwicklungsstand und das Anwendungsspektrum derartiger Modelle findet sich in: G. Orcutt, J. Merz, H. Quinke, Microanalytic Simulation Models to Support Social and Financial Policy, Amsterdam: North Holland.
[2]) Gesetz über die Statistik für Bundeszwecke (BStatG) vom 14. März 1980.
[3]) Vgl. M. Kaase, H. J. Krupp, M. Pflanz, E. K. Scheuch und S. Simitis (Hrsg. 1980), Datenzugang und Datenschutz – Konsequenzen für die Forschung, Königstein/Taunus.

Um die Konstruktion mikroanalytischer Simulationsmodelle (Mikromodelle) zu ermöglichen hat die GMD ein Forschungsprojekt AIMIPH [4]) durchgeführt zur Prüfung, ob und unter welchen Umständen die Datensätze eines MDF nicht mehr den zugehörigen Personen zuzuordnen sind. In diesem Bericht werden einige Ergebnisse des Projektes vorgestellt. Im zweiten Abschnitt wird ein mathematisch-statistisches Verfahren zur Abschätzung des Re-Identifikationsrisikos an Hand eines Beispiels erläutert. Im dritten Abschnitt wird unter realistischen Randbedingungen das Re-Identifikationsrisiko realer MDF bestimmt. Der letzte Abschnitt enthält eine kurze Zusammenfassung und Diskussion der Ergebnisse. [5])

1 Abschätzung des Re-Identifikationsrisikos

Ein Datensatz des MDF ist re-identifizierbar im Sinne des Bundesstatistikgesetzes, wenn er dem Auskunftspflichtigen zugeordnet werden kann. Beispielsweise kann ein Angreifer die ihm bekannten Merkmale einer Person – der Zielperson – mit den im MDF erfaßten Merkmalen vergleichen. Kann er so eine eindeutige Zuordnung zu einem der Datensätze erreichen, so ist der Personenbezug wiederhergestellt und der Datensatz re-identifiziert. Der Angreifer kann auf diese Weise Kenntnis von weiteren Merkmalen der Zielperson im Datensatz erhalten. Ist die Zuordnung des Datensatzes zur Zielperson nur mit einer gewissen Wahrscheinlichkeit eindeutig, so besteht nur ein entsprechendes Re-Identifikationsrisiko für diesen Datensatz.

Die weitere Darstellung geht von folgender Ausgangssituation aus:

- Das MDF ist eine Menge einzelner Datensätze $y_1, \ldots, y_m$.
- In der Regel ist das MDF eine Stichprobe aus einer größeren Grundgesamtheit mit N Elementen. Der Auswahlsatz der Stichprobe ist m/N.
- Der Angreifer möchte im Rahmen einer gezielten Suche weitere Daten über eine einzelne Zielperson aus dem MDF erhalten. Das „Zusatzwissen" des Angreifers bestehe damit allein aus dem Datensatz z_1 der Zielperson.

Beispiel für ein im Projekt näher untersuchtes MDF ist die Einkommens- und Verbrauchsstichprobe (EVS). Sie weist für etwa 50000 Privathaushalte mehrere hundert Merkmale zu Einkommen, Verbrauch, Vermögen, Wohnung etc. nach. Sie ist eine Stichprobe aus den etwa 22 Millionen deutschen Privathaushalten und enthält damit etwa jeden 500. Haushalt.

---

[4]) „Konstruktion und Erprobung eines anonymisierten integrierten Mikrodatenfiles der bundesdeutschen Privathaushalte". Die diesem Bericht zugrundeliegenden Arbeiten wurden mit Mitteln des Bundesministeriums für Forschung und Technologie (Förderungskennzeichen IT 3053) gefördert. Die Verantwortung für den Inhalt liegt jedoch allein beim Autor.

[5]) Eine ausführliche Darstellung des Projektes findet sich in: G. Paaß, und U. Wauschkuhn (1985), Datenzugang, Datenschutz und Anonymisierung – Analysepotential und Identifizierbarkeit von anonymisierten Individualdaten, München: Oldenbourg Verlag. Ein methodisch orientierter Überblick wird gegeben in: G. Paaß, (1986), Identifikationsrisiko und Anonymisierbarkeit von Mikrodatenfiles. Allgemeines Statistisches Archiv, (erscheint demnächst).

Tabelle 1: Beispiel für eine Re-Identifikation

	Datensatz der Zielperson ...	
	im Zusatzwissen	im MDF
Identifikations-Variable		
Name	Markus Müller	?
Adresse	Wahlscheid	?
Straße	Kirchstr. 33	?
Gemeinsame Variablen		
Einkommen	45000	45053
Konsumausgaben	27500	29300
Nutzmerkmale		
Schulden	?	90000 DM
Krankheiten	?	Leberschaden

Tabelle 1 verdeutlicht die Vorgehensweise des Angreifers bei einer Re-Identifikation. Der Angreifer kennt einige Merkmale der Zielperson, welche auch im MDF vorkommen. Im Beispiel handelt es sich um das Einkommen und den Konsum. Er kann nun die Merkmalswerte der Datensätze im MDF mit den entsprechenden Werten der Zielperson vergleichen. Hierbei muß er berücksichtigen, daß sowohl bei der Erhebung des MDF als auch bei der Sammlung des Zusatzwissens Erhebungsfehler auftreten können, welche eine Abweichung zwischen den Datensätzen der gleichen Zielperson zur Folge haben können. In Tabelle 1 ist beispielsweise das Einkommen in beiden Datenbeständen nahezu korrekt erfaßt, die Angaben zum Konsum unterscheiden sich hingegen relativ stark. Gibt es nun einen Datensatz im MDF, der zu den Angaben des Zusatzwissens „paßt" und kann der Angreifer mit hoher Sicherheit ausschließen, daß dieser Datensatz zu einer anderen Person der Bevölkerung gehört, so ist der Datensatz mit der entsprechenden Sicherheit re-identifiziert. In einem der folgenden Abschnitte wird untersucht, unter welchen Umständen die konkreten Angaben in Tabelle 1 zu einer Re-Identifikation führen können.

Die Sicherheit, mit der eine Re-Identifikation möglich ist, hängt von einer Reihe von Randbedingungen ab. Offenbar steigt das Re-Identifikationsrisiko mit der Anzahl der gemeinsamen Merkmale, denn bei einer höheren Zahl dieser Merkmale können potentielle „Doppelgänger" mit ähnlichen Merkmalen eher ausgeschlossen werden. Darüberhinaus spielt nicht nur die Anzahl der gemeinsamen Merkmale sondern der Informationsgehalt – die Aussagekraft der Merkmalswerte – eine entscheidende Rolle, da z. B. eine Kinderzahl von 8 eher zu einer Re-Identifikation führt als eine Kinderzahl von 2. Das Risiko wird geringer, wenn nur eine Stichprobe mit kleinem Auswahlsatz zur Verfügung steht, da dann die Anzahl der potentiellen Doppelgänger steigt. Schließlich sinkt das Re-Identifikationsrisiko, falls sich das Ausmaß der Erhebungsfehler vergrößert, da dann potentiell mehr Datensätze des MDF für eine Re-Identifikation in Frage kommen.

Erhebungsfehler können durch falsche Angaben der Befragten oder durch Übermittlungs- und Kodierfehler verursacht werden und können sowohl im MDF als auch im Zusatzwissen

vorhanden sein. Bei der Re-Identifikation von Datensätzen haben außerdem Faktoren, welche einen Unterschied zwischen den Variablenwerten einer Person im Zusatzwissen und im MDF zur Folge haben, die gleiche Wirkung wie Erhebungsfehler und werden daher diesen zugeschlagen. Hierzu gehören Differenzen auf Grund unterschiedlicher Variablendefinitionen oder Erhebungszeitpunkte sowie Unterschiede infolge etwaiger Maßnahmen zur Anonymisierung des MDF.

Die Re-Identifikationsmöglichkeiten bei Erhebungsfehlern oder Stichprobenziehung sollen im Rahmen des obigen Beispiels verdeutlicht werden. Das MDF enthalte die Angaben von vier Personen, deren Wertekombinationen der gemeinsamen Variablen in Abbildung 2 durch *A*, *B*, *C* und *D* markiert sind. Infolge von Erhebungsfehlern können die zugehörigen Angaben im Zusatzwissen um die MDF-Werte streuen. Daher sind in Abbildung 2 exemplarisch einige mögliche fehlerbehaftete Werte des Zusatzwissens eingezeichnet, die je nach der Zugehörigkeit zu einem Datensatz des MDF durch *a*, *b*, *c* oder *d* gekennzeichnet sind.[6]) Der Datensatz z_1 der Zielperson ist einer dieser möglichen Werte.

Das Ausmaß der Streuung der Erhebungsfehler wird in der Abbildung charakterisiert durch eine Streuungsellipse, welche im Mittel gerade 50% der fehlerhaften Werte enthält. Überlappen sich die Streuungsellipsen (wie etwa bei den Datensätzen *B* und *C*), so ist, falls z_1 in diesem Überlappungsbereich liegt, eine sichere Zuordnung von z_1 durch den Angreifer nicht möglich, da z_1 von zwei oder mehreren Datensätzen des MDF stammen könnte. Befindet sich hingegen in der „Umgebung" von z_1 nur ein einziger Datensatz des MDF, so kann z_1 diesem Datensatz mit einiger Sicherheit zugeordnet werden, wie etwa bei *A* in unserem Beispiel. In jedem Fall aber bleibt noch eine geringe Wahrscheinlichkeit, daß z_1 infolge von Erhebungsfehlern großen Ausmaßes zu einem anderen Datensatz gehört, so daß eine vollständig sichere Re-Identifikation unmöglich ist.

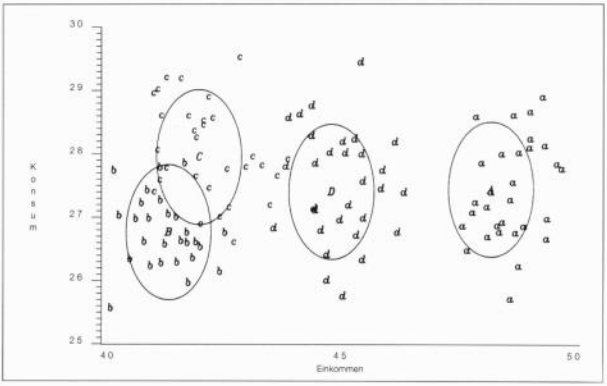
Abbildung 3 zeigt für das oben diskutierte Beispiel die durch * bezeichnete mögliche Position derartiger nicht im MDF erfaßten Personen der Grundgesamtheit.[7]) Ist in der Umgebung eines MDF-Datensatzes mit hoher Wahrscheinlichkeit ein solcher Datensatz der Grundgesamtheit enthalten, so verhindern, wie oben diskutiert, die Erhebungsfehler die Re-Identifikation einer Zielperson z_1 in diesem Bereich. Offenbar sind die Chancen für eine eindeutige Zuordnung um so höher, je mehr Personen aus der Grundgesamtheit in der Stichprobe erfaßt sind.

Für eine Re-Identifikation muß der Angreifer zunächst eine Vorstellung von dem Ausmaß und der Struktur der Abweichungen zwischen den Datensätzen des MDF und des Zusatzwissens entwickeln und diese durch eine Verteilung der Abweichungen charakterisieren. Für einen beliebigen Datensatz y_i des MDF beschreibt diese Fehlerverteilung die Wahrscheinlichkeit, mit der ein Variablenvektor z aus dem Datensatz y_i hervorgehen kann. In Abbildung 2 sind die Fehlerverteilungen der einzelnen Datensätze des MDF durch Streuungsellipsen charakterisiert. Sind keine Erhebungsfehler vorhanden, so ordnet die Fehlerverteilung sämtlichen Abweichungen den Wahrscheinlichkeitswert 0 zu.

---

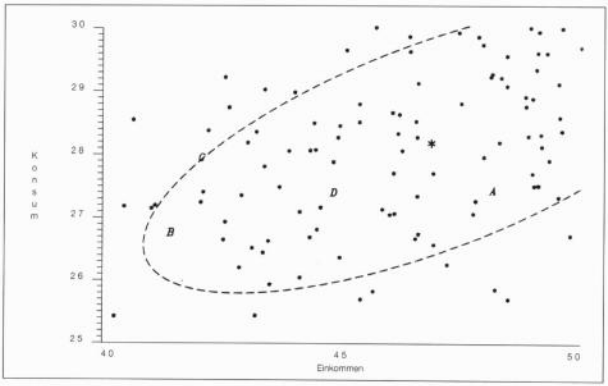
[6]) Es wurden in diesem Beispiel unabhängige und normalverteilte Erhebungsfehler angenommen.
[7]) Die Grundgesamtheit wurde als normalverteilt angenommen und eine Korrelation beider Variablen von 0.9 unterstellt.

Abbildung 2: Datensätze des MDF und zugehörige Verteilungen der Erhebungsfehler



A, B, C, D: Datensätze des MDF. ——: Streuungselipse der Fehler.
a, b, c, d: durch Erhebungsfehler aus *A, B, C* bzw. *D* hervorgegangen.

Abbildung 3: Verteilung der Datensätze in der Grundgesamtheit



A, B, C, D: Datensätze des MDF
★: weitere potentielle Datensätze der Grundgesamtheit.
×: Mittelwert der Grundgesamtheit. —— ——: zugehörige Streuungselipse.

Für einen Re-Identifikationsversuch stehen dem Angreifer folgende Informationen zur Verfügung:

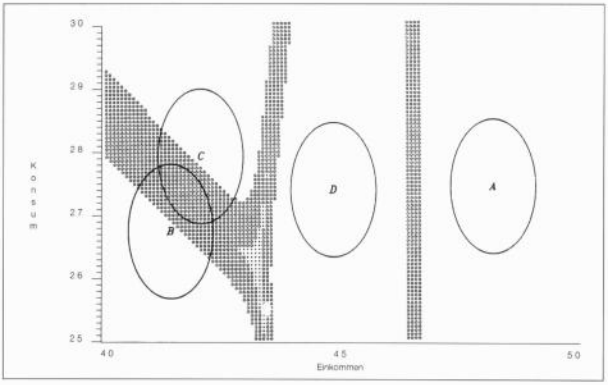
- Die Datensätze y_i des MDF.
- Der Auswahlsatz des MDF.
- Der Datensatz z_1 des Zusatzwissens.
- Eine ungefähre Kenntnis der Verteilung der Abweichungen.

Es ist nicht möglich, all diese Faktoren in eine Formel einzubringen und das Re-Identifikationsrisiko analytisch zu bestimmen. Vielmehr wurden alle Randbedingungen in plausiblen Szenarios festgelegt. Unter deren Annahme kann man mit Hilfe eines im Projekt AIMIPH entwickelten Verfahrens denjenigen Datensatz des MDF mit der höchsten Wahrscheinlichkeit der Zugehörigkeit zur Zielperson bestimmen. Diese Zuordnungswahrscheinlichkeit $p_{max} = p_{max}(z_1)$ berücksichtigt sämtliche dem Angreifer zur Verfügung stehenden Informationen. Der Angreifer wird eine Zuordnung je nach dem geplanten Verwendungszweck der Nutzmerkmale erst als endgültig ansehen, wenn die Zuordnungswahrscheinlichkeit eine gewisse Schwelle p_0 (z. B. $p_0 = 0.90$) überschreitet, denn dann ist das Risiko einer Fehlzuordnung durch $1 - p_0$ begrenzt. Die Zuordnungswahrscheinlichkeit ist – vom Gesichtspunkt des Datenschutzes – das Re-Identifikationsrisiko der Zielperson unter den betrachteten Randbedingungen.

Abbildung 4 verdeutlicht die Höhe der Zuordnungswahrscheinlichkeit p_{max} in unserem Beispiel bei Erhebungsfehlern unter der Annahme, daß das MDF sämtliche 4 Personen der Grundgesamtheit umfaßt. Die Bereiche, in denen p_{max} unter 0.50 bzw. zwischen 0.50 und 0.80 liegt, sind durch „o" bzw. „." gekennzeichnet. In den restlichen Flächen liegt ein hohes Re-Identifikationsrisiko von über 0.80 vor. Zur Orientierung wurden die Positionen der MDF-Datensätze und die zugehörigen Streuungsellipsen der Erhebungsfehler eingezeichnet. Wie zu erwarten, ist das Re-Identifikationsrisiko gerade in den Überlappungsbereichen niedrig. Trotz der Erhebungsfehler ist aber die Region, in der eine Re-Identifikation mit hoher Sicherheit noch möglich ist, relativ groß. In Abbildung 5 ist p_{max} für den gleichen Sachverhalt in einer dreidimensionalen Darstellung wiedergegeben. Die Überlappungsbereiche erscheinen hier als scharfe Einschnitte, während die Regionen, in denen eine Zuordnung möglich ist, ein „Plateau" mit Werten nahe bei 1 bilden.

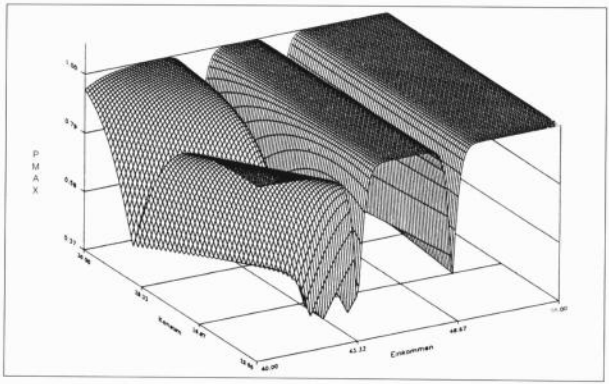
Ist das MDF eine Stichprobe mit einem Auswahlsatz von beispielsweise ½, und liegt für eine Zielperson keine Vorabinformation darüber vor, ob ihre Daten im MDF enthalten sind, so wird der ihr zugehörige Datensatz im Mittel nur in jedem zweiten Fall tatsächlich in der Stichprobe vorhanden sein, weshalb im Mittel nur höchstens jede zweite gezielte Suche erfolgreich sein kann. Der Effekt der Stichprobeneigenschaft soll im Rahmen des obigen Beispiels demonstriert werden. Hierbei wurde angenommen, daß die Grundgesamtheit insgesamt acht Personen umfaßt (Auswahlsatz ½). Die Ergebnisse sind in Abbildung 6 dargestellt. Der Bereich, in dem eine eindeutige Re-Identifikation nicht mehr möglich ist, hat sich infolge der Stichprobeneigenschaft stark vergrößert. Lediglich in relativ kleinen Gebieten in der Nähe der erfaßten Datensätze *A*, *B*, *C* und *D* des MDF kann eine Re-Identifikation noch mit ausreichender Sicherheit erfolgen. Abbildung 7 zeigt eine dreidimensionale Darstellung

Abbildung 4: Re-Identifikationsrisiko bei Erhebungsfehler



A, B, C, D : Datensätze des MDF. ___: Streuungsellipse der Fehler.
„o" bzw. „+": Re-Identifikationsrisiko $\underline{p}_{max}(z) < 0.50$ bzw. $0.50 \leq \underline{p}_{max}(z) < 0.80$.

Abbildung 5: Risikofunktion bei Erhebungsfehlern



der Risikofunktion für die gleiche Situation. Hier wird deutlich, daß das mittlere Niveau des Re-Identifikationsrisikos gegenüber Abbildung 5 wesentlich geringer ist. Die Berücksichtigung nicht erfaßter Personen der Grundgesamtheit führt zudem zu einer „Glättung" der Risikofunktion, da die genauen Wertekombinationen dieser Personen nicht bekannt sind.

In unserem Beispiel sinkt das Re-Identifikationsrisiko stark ab, wenn das MDF eine 50%-Stichprobe aus der Grundgesamtheit ist. Diese Verminderung des Risikos kann aber durch eine Vergrößerung der Anzahl der gemeinsamen Variablen – und damit deren Informationsgehalt - ausgeglichen werden. Denn falls von einer Person eine größere Zahl von Variablen bekannt ist, besitzt sie eine umfangreiche „Umgebung" von Wertekombinationen, die „näher" zu ihr liegen als zu jeder anderen Person der Grundgesamtheit. Selbst bei hohen Erhebungsfehlern wird dann der größte Teil der „Fehlerwerte" innerhalb dieser Umgebung liegen, so daß eine Re-Identifikation möglich bleibt.

Die meisten Erhebungen umfassen gleichzeitig sowohl diskrete Merkmale, z. B. Familienstand und Zahl der Kinder, als auch kontinuierliche Angaben, wie Einkommen oder Lohnsteuer, zwischen denen eine Vielzahl komplexer Abhängigkeiten und Zusammenhänge bestehen. Im Gegensatz zu der in den Beispielen untersuchten Situation ist daher in der Praxis die Verteilung der Grundgesamtheit nicht bekannt und nicht einmal in geschlossener Form darstellbar. Dieses auf die Stichprobeneigenschaft zurückzuführende Problem stellt die eigentliche Schwierigkeit bei der Bstimmung des Re-Identifikationsrisikos dar. Das in der GMD entwickelte Verfahren zur Berechnung des Re-Identifikationsrisikos benutzt daher einen „nichtparametrischen" Ansatz. Es beruht auf einer Reihe von Annahmen. Wie Testrechnungen zeigen, liefert es trotz der Annahmen und der unvermeidlichen Schätzunsicherheiten bei der Bestimmung des Re-Identifikationsrisikos zuverlässige Ergebnisse, da insbesondere bei vielen gemeinsamen Merkmalen die Merkmalsausprägungen einen hohen Grad an Information über die Zugehörigkeit der Datensätze zueinander enthalten, so daß sich die potentiellen Fehlerquellen nicht auswirken.

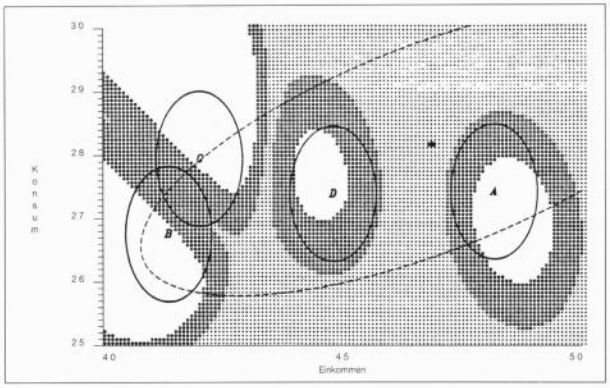
2 Ergebnis der Re-Identifikationsexperimente

Im Projekt AIMIPH wurden Re-Identifikationsrisiken exemplarisch für zwei reale umfangreiche Datenbestände evaluiert, die EVS 1978 und der Mikrozensus (MZ) 1978, welche viele der in sozialwissenschaftlichen und ökonomischen Untersuchungen benötigten Variablen enthalten.[8])

Die Gefährdung dieser Datenbestände wurde in „Re-Identifikationsexperimenten" bestimmt. Hierzu wurden zunächst verschiedene plausible Szenarien spezifiziert, in denen die relevanten Randbedingungen (Art und Motiv des Datenangriffs, Variable des Zusatzwissens, Ausmaß der Erhebungsfehler) festgelegt wurden. Da es sich von selbst verbot, die Teilnehmer an den Erhebungen tatsächlich zu re-identifizieren, wurden entsprechend der angenommenen Verteilung der Erhebungsfehler für jedes Szenario 100 synthetische Ziel-

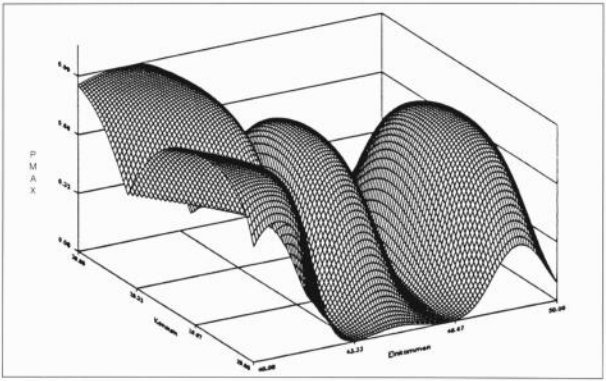
[8]) Teilstichproben der Datenbestände mit eingeschränktem Variablenumfang wurden der GMD in formal anonymisierter Form ausschließlich für die Zwecke des Projektes AIMIPH vom Statistischen Bundesamt zur Verfügung gestellt.

Abbildung 6: Re-Identifikationsrisiko bei Erhebungsfehlern und Stichprobeneigenschaft



A, B, C, D: Datensätze des MDF. ——: Streuungsellipse der Fehler.
★: Mittelwert der Grundgesamtheit. – – –: zugeh. Streuungsellipse.
„o" bzw. „.": Re-Identifikationsrisiko $p_{max}(z) < 0.50$ bzw. $0.50 \leq p_{max}(z) < 0.80$.

Abbildung 7: Risikofunktion bei Erhebungsfehlern und Stichprobeneigenschaft



personen aus den Datensätzen des MDF erzeugt. Für jede dieser Zielpersonen wurde – aus der Sicht eines potentiellen Angreifers – das Re-Identifikationsrisiko geschätzt. Hierbei wurde auch die Situation simuliert, daß der zur Zielperson gehörige Datensatz nicht im MDF enthalten war. Da die tatsächliche Zugehörigkeit der Datensätze bekannt war, konnte nach Abschluß der Experimente die Anzahl der korrekt re-identifizierten, der falsch zugeordneten und der nicht re-identifizierten Datensätze des MDF bestimmt werden.

Es wurden sechs Szenarien untersucht, welche das Spektrum der plausiblen Angriffssituationen hinsichtlich Anzahl der gemeinsamen Merkmale, Art des Datenangriffs und der Höhe der Erhebungsfehler abdecken sollten. Das Re-Identifikationsverfahren erwies sich als bemerkenswert robust, so daß es fast gar nicht zu Fehlzuordnungen kam.

In Tabelle 8 werden die wichtigsten Ergebnisse der Re-Identifikationsexperimente wiedergegeben. Hierbei wurde der Fall untersucht, daß die Erhebungsfehler „Normalniveau" besitzen und dem Angreifer nicht bekannt ist, ob die gesuchte Person im MDF erfaßt ist. Nimmt man die Anzahl der gemeinsamen Variablen als eine grobe Maßzahl für den Informationsgehalt der gemeinsamen Variablen, so ist dieser Informationsgehalt die wesentliche Determinante für den Anteil der re-identifizierbaren Datensätze im MDF. Bei wenigen gemeinsamen Merkmalen, d. h. niedrigem Informationsgehalt, können überhaupt keine Datensätze re-identifiziert werden. Dies ist beispielsweise bei dem Adreßverlags- und Konzern-Szenario der Fall. Bei dem höheren Informationsgehalt der gemeinsamen Variablen im Staatsanwalt-, Steuerfahndungs- und Kripo-Szenario kann trotz der Erhebungsfehler ein Anteil von etwa 10% bis zu 60% der Datensätze des MDF korrekt re-identifiziert werden. Damit ist unter diesen Umständen ein Großteil der Datensätze des MDF re-identifikationsgefährdet.

Trotzdem sind gemäß der rechten Spalte der Tabelle 8 die Erfolgschancen des Angreifers nur gering. Wegen der Stichprobeneigenschaft ist bei 10000 zufällig aus der Grundgesamtheit ausgewählten Zielpersonen der zugehörige Datensatz nur in 20 Fällen in der EVS vorhanden, so daß unter diesen Umständen lediglich 1 bis maximal 11 von 10000 Re-Identifikationsversuchen zum Erfolg führen. Die Stichprobeneigenschaft bildet daher bei umfangreichem Zusatzwissen den hauptsächlichen Schutzfaktor vor Re-Identifikationen.

Tabelle 8: Ergebnis der Re-Identifikationsexperimente*)

Szenario (Anzahl der gemeinsamen Variablen)	Prozentsatz der ... re-identifizierbaren Datensätze im MDF	Prozentsatz der ... erfolgreichen Re-Identifikationsversuche
Staatsanwalt ..... (68)	56	0.11
Steuerfahndung ..... (45)	63	0.11
Kripo ..... (15)	7	0.01
Adreßverlag ..... (7)	0	0.00
Konzern ..... (11)	0	0.00

*) Sicherheitsschwelle p_0 = 99.9%. Auswahlsätze: EVS 0.2%, MZ 1%. Gezielte Suche, Zusatzwissen: 1 Datensatz z_t.

Besteht das Zusatzwissen aus einer großen Datenbank, die einen hohen Anteil der Gesamtbevölkerung erfaßt (z. B. die Datei mit den Einkommensteuererklärungen von Nordrhein-Westfalen mit 25% der Bundesbürger), so kann die Suchrichtung umgekehrt werden. Der Angreifer geht von einer „Zielperson" im MDF aus, und versucht, dieser den zugehörigen Datensatz aus dem Zusatzwissen zuzuordnen. Die Modellrechnungen für die Datenbank der Steuererklärungen zeigen, daß bei einem solchen Massenfischzug bis zu 17% der Re-Identifikationsversuche erfolgreich sind. Damit können derartig umfangreiche Datenbänke des Zusatzwissens eine reale Bedrohung für die Anonymität der Datensätze beinhalten.

Zusammenfassung und Diskussion

Ausgangspunkt der Untersuchung war die bestehende Unsicherheit, unter welchen Umständen statistische Einzelangaben ohne Re-Identifikationsmerkmale, wie Name und Adresse, als anonym betrachtet werden können. Im Rahmen des Projektes AIMIPH wurde hierzu ein Verfahren entwickelt, welches unter gegebenen konkreten Randbedingungen die Einstufung einzelner vorgegebener Datensätze als re-identifikationsgefährdet oder sicher erlaubt. Hierbei können mögliche Erhebungsfehler in den Daten, der Effekt von Anonymisierungsmaßnahmen sowie die Stichprobeneigenschaft des MDF berücksichtigt werden. Durch die Hochrechnung einer Reihe von Einzelergebnissen auf den gesamten Datenbestand läßt sich der Prozentsatz der re-identifikationsgefährdeten Datensätze des Datenbestandes abschätzen. Die Re-Identifikationsexperimente für reale MDF zeigen, daß bei umfangreichem Zusatzwissen eine Re-Identifikation von Datensätzen selbst bei Erhebungsfehlern und in Stichproben mit sehr geringem Auswahlsatz möglich ist. Zwar ist in Stichproben die Chance gering, daß der in Frage kommende Datensatz im MDF enthalten ist. Ist dies allerdings der Fall, so ist bei umfangreichem Zusatzwissen eine Re-Identifikation mit hoher Sicherheit möglich.

Im Verlauf des Projektes AIMIPH wurde daraufhin geprüft, ob durch eine Modifikation der Merkmalswerte des MDF eine Re-Identifikation der Datensätze verhindert werden kann. Bei der Untersuchung derartiger Anonymisierungsverfahren zeigte sich, daß dieses Vorgehen nur bei MDF mit relativ wenigen Merkmalen erfolgversprechend ist. Denn selbst bei sehr aufwendigen Verfahren hatten die Modifikationen starke Verzerrungen des multivariaten statistischen Zusammenhangs zur Folge, welche die MDF für komplexere Auswertungen unbrauchbar machten.[9])

Damit hängt die Möglichkeit der Freigabe von Datenbeständen mit Einzelangaben als Public Use Files im wesentlichen vom Informationsgehalt des Zusatzwissens ab:

- Ist der Informationsgehalt des Zusatzwissens gering (in den untersuchten Szenarien bei weniger als 15 Merkmalen), so kann ein Datenbestand ohne Auflagen als Public Use File freigegeben werden.

[9]) Paaß, Wauschkuhn (1985), a. a. O., S. 209–278

- Bei mittlerem Informationsgehalt (in den untersuchten Szenarien bei 15 bis etwa 30 Merkmalen) kann das Re-Identifikationsrisiko durch begrenzte Anonymisierungsmaßnahmen bei einzelnen gefährdeten Datensätzen beseitigt werden ohne daß die Qualität der Daten zu stark beeinträchtigt wird.

- Bei einem hohen Informationsgehalt des Zusatzwissens ist eine ausreichende Anonymisierung mit einer starken Beeinträchtigung des Analysepotentials verbunden. Daher können unter den derzeitigen Weitergaberegelungen umfangreiche MDF nicht in befriedigender Qualität als Public Use Files verfügbar gemacht werden, da sich bei einer Weitergabe an einen uneingeschränkten Empfängerkreis ein hohes Zusatzwissen nicht ausschließen läßt.

Um auch MDF mit vielen Merkmalen der Wissenschaft zugänglich machen zu können, muß eine Verknüpfung der MDF mit solchen Datenbanken verhindert werden, die umfangreiches potentielles Zusatzwissen enthalten.

Als Ausweg bietet sich an, den Empfängerkreis auf Einrichtungen der unabhängigen wissenschaftlichen Forschung zu beschränken. Bei diesen ist – auch nach Einschätzung des Bundesverfassungsgerichts – regelmäßig kaum Zusatzwissen vorhanden.[10]) Durch zusätzliche Auflagen, wie ein Verbot der Weitergabe und geeignete Zugangssicherungen, könnte die Abschottung der Mikrodaten von umfangreichem Zusatzwissen gewährleistet werden. Unter diesen Umständen wäre bei den Datenempfängern allenfalls von einem systematischen Zusatzwissen vom Informationsgehalt des Adreß- oder Konzernszenarios auszugehen, bei welchem unter den getroffenen Annahmen eine Re-Identifikation nicht möglich war. Der geringen Chance von Zufallstreffern mit dem Zusatzwissen einzelner Nachbarn oder Verwandten ließe sich durch ein ausdrückliches Re-Identifizierungsverbot begegnen. Es besteht Aussicht, daß derartige Überlegungen bei der anstehenden Novellierung des Bundesstatistikgesetzes und des Datenschutzgesetzes berücksichtigt werden.

[10]) Bundesverfassungsgericht, 1983: Urteil des Ersten Senats vom 15. Dezember 1983 – 1BvR 209/83 u. a. – Volkszählungsgesetz teilweise verfassungswidrig, Europäische Grundrechte Zeitschrift, S. 595.

Dr. Joachim Kühn*)
Statistisches Bundesamt, Wiesbaden

Automatisierte Anonymisierungsverfahren für Kurzbandsätze

Einführung

Viele Aufgabenstellungen, die die Nutzer der amtlichen Statistik bearbeiten, erfordern weitgehend die Auswertung von Individualmaterial. Die gesetzliche Lage sowie das bei der Bevölkerung entwickelte Datenschutzbewußtsein erfordern seitens der amtlichen Statistik in viel höherem Maße eine Garantie der statistischen Geheimhaltung, als es noch vor einiger Zeit denkbar war. Eine Weitergabe personenbezogener amtlicher statistischer Daten ohne Namen und Anschrift und ohne weitere anonymisierenden Maßnahmen ist z. Z. undenkbar. Alle anonymisierende Maßnahmen führen zu einer mehr oder weniger starken Einschränkung des Analysepotentials des anonymisierten Materials im Vergleich zum Ausgangsmaterial. Es liegt auf der Hand, daß das Anonymisierungsproblem um so größer wird, je mehr Merkmale und je mehr Merkmalsausprägungen je Person nachgewiesen werden müssen. Bereits einfache Tabellierungen zeigen, daß im Mikrozensus etwa 30% der Personensätze einzig sind, die bei einer Anonymisierung auf jeden Fall „modifiziert" werden müßten. Bei vielen Auswertungswünschen der Nutzer der amtlichen Statistik werden nur ausgewählte Merkmale einer statistischen Erhebung benötigt. Es ist klar, daß bei wenigen Merkmalen und bei wenigen Merkmalsausprägungen die Möglichkeiten der Anonymisierung eines solchen Teilmaterials bei zumutbarem Informationsverlust steigen. In vorliegendem Papier wird über Experimente zur Anonymisierung von solchen sogenannten Kurzbandsätzen berichtet.

1 Einige Hinweise zu dem vorzustellenden Anonymisierungsverfahren

Bei verschiedenen Statistiken werden Merkmale von natürlichen Personen erhoben. Es wird hier von der Vorstellung ausgegangen, daß jeder Person ein Satz von Merkmalsausprägungen entspricht. Je nach Anzahl der Merkmale sowie nach Zahl der vorkommenden Merkmalsausprägungen sind in einem solchen Material mehr oder weniger Sätze mehrfach vorhanden. Ziel eines sinnvollen Anonymisierungsverfahrens ist es, Einzelsätze so zu verändern, daß in dem dann entstehenden anonymisierten Material kein Satz einzig bleibt, daß aber der statistische Informationsgehalt möglichst wenig verfälscht wird.

Der Grundgedanke des Verfahrens besteht darin, daß man in einem ersten Schritt das zu anonymisierende Material zu Gruppen zusammenfaßt, deren Individuen eine „verwandte"

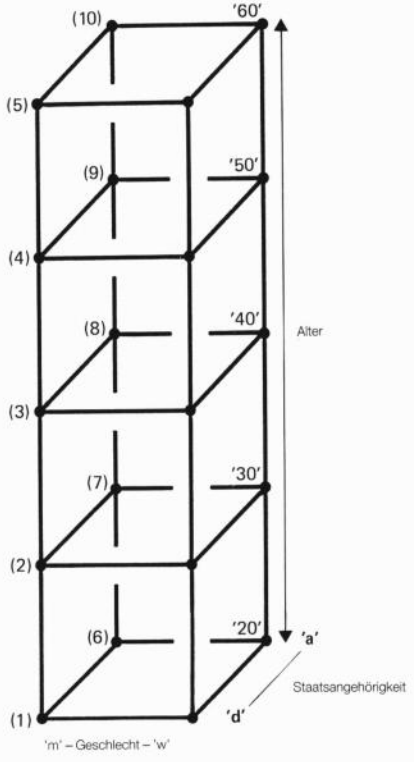
*) Unter Mitarbeit von Michael Radtke.

Struktur haben. Im zweiten Schritt ersetzt man diese Gruppen durch einen charakteristischen (evtl. künstlichen) Repräsentanten, der die gesamte Masse der betreffenden Gruppe enthält. Der Grundgedanke entspricht also dem Vorgehen, das man bei der Konstruktion von sogenannten Tripelaggregaten wählt. Das Hauptproblem bei derartigen Konstruktionen besteht hier naturgemäß in der Wahl eines geeigneten Abstandes, durch den ja die Umgebung eines Satzes bzw. die Verwandtschaft zwischen Sätzen definiert wird. Ein automatisch gut durchzuführendes Verfahren erhält man, wenn man die Verwandtschaft durch Gleichheit in ausgewählten Merkmalen definiert; „benachbarte Sätze" werden dann durch Sortierung des Materials nach diesen Merkmalen zusammengeführt. Die Sortierfolge ist hierbei durch die Bedeutung der einzelnen Merkmale für die geplante Auswertung gegeben. Dasjenige Merkmal, dessen Verteilung durch die Anonymisierungsmaßnahmen am wenigsten gestört werden darf, muß oberster Sortierbegriff werden. Die Problematik des durch die Sortierung gewonnenen Nachbarschaftsbegriffs erkennt man leicht aus Abbildung 1, Seite 103.

Die drei Merkmale Geschlecht, Staatsangehörigkeit (mit zwei Ausprägungen), Altersgruppen (mit fünf Ausprägungen) sind jeweils rechtwinklig zueinander aufgetragen. Oberster Sortierbegriff ist das Geschlecht, zweiter Sortierbegriff die Staatsangehörigkeit und letzter Sortierbegriff das Alter. Die möglichen, durch die Sortierung gegebenen Positionen sind durchnumeriert, und man erkennt klar, daß z. B. Position 7 nicht zur Nachbarschaft der Position 2 gehört, obwohl für eine Auswertung, bei der die Staatsangehörigkeit von untergeordneter Bedeutung ist, die beiden Positionen benachbart sein müßten. Natürlich kann man sich durch eine dem Auswertungsziel besser angepaßte Sortierung helfen.

Für eine gewählte Sortierung ist das weitere Vorgehen folgendermaßen: Unter der in dem Material entstandenen Reihenfolge werden zunächst gleiche Sätze zusammengefaßt und durch einen entsprechenden Fallzähler gekennzeichnet. In einem zweiten Schritt werden dann „benachbarte" Sätze zufällig aggregiert, so daß alle nun entstehenden Fallzähler größer oder gleich einer vorzugebenden Grenzzahl K werden. Dies erreicht man, indem man sukzessiv bei dem nach dem ersten Schritt voraggregierten Material bei benachbarten Einzelsätzen die „unwichtigen" Merkmale durch geeignete Zufallsexperimente anpaßt, bis eine Aggregation von jeweils mindestens K Sätzen möglich wird. Es liegt auf der Hand, daß man den entsprechenden Algorithmus so aufbaut, daß jeweils benachbarte Gruppen von Sätzen bearbeitet werden, bei denen die Summe der Fallzähler nach dem ersten Schritt die jeweils kleinstmögliche Zahl größer als K ergibt.

Abbildung 1



( ) = hierarchische Position

Statistisches Bundesamt 86 0917

2 Experimente zur Anonymisierung eines Kurzbandsatzes aus dem Mikrozensus 1982

Es wurde versucht, einen anonymisierten Kurzbandsatz aus dem Mikrozensus 1982 mit den folgenden Merkmalen zu erstellen:

Merkmal	Zahl der Ausprägungen
Geschlecht	(2)
Altersgruppe	(33)
Staatsangehörigkeit	(2)
Erwerbstätigkeit in der Berichtswoche	(4)
Wirtschaftszweig	(97)
Beruf	(335)
Stellung im Beruf	(11)
Normal geleistete Arbeitszeit	(12)
Nettoeinkommen	(14)
Tätigkeitsmerkmale	
überwiegend ausgeübte Tätigkeit	(11)
vorwiegender Arbeitsplatz	(11)
Stellung im Betrieb	(11)
Allgemeinbildender Schulabschluß	(6)
Abgeschlossene berufsbildende Schule	(8)

Es wurden nun zum Vergleich drei Sortierreihenfolgen dadurch gewählt, daß einmal die oben angegebene Reihenfolge der Merkmale als Sortierreihenfolge gewählt wurde (Version A), daß zweitens die Sortierreihenfolge dadurch verändert wurde, daß der Beruf als drittletzter Sortierbegriff gewählt wurde (Version B) und daß schließlich der Beruf oberster Sortierbegriff wurde (Version C).

Zur Bewertung des Informationsgehaltes der drei anonymisierten Materialien wurden diese mit dem Originalmaterial mit folgenden Ergebnissen verglichen:

– Bei Anpassungstests der jeweiligen Verteilungen der Einzelmerkmale beim anonymisierten und beim Originalmaterial ergab sich, daß nur in ganz wenigen Fällen die Annahme der Gleichheit der Verteilung verworfen werden konnte.

Die folgende Tabelle zeigt die Ergebnisse für einige Randverteilungen.

Anpassungstests für Randverteilungen

Merkmal	Freiheitsgrade	95%-Quantil	Testgröße Version A	B	C
Altersgruppe	32	46,19	0,07	0,09	6,22
Erwerbstätigkeit	3	7,81	1,01	1,55	0,94
Wirtschaftszweig	96	119,87	17,44	30,57	98,22
Beruf	331	374,42	326,43	511,21	10,99
Stellung im Beruf	10	18,31	7,86	3,41	4,16
Arbeitszeit	11	19,68	19,66	13,40	11,57
Nettoeinkommen	13	22,36	11,59	13,18	16,04
Ausgeübte Tätigkeit	10	18,31	9,06	9,00	15,28
Arbeitsplatz	10	18,31	5,43	6,02	6,13
Stellung im Betrieb	10	18,31	19,81	10,04	9,47
Schulabschluß	5	11,07	12,21	14,16	6,22
Berufsbild. Schulabschluß	7	14,07	6,78	7,33	7,42

- Die Abbildungen 2, 3, 4 geben einen Überblick über die mittleren relativen Fehler der einzelnen Merkmale, die durch die Anonymisierung bei den verschiedenen Versionen entstehen. Auf die Berücksichtigung des Stichprobenfehlers des Mikrozensus ist in dieser Graphik verzichtet worden. Der Einfluß der Sortierung ist deutlich für das Merkmal Beruf zu erkennen. Der mittlere relative Fehler fällt von 8,6% bei Version B bzw. 7,3% bei Version A auf 1,6% bei der für das Merkmal Beruf günstigen Version C. Auch die Varianz des mittleren relativen Fehlers nimmt entsprechend ab. Der maximal auftretende Anonymisierungsfehler geht ebenfalls von 100% bei Version B und 61,1% bei Version A auf 50% bei Version C zurück.

- Die Abbildungen 5, 6, 7 zeigen die Abschätzung der Vertrauensgrenzen für die theoretisch durch die Anonymisierung zu erwartenden maximalen relativen Fehler:

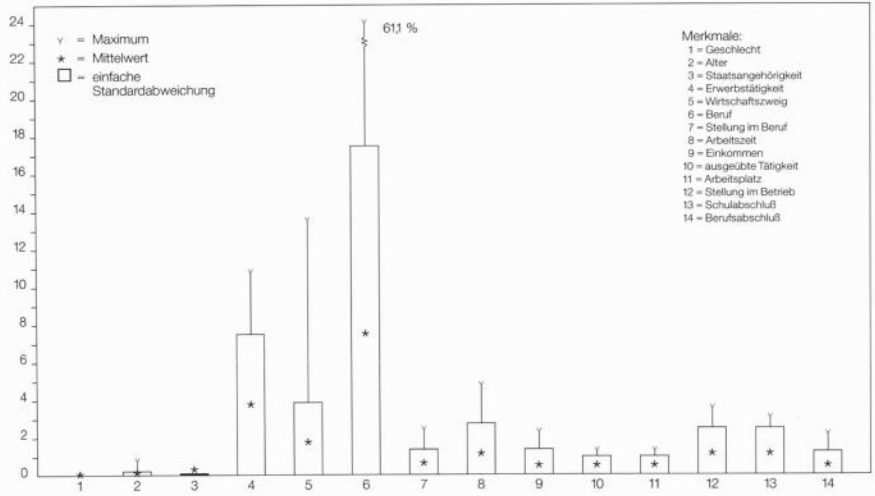
$$|f| < 2{,}58 \cdot \sqrt{2/n}$$

sowie die experimentell ermittelten Fehler für ein Merkmal bei den verschiedenen Versionen. Diese Graphiken zeigen deutlich den Einfluß der Sortierung. Wenn bei der Version C auch bei gering besetzten Ausprägungen des Merkmals Beruf relativ große Fehler auftreten können, so sind die Fehler bei allen höher besetzten Ausprägungen praktisch zu vernachlässigen.

Abbildung 2

Mittlerer relativer Fehler in Prozent nach der Anonymisierung

(Version A)



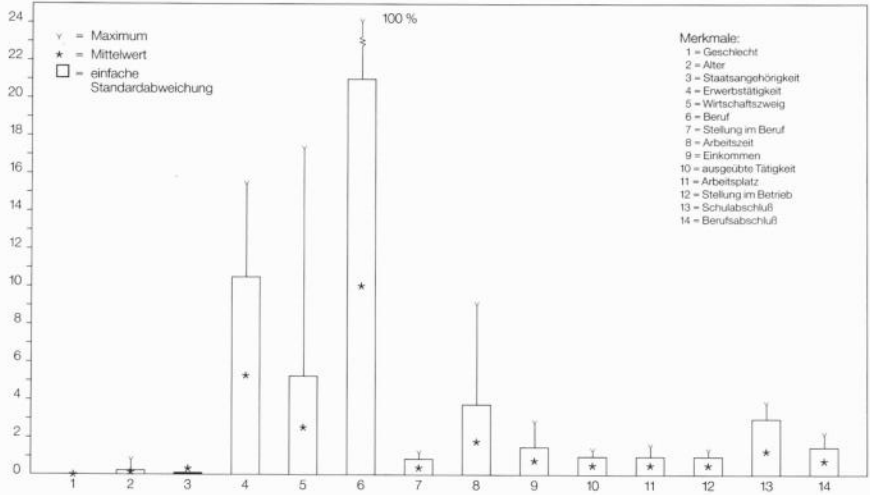
Statistisches Informationssystem

11. 06. 86

Abbildung 3

Mittlerer relativer Fehler in Prozent nach der Anonymisierung

(Version B)



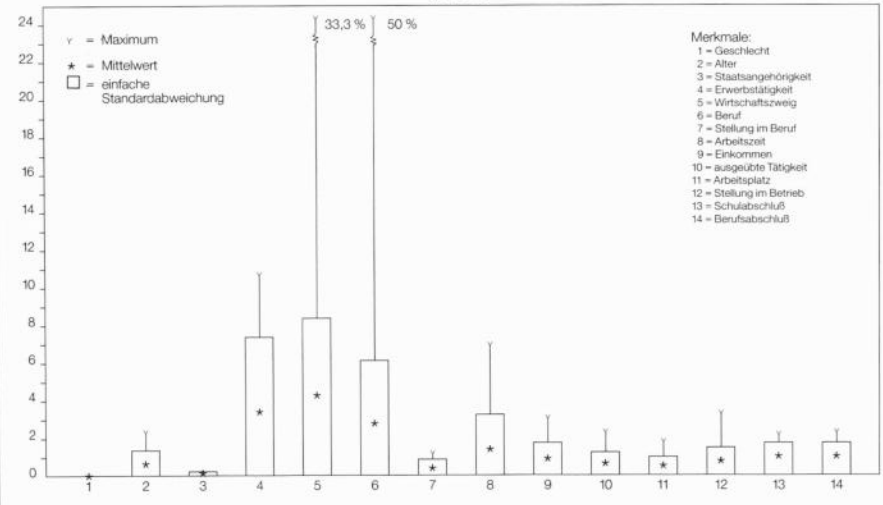
Statistisches Informationssystem

13. 06. 86

Abbildung 4

Mittlerer relativer Fehler in Prozent nach der Anonymisierung

(Version C)



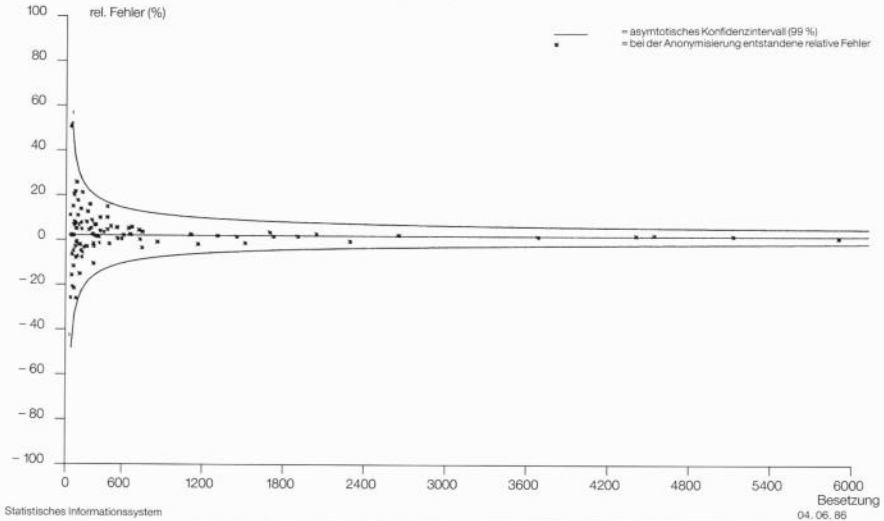
Statistisches Informationssystem

05. 06. 86

Abbildung 5

Relativer Fehler nach Besetzungszahlen für das Merkmal Beruf nach der Anonymisierung

(Version A)



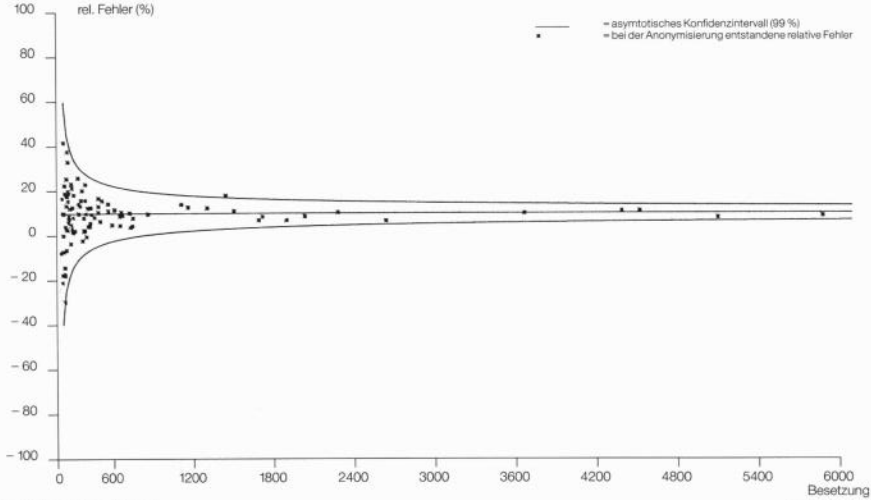
Statistisches Informationssystem

04. 06. 86

Abbildung 6

Relativer Fehler nach Besetzungszahlen für das Merkmal Beruf nach der Anonymisierung

(Version B)



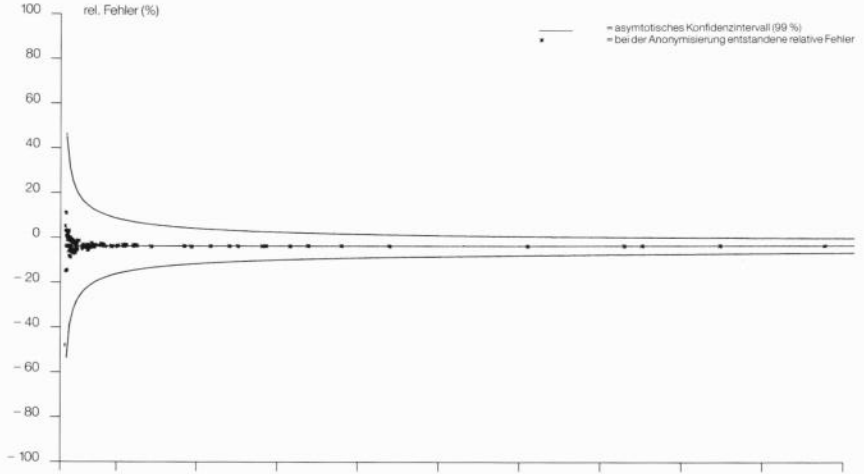
Statistisches Informationssystem

04. 06. 86

Abbildung 7

Relativer Fehler nach Besetzungszahlen für das Merkmal Beruf nach der Anonymisierung

(Version C)



Statistisches Informationssystem

05. 06. 86

Ausblicke

Bei dem hier kurz angedeuteten Verfahren könnte noch der Einfluß verschiedener Abstandsfunktionen auf die Anonymisierungsmaßnahmen untersucht werden. Hierbei könnte man auch verschiedene Abstandsfunktionen auf disjunkte Teilmengen des zu anonymisierenden Materials anwenden, die garantieren, daß ein Satz einer Teilmenge nur innerhalb seiner eigenen Teilmenge zur Aggregation verwandt werden darf, daß also eine Aggregation verschiedener Sätze aus verschiedenen solchen Teilmengen nicht vorkommen darf. Von einer derartigen Verfeinerung des besprochenen Verfahrens kann sicher ein weiterer Fortschritt erwartet werden. Allerdings zeigen derartige Vorgehensweisen, daß eine Anonymisierung nicht nur in hohem Maße von dem gewünschten Auswertungszweck, sondern auch von der Beschaffenheit des zur Anonymisierung vorgelegten Materials abhängt.

Schließlich sollte untersucht werden, welche weitergehenden Anonymisierungsmöglichkeiten sich durch das Zulassen von nicht ganzzahligen Hochrechnungsfaktoren ergeben.

Literaturhinweise

Dalenius, T. (1981): A Simple Procedure for Controlled Rounding, Statistik tidskrift 3/81.

Kühn, J., Pfrommer, F., und Schrey, E. (1984): Zur technischen Weiterentwicklung des Statistischen Informationssystems des Bundes, Wirtschaft und Statistik 12/84, S. 981ff.

Paaß, G. (1982): Statistical match with additional information, Interner GMD-Bericht.

Paaß, G., und Wauschkuhn, U. (1984): Datenzugang, Datenschutz und Anonymisierung; Analysepotential und Identifizierbarkeit von anonymisierten Individualdaten, München.

Dr. Hans-Peter Kirschner
Statistisches Landesamt, Berlin

Stichprobenverfahren und Auswahlsätze als Mittel der Anonymisierung

Vorbemerkung

Auswahlsatz und Informationsgehalt der Daten sind wichtige Parameter, von denen bekanntermaßen das Reidentifikationspotential, das in einem als Stichprobe vorliegenden Mikrodatenfile (MDF) enthalten ist, in hohem Maße abhängt.

Obwohl auch dem jeweils zugrundeliegenden Stichprobenplan eine Parametereigenschaft in diesem Sinne zuzubilligen ist, gibt es bisher keine Untersuchung, die für komplexe Stichprobenpläne darstellt, in welcher Weise diese zur Anonymisierung eines MDF beitragen können. In den folgenden Abschnitten wird eine erste Annäherung an diesen Problemkreis vorgestellt. Dabei wird durchgehend und ohne weitere Erläuterungen Terminologie verwendet, wie sie etwa in Paaß und Wauschkuhn (1984) oder in Schlörer (1980) enthalten ist.

Einleitung

In dem von der Gesellschaft für Mathematik und Datenverarbeitung (GMD) vorgelegten Schlußbericht des Forschungsvorhabens AIMIPH – Konstruktion und Erprobung eines anonymisierten integrierten MDF der Bundesdeutschen Privathaushalte – (Paaß und Wauschkuhn (1984)) wird sehr deutlich, welchen hohen Stellenwert Auswahlsätze von Stichproben bei der Beurteilung von Deanonymisierungsmöglichkeiten bzw. Reidentifikationspotentialen haben können. Es wird aber auch zugleich gezeigt, daß Auswahlsatz und Informationsgehalt der erhobenen Merkmale zusammen, d. h. in ihrem Wechselspiel zu betrachten sind, wenn es gilt, Deanonymisierungsrisiken abzuschätzen. Es soll im folgenden nun dargestellt werden, wie neben Auswahlsatz und Informationsgehalt auch die Art der Datenerhebung an sich, d. h. der Stichprobenplan, einen Einfluß auf Möglichkeiten zur Einkreisung bzw. Reidentifikation haben kann. Selbstverständlich wird die ohnehin komplexe Fragestellung nach Mitteln der Anonymisierung von MDF noch schwieriger zu beantworten sein, wenn man in Gestalt des Stichprobenplans einen weiteren „Modellparameter" einführt, der dann in vielfältiger Weise variieren kann und zudem in Wechselwirkung mit den anderen Einflußgrößen steht. Unter solchen Umständen wird man also versuchen müssen, entweder Szenarios komplexer Natur per Simulation anzugehen, oder aber das Problem so lange in nicht triviale „Unterprobleme" zu zerlegen, bis von denen wenigstens in erster Näherung eines einer theoretischen Lösung zugänglich ist.

Hier wird schwerpunktmäßig der zweite Weg gewählt und es sei vorbereitend an einem Beispiel demonstriert, wie zum einen reales Geschehen sich überaus verwickelt darstellen kann und wie zum anderen gerade dieser Umstand sich gut zur präzisierenden Diskussion zentraler Begriffe aus der Anonymisierungsdebatte eignet.

1 Der ALLBUS: reidentifikationsgefährdet?

Vorwiegend im Bereich der empirischen Sozialforschung wird eine Vielzahl von MDF mit Fallzahlen zwischen 2000 und 3000 für praktisch jedermann vorgehalten; diese Files zeichnen sich durch einen sehr umfangreichen demographischen Teil aus und enthalten pro Befragtem typischerweise die Regierungsbezirkskennung sowie in sieben Stufen die politische Gemeindegrößenklasse und einen Boustedt-Index. Die bekannten ALLBUS-MDF gehören dazu und deren „Verteilerstelle" ist das Zentralarchiv für empirische Sozialforschung der Universität Köln. Bezieht man die Stichprobengröße von 3000 auf die bei solchen Umfragen übliche Grundgesamtheit der wahlberechtigten Deutschen in Privathaushalten, ergibt sich ein Auswahlsatz von rd. 0.7 Zehntausendstel (0.000068) in Kombination allerdings mit einer umfangreichen Demographie und scheinbar ungefährlichen regionalen Bezügen, nämlich Regierungsbezirk und politische Gemeindegrößenklasse. Insbesondere des verschwindend kleinen Auswahlsatzes wegen würde man nun – trotz des hohen Informationsgehalts des soziodemographischen Teil-MDF dieser Art nur ein sehr geringes Reidentifikationspotential zubilligen. Leider ist dies schlicht falsch. Irgend jemand, der unter hohem Einsatz von Mitteln Deanonymisierung betreiben will, wird bei Dateien dieser Art eine reelle Erfolgschance haben. Nehmen wir an, daß dieser Angreifer – ohne zunächst personenspezifisches Angriffswissen zu benutzen – aus der Kombination von Regierungsbezirk und politischer Gemeindegrößenklasse im ALLBUS 1984 u. a. herausfindet, daß sieben der Interviews in Aschaffenburg oder Schweinfurt stattgefunden haben müssen: mit einer Einwohnerzahl zwischen 50000 und 100000 in Kombination mit dem entsprechenden Boustedt-Index gibt es nämlich nur diese beiden Städte im Regierungsbezirk Unterfranken. Auf dieselbe Weise könnte er z. B. feststellen, daß fünf der Interviews in Hildesheim (Regierungsbezirk Hannover) stattgefunden haben müssen. Für beide Demaskierungen auf Gemeindeebene gilt im übrigen, daß pro herausgefundener regionaler Einheit jeweils mehr als 100000 Einwohner vorhanden sind – die entsprechende Regel „nicht weniger als 100000" ist also nicht verletzt! Gefährlich im Sinne der Deanonymisierung und – natürlich – hocherfreulich für den Demaskierer ist hingegen der Umstand, daß beide Regionaleinheiten aus administrativen Raumeinheiten bestehen, für die gerade deshalb umfangreiches Angriffswissen, etwa berufsbezogen, vorhanden ist. Man denke an Rechtsanwälte, Ärzte, Handwerker usw. Zum Tragen kommt hier in voller Schärfe der große Informationsgehalt der erhobenen Demographie. Diese enthält u. a.:

- Geburtsdatum, Geschlecht, Familienstand
- Anzahl der Kinder
- Komplette Haushaltsstruktur
- Erwerbstätigkeit im Detail
- Erwerbstätigkeit Vater/Mutter
- Schulbildung im Detail
- Schulbildung Ehepartner im Detail
- Schulbildung Vater/Mutter

– Heiratsjahre, Scheidungsjahre, Todesjahre Ehepartner

– Nettoeinkommen monatlich

– überwiegender Lebensunterhalt

– Religionsgemeinschaft, auch Ehepartner

– Zugehörigkeit zu einer Organisation

– Art des Wohnverhältnisses.

Auf dieser Wissensbasis sind also der Findigkeit des Demaskierers praktisch keine Grenzen gesetzt, wobei noch hinzukommt, daß er aus veröffentlichter Literatur weiß, daß Fälle in der Datei mit gleicher Klumpenkennung eng beieinander wohnen müssen, nämlich innerhalb eines Wahlbezirks.

Kennzeichnend für dieses Beispiel ist als erstes die Mehrstufigkeit des Deanonymisierungsprozesses, wobei u. a. auf verschiedenen Aggregationsebenen (Gemeinde/Personen) das Schema „Mikrodatenfile versus Identifikationsfile" im Sinne der Hintertreppenidentifikation zum Tragen kam. Als zweites ist zu nennen, daß der sehr niedrige Auswahlsatz mehr als ausgeglichen wird durch die regionalspezifischen Angaben in Kombination mit der komfortabel ausgestatteten Demographie. Der Stichprobenplan an sich spielte in diesem (realisierbaren) Szenario ersichtlich eher eine hilfreiche Rolle bei der Demaskierung (Stichwort: Wahlbezirk) und nicht zuletzt zeigt dieses Beispiel, daß keineswegs nur Datensammlungen der amtlichen Statistik von der Anonymisierungsdebatte betroffen sind.

Wollte man das Vorgehen in diesem Beispiel mathematisch-modellhaft abbilden, um das zweifelsohne vorhandene Risikopotential zu berechnen, würde man auf große Schwierigkeiten stoßen, weil die Situation sehr komplex ist. Allein den Faktor „Findigkeit des Angreifers" zu beschreiben, wäre praktisch nicht befriedigend möglich, wie überhaupt kompliziertere Einkreisungsverfahren sich theoretisch – quantifizierender Betrachtung meist entziehen.

Reduktion der Komplexität würde hier bedeuten, das Muster „Mikrodatenfile versus Identifikationsfile" herauszulösen und nach Chancen bzw. Risiken von Einkreisungsversuchen zu fragen, nach einer Hintertreppeneinkreisung also. In unserem Beispiel bestünde diese etwa auf Personenebene darin, daß der Angreifer oder Demaskierer angesichts der beiden Facharbeiter, die er vielleicht in seinen fünf Hildesheimer Interviews gefunden hat und die angegeben haben, einer bestimmten Branche und auch einer Gewerkschaft anzugehören, auf kriminellem Wege ein Mitgliederverzeichnis als Identifikationsfile dieser Gewerkschaft beschafft; dessen Grunddaten würde er komplett als Überschneidungswissen benutzen können. Es wird sich ergeben, daß nur noch wenige Personen, deren Identifikatoren der Demaskierer nun besitzt, sich hinter den beiden Interviews verbergen, die Einkreisung also bereits weit fortgeschritten ist.

Im folgenden wird nun die Situation „Hintertreppeneinkreisung bei gegebenem Mikrodatenfile und Identifikationsfile" unter der Vorgabe betrachtet, daß

– der MDF eine formal anonymisierte Zufallsstichprobe aus einer irgendwie definierten Grundgesamtheit sein soll und

– der Identifikationsfile aus einem Element bestehen möge (man sucht also sein Wissen über eine Person mit Hilfe des MDF zu erweitern).

Letzteres bedeutet keine Einschränkung, läßt jedoch bei der Diskussion wesentliche Punkte besser hervortreten. Des weiteren wird vorausgesetzt, daß der Angreifer nicht weiß, ob bestimmte Einheiten (die Einheit) aus seinem Identifikationsfile im MDF enthalten sind (ist) oder nicht. Diese Konstellation läßt sich immer über die Ziehung einer Substichprobe aus dem MDF erzwingen, muß also hier nicht weiter problematisiert werden.

Bevor im nächsten Abschnitt die mathematisch-formale Behandlung des Einflusses von Stichprobenplänen auf Chancen zur Deanonymisierung begonnen wird, sei eine kurze Nebenbemerkung zum Begriff Hintertreppeneinkreisung vorangeschickt. Diese Sprechweise ist selbstverständlich bis zu einem gewissen Grade synonym zum Gebrauch des Begriffs der probabilistischen Reidentifikation. Hier wird jedoch diese Terminologie insofern nicht verwendet, als insbesondere bei der Nutzung anonymisierter Einzelangaben aus Mikrodatenbeständen der amtlichen Statistik in Form von Stichproben und mit eingeschränkter Variablenzahl, die erfolgreiche Suche nach einer bestimmten Merkmalskombination im Wertevorrat des MDF noch lange nicht bedeuten muß, den Träger dieser Merkmalskombination identifiziert zu haben. Wohnt dieser in Norddeutschland, kann der record im MDF gerade von einem seiner Doppelgänger in Süddeutschland stammen. Man kreist also höchst deterministisch ein und identifiziert keineswegs sofort in dem Sinne, daß man den richtigen Identifikator ermittelt.

2 Stichprobenziehung und Deanonymisierungswahrscheinlichkeit

Will man für die (eingeschränkte) Problemstellung „Hintertreppeneinkreisung bei gegebenem Mikrodatenfile und Identifikationsfile" einsehen, wieso eigentlich ausgerechnet das Stichprobenverfahren, das den MDF erzeugt, auch anonymisierend wirken kann und wie dies mit dem Auswahlsatz, also der Stichprobengröße, zusammenhängt, ist eine formalisiertere Betrachtungsweise angezeigt. Das folgende Schema möge zur Veranschaulichung des gedanklichen Ansatzes dienen:

$$\text{Grundgesamtheit} = \{ e_1, e_2, \ldots, e_N \}$$

$$\text{Mikrodatenfile (MDF)} = \{ e_{i1}, e_{i2}, \ldots, e_{in} \} \quad \text{Angriffsbasis}$$

$$\begin{array}{l|l} e_{i1} & (\underset{\sim}{y}_1 \mid \underset{\sim}{z}_1') \\ e_{i2} & (\underset{\sim}{y}_2 \mid \underset{\sim}{z}_1') \\ & \qquad (\underset{\sim}{z}(e^*), e^* \mid \underset{\sim}{v}(e^*)) \\ e_{in} & (\underset{\sim}{y}_n \mid \underset{\sim}{z}_1') \end{array}$$

Die Elemente der Grundgesamtheit werden hier über ihre Identifikatoren dargestellt, nicht über ihre Werte; gleiches gilt für den MDF. Die Wertekombinationen des Überschneidungswissens werden mit $\underline{z}$ bezeichnet, wobei ~ andeuten soll, daß es sich um Vektoren handelt.

Die Symbole $\underline{z}'$ bezeichnen die Wertevektoren des MDF, und zwar in dem Sinne, daß sie genau dasjenige wiedergeben, was zum Zeitpunkt der Erhebung dieser Daten der Interviewer bzw. der Befragte zu Papier gebracht haben. Der Begriff des „wahren Wertes" wird hier bewußt vermieden, da er die Diskussion zum einen in formaler Hinsicht nur unnötig erschweren würde und zum anderen seine Definition ohnehin methodisch nicht unumstritten ist. Im gleichen Sinne ist $\underline{z}(e^*)$ diejenige Wertekombination, die dem Angreifer zum Zeitpunkt seines Angriffs zusammen mit dem Identifikator e^* zur Verfügung steht. Wie diese aus dem Wertevorrat, der zur Zeit der Erhebung des Mikrodatenfiles vorhanden war, entstanden ist, soll hier nur insoweit interessieren, als wir diesem Vorgang eine zufällige Komponente zubilligen wollen.

Die Werte $\underline{y}$ und $\underline{v}$ sind in der Darstellung hinzugenommen worden, um die zweite Runde der Hintertreppeneinkreisung anzudeuten: Hat man erst einmal zu $\underline{z}(e^*)$ „passende" Merkmalskombinationen gefunden, wird man über Korrelationen zwischen den $\underline{v}$- und den $\underline{y}$-Werten fortfahren bei der Eingrenzung der Zahl der „ähnlichen" records. Die Doppelleiste zwischen den Identifikatoren und den Werten soll betonen, daß man auf MDF-Seite lediglich mit einem Wertevorrat arbeitet, die Identifikatoren hingegen unbekannt sind.

Zu jeder Wertekombination der Angriffsbasis gehört die Elementarmenge $E(\underline{z}(e^*))$, die in Anlehnung an Schlörer (Schlörer [1980: 121]) wie folgt definiert wird:

$$E(\underline{z}(e^*)) = \text{alle } e \text{ in der Grundgesamtheit mit } \underline{Z}(e) = \underline{z}(e^*).$$

Sie bezeichnet also die $\underline{z}$-Doppelgänger, die e^* zur Zeit des Angriffs besitzt. Diese Menge und auch ihre Größe sind in der Regel nicht bekannt, und ihre Größe ist zudem meist schwer schätzbar, da sie fast immer sehr klein ist: man wird kaum einen Angriff über eine Wertekombination starten, von der man z. B. annehmen muß, daß sie in der Population 30000 mal vorkommt.

Das für den Angreifer wünschenswerte Ergebnis der Einkreisung läßt sich als Ereignis formulieren, dem man eine Wahrscheinlichkeit zuordnen kann:

„Wenigstens ein e aus $E(\underline{z}(e^*))$ liegt im MDF
und $\underline{z}'(e)$ ist ähnlich zu $\underline{z}(e^*)$."

Wichtig dabei ist, daß zum einen die Stichprobe MDF wenigstens eine Einheit aus der Elementarmenge trifft und daß zum anderen die Annahmen zur Entstehungsgeschichte von $\underline{z}(e^*)$ in Form von „ähnlich sein" zutreffen. Man kann also sagen, daß stichprobenbezogener Anteil und subjektiv bestimmter definitorischer Anteil die Wahrscheinlichkeit einer gelungenen Einkreisung bestimmen, diese Wahrscheinlichkeit aber stets kleiner oder gleich ist der Wahrscheinlichkeit für den stichprobenbezogenen Anteil allein und dieser soll im folgenden näher untersucht werden.

Will man darstellen, welchen Einfluß die Art der Stichprobenziehung auf den Einkreisungserfolg hat, reduziert sich also unter den gegebenen Rahmenbedingungen die Frage auf das Problem zu bestimmen, mit welcher Wahrscheinlichkeit wenigstens ein Element der Elementarmenge von z(e*) im MDF vorkommt. Plausiblerweise wird diese Größe bestimmt sein durch die Art der Stichprobenziehung und natürlich den Auswahlsatz. Randomisiert der Stichprobenplan perfekt, ist also jede Kombination von Einheiten der Zielpopulation in der Stichprobe möglich, sollte die Wahrscheinlichkeit allein vom Auswahlsatz abhängen, denn die Randomisierung ist blind gegen Gruppenzugehörigkeit zur Elementarmenge. Ist die Randomisierung hingegen nicht total, können also z. B. drei verschiedene Personen aus einem Klumpen nicht gemeinsam in die Stichprobe gelangen, weil nach dem Stichprobenplan in eben diesem Klumpen nur zwei Personen gezogen werden, wird es darauf ankommen, wie die nicht-totale Randomisierung Größe und Struktur der Elementarmenge beeinflußt. Um diese Art von Plausibilitäten quantifizieren zu können, seien einige zusätzliche Bezeichnungen eingeführt: Zunächst sei daran erinnert, daß der MDF als eine Zufallsstichprobe vorausgesetzt wird. Der Auswahlsatz sei mit f=n/N bezeichnet, mit N als Größe der Grundgesamtheit bzw. n als Stichprobengröße.

Die Inklusionswahrscheinlichkeit für eine bestimmte Einheit mit dem Identifikator e sei $\pi(e)$, also

$$\pi(e)=P \text{ (Die Einheit e liegt im MDF).}$$

Die Anzahl der Einheiten in der Elementarmenge E(z*) sei m(E). Der Mittelwert der Auswahlwahrscheinlichkeiten für die Elemente aus E sei π (E); E=E(e(z*)). π sei der Mittelwert aller Auswahlwahrscheinlichkeiten (man beachte, daß für wichtige Stichprobenpläne π=f). Man kann nun recht einfach die folgende Gleichung herleiten (auf Beweise wird an dieser Stelle und auch im folgenden verzichtet):

(1) P (wenigstens eine Einheit aus E liegt im MDF)=
=m(E)*f+m(E)*(π(E)−π)+Rest
Rest=Rest (Stichprobenplan)

In dieser Gleichung wird deutlich, warum zum einen der Auswahlsatz bei der Anonymisierung eine so große Rolle spielt und auf welche Weise dabei der Stichprobenplan zum Tragen kommt. Der erste Summand für die Wahrscheinlichkeit positiver Einkreisung hängt einzeln linear vom Auswahlsatz und der Größe der Elementarmenge m(E) ab. Bei vollständiger Randomisierung verschwinden die anderen Terme und man erhält das vorhersagbare Resultat.

Der Vollständigkeit halber sei dies explizit angegeben für den Fall der uneingeschränkten Zufallsauswahl, bei der man mit einem einfachen und direkten kombinatorischen Argument die folgende „Einschachtelung" beweisen kann; leider versagt dieser Beweisgang bei komplizierteren Stichprobenplänen:

$$1-(1-f)^m <$$
$$P \text{ (wenigstens eine Einheit aus E im MDF)} <$$
$$1-\left\{1-[f/1-m/N]\right\}^m.$$

Gleichung (1) zeigt, daß die fragliche Deanonymisierungswahrscheinlichkeit durchaus kleiner werden kann als der erste Summand allein, da beide restlichen Summanden ein Vorzeichen haben können. So wird bei nahezu verschwindendem Rest die Tatsache, daß die Elemente der Elementarmenge im Schnitt kleinere Auswahlwahrscheinlichkeiten haben könnten, für eine geringere Deanonymisierungswahrscheinlichkeit sorgen. Natürlich kann auch der umgekehrte Fall eintreten, der den Angreifer eher begünstigt. Entscheidend sind dabei dessen Kenntnisse über die Elementarmenge zusammen mit der Art des Stichprobenplans. Allgemeiner kann man sagen, daß bei festgehaltener Elementarmenge die Deanonymisierungswahrscheinlichkeit von der Vergleichsgröße (Auswahlsatz*Größe der Elementarmenge) abweicht, wenn „innere" Eigenschaften der Elementarmenge zusammenhängen mit der Art der Stichprobenziehung. Ein Beispiel möge dies verdeutlichen:

- Man ziehe aus einer Substichprobe der Größe 10000 aus den Auswahlbezirken des Mikrozensus je genau einen Haushalt zufällig und aus diesen zufällig dann genau eine Person. Dann wird der Stichprobenanteil der Deanonymisierungswahrscheinlichkeit für solche Elementarmengen sinken, die aus Personen in vorwiegend größeren Haushalten bestehen. Solche Personen haben nämlich ersichtlich eine relativ geringere Chance, in die Stichprobe zu gelangen. Der Angreifer würde also unter diesem Stichprobenplan für ganz spezielle Elementarmengen geringere Chancen besitzen, als wenn etwa eine uneingeschränkte Zufallsauswahl vorliegen würde. Es bleibt die Frage: wieviel geringer denn?

Unabhängig vom Stichprobenplan für den Mikrozensus läßt sich für die nicht unrealistische Situation

- daß in keinem Klumpen (oder: Auswahlbezirk) mehr als eine Person der Elementarmenge vorhanden ist und

- daß die Stichprobengröße pro Klumpen gleich Eins ist,

eine formale Darstellung wie folgt geben:

$$P(\text{wenigstens eine Einheit aus E im MDF}) = 1-(1-q^*f)^m$$

q=mittlere Klumpengröße/mittlere Klumpengröße in E.

Bezieht man die Formel auf die eben geschilderte Substichprobe aus dem Mikrozensus, kann man z. B. schließen, daß für eine Einheitsmenge von Personen, die alle in Haushalten der Größe vier leben, die Deanonymisierungswahrscheinlichkeit sich ungefähr halbiert, verglichen mit der Standardsituation uneingeschränkter Zufallsauswahl. Umgekehrt verdoppelt sie sich jedoch für Einheitsmengen, die nur aus Personen in Einpersonenhaushalten bestehen. Der Stichprobenplan führt also zu einer Verbreiterung des Wertebereichs für den stichprobenabhängigen Anteil an der Deanonymisierungswahrscheinlichkeit, und zwar um den Zentralterm m(E)*f herum. Kennt der Angreifer nicht den Stichprobenplan, ist dies für ihn natürlich von Nachteil, weil er so damit rechnen muß, nach einer Elementarmenge zu suchen, die u. U. nur mit sehr geringer Wahrscheinlichkeit in der Stichprobe

vertreten ist. Zu beachten ist jedoch, daß in dem Beispiel für einen Substichprobenplan aus dem Mikrozensus recht gewaltsam eine Disproportionierung der Daten eingeführt wurde, die sich dann entsprechend deutlich in den Wahrscheinlichkeiten bemerkbar machte. Stichprobenpläne, die selbstgewichtend sind, müssen dagegen bei weitem nicht solch eine Wirkung haben.

Schlußfolgerungen

Ganz allgemein kann man folgendes Fazit ziehen: Stichprobenplan und Auswahlsatz sind in ihrer Kombination als Mittel der Anonymisierung insbesondere dann tauglich, wenn der Stichprobenplan in seinem Abbildungsverhalten für die Populationsparameter abweicht von uneingeschränkter Zufallsauswahl. Tut er dies nicht, bleibt als Mittel der Anonymisierung von beiden Möglichkeiten eben nur der Auswahlsatz und die Resultate z. B. aus Paaß und Wauschkuhn (1984) können ohne weiteren Kommentar herangezogen werden.

Zu berücksichtigen bleibt allerdings, daß auch durch geschicktes und damit „verschleierndes" Ziehen von Substichproben das Deanonymisierungsrisiko nicht verschwindet. Man erhält für bestimmte Szenarios auf diese Weise für die Restrisiken u. U. einige Nullen hinter dem Komma mehr; jedoch wird eine Freigabeentscheidung für einen Mikrodatenfile sich weniger daran orientieren können, wie klein diverse Wahrscheinlichkeiten sind, wenn die „Verlustfunktion" – in diesem Falle für denjenigen, der den MDF herausgibt – im übertragenen Sinne dann den Wert unendlich annehmen kann, wenn doch einmal eine Einkreisung erfolgreich war. Man könnte diese Situation durchaus vergleichen mit Restrisiken etwa bei konventionellen Atomkraftwerken.

Sinnvoll eingesetzte Stichprobenverfahren können zweifelsohne dazu beitragen, einen MDF gegen Einkreisungsangriffe zu schützen. Absolute Sicherheit gegen Reidentifikationen ist im allgemeinen mit solchen Methoden jedoch nicht zu erreichen. Entscheidungen über die Freigabe eines MDF sollten sich also keinesfalls allein auf eher technische Maßnahmen wie Stichprobenziehungen stützen; eine Güterabwägung hinsichtlich des Restrisikos unter übergeordneten Kriterien bleibt nach wie vor unumgänglich.

Literaturhinweise

Paaß, G., und Wauschkuhn, U., (1984): Schlußbericht des Forschungsvorhabens AIMIPH, Konstruktion und Erprobung eines anonymisierten integrierten Mikrodatenfiles der Bundesdeutschen Privathaushalte. Gesellschaft für Mathematik und Datenverarbeitung. St. Augustin/Darmstadt 1984.

Schlörer, J., (1980): Datenorientierte Verfahren der Anonymisierung, in: M. Kaase, H.-J. Krupp et al.: Datenzugang und Datenschutz. Athenäum Verlag, Königstein/Ts. 1980, S. 118-142.

Prof. Dr. Erwin K. Scheuch
Universität Köln

Risiko - Interpretation beim Datenschutz

1 Veränderungen im Klima für Forschung

Wenn hier über Erfahrungen mit dem Datenschutz berichtet wird, wenn Begründungen für die Beeinträchtigung von Forschungsfreiheit zitiert und mit den Ansichten von Datenschützern gehadert wird, so ist das alles nur ein Teil des Problems. Die Vorfälle und Auffassungen werden bedeutsam vor dem Hintergrund einer Klima-Änderung: Nicht in der Bevölkerung allgemein, wohl jedoch unter geisteswissenschaftlich Gebildeten, verstärkt sich ein empiriefeindliches Klima.[1]) Empirische Sozialforschung hat da keinen Vertrauensvorschuß, sondern steht unter Rechtfertigungsdruck in jedem Einzelfall.

„Datenskandal in Schweden" war die Überschrift einer von der Deutschen Presse Agentur (dpa) am 25. 2. 1986 verbreiteten Meldung, die eine Kampagne von „Dagens Nyheter" zusammenfaßte.[2]) 1966 hatte Carl Gunnar Jansson eine Datei von 15000 Jugendlichen aus Stockholm begonnen, deren Lebenslauf er verfolgen wollte. Sozialwissenschaftliche Forschung ist im Normalfall eine Momentaufnahme. Wenn wir in einem gegebenen Moment feststellen, welche Stellung eine Person im Gefüge der sozialen Schichtung hat oder wie groß der Abstand im Prestige von einer Berufsgruppe zur nächsten ist, so ist daraus noch nicht viel über die Starrheit eines Schichtungssystems abzuleiten. Hier werden als Grundlage für Aussagen Lebensläufe notwendig. Überhaupt dürften wir mit den Momentaufnahmen soziales Leben zu starr, zu mechanistisch abbilden. Soweit wir in der Sozialwissenschaft aber Lebenslaufdaten erhalten, sind dies retrospektive Daten. Das Projekt „Metropolit" von Carl Gunnar Janson war das weltweit einzige mit prospektiven Daten. Zu Recht hieß es in der dpa-Meldung: „Mit dem Projekt sollte der Lebensweg aller in Stockholm geborenen Schweden des Jahrgangs 1953 bis zum Tode erforscht werden . . ." Ein einmaliges Datenmaterial, dessen Existenz Sozialwissenschaftlern mit den Spezialitäten Demographie, Devianzsoziologie und Sozialökologie weltweit bekannt war.

„Wie in Stockholm bekannt wurde, enthielt das . . . Forschungsprojekt „Metropolit" auch eine heimlich angelegte Sammlung von Daten über politische Aktivitäten . . ." Tatsächlich hatte Prof. Jansson auch Daten über politisches Verhalten in seine Datei aufgenommen.

---

[1]) Kennzeichnend hierfür ist Hoimar von Ditfurth: So laßt uns denn ein Apfelbäumchen pflanzen – Es ist soweit, Hamburg 1985. In der Kontroverse über Tierversuche im Frühjahr 1986 wurde offensichtlich, daß die Notwendigkeit weiteren empirischen Wissens gering geschätzt wird, wenn einmal eine Streitfrage als Moralentscheid definiert wurde; vgl. Erwin K. Scheuch, Das Tier als Partner des Menschen, in: Studium Generale III, Tierärztliche Hochschule Hannover 1986. Ein Wissen, das nicht ökologisch-grün ist, wird vom Hamburger Wissenschaftssenator Meyer-Abich sogar als Zerstörungswissen qualifiziert; so in seiner Antrittsrede vor dem Senat der Universität Hamburg. Siehe auch Günter Altner et al. (Hrsg.), Manifest zur Versöhnung mit der Natur, Neukirchen-Vluyn 1984.

[2]) Abgedruckt u. a. in Süddeutsche Zeitung, 15. Februar 1986. Eine ausführlichere Darstellung bringt Hannes Gamillscheg: „Forscher registrierte heimlich Lebensdaten eines ganzen Jahrgangs", Kölner Stadt-Anzeiger, 12. Februar 1986, S. 6. Wie unangemessen die Bezeichnung „Datenskandal" war, ergibt sich auch daraus, daß Professor Janson seine Datenbank mit ausdrücklicher Zustimmung der Datenschutzbehörde anlegte.

Erhalten hatte er diese Daten, wie alle Daten der Datei Metropolit, von verschiedenen Behörden in Stockholm. Selbstverständlich war die Existenz von „Metropolit" auch den schwedischen Datenschützern bekannt und seit es sie gibt von ihnen genehmigt worden. Ein Problem wurde nicht gesehen – wie sollte es auch in einem Land, in dem Steuerdaten publiziert werden. Ein „Datenskandal" wurde daraus erst durch eine Veröffentlichung, in der die Datei als „geheim" angelegt qualifiziert wurde. Das ist nach normalem Wortverständnis von Geheim unzutreffend, nicht aber nach dem Wortverständnis des Journalisten. Der sah das Geheime bei der Anlage der Datei darin, daß „ohne Informierung der Betroffenen" Angaben von Behörden an die Universität gegeben wurden. Der Ausdruck „Betroffene" zeigt bereits aus welcher Schauweise heraus dieser Vorgang als Skandal gewertet wurde: Der Bürger wird beobachtet!

Es gibt zwar kein Land wie das unsrige, in dem sogar eine Volkszählung einem erheblichen Teil der Publizistik als bedrohliches Unternehmen erscheint, aber ein Aspekt der bundesdeutschen Hysterie, wie sie auch anläßlich des maschinenlesbaren Personalausweises wieder akut wurde, ist mit unterschiedlicher Intensität in allen protestantisch geprägten Industriegesellschaften anzutreffen: Hysterie gegenüber dem Beobachtetwerden. Erkenntnisgewinn für Wissenschaft und der Verweis, daß Wissenschaftler ja keine Individuen als Einzelperson beobachteten, sondern an Personen lediglich ein kategoriales Interesse haben, daß Menschen dem Sozialwissenschaftler zumal nur als Merkmalsträger gelten, haben für die Hysterisierten keinen Belang. Teilweise im Gegenteil: Es wirkt beleidigend, als Merkmalsträger behandelt zu werden, wo man in sich selbst doch sonst ein zur Reflektion berufenes raisonnierendes Wesen sieht. Überhaupt gilt in breiten Kreisen der geisteswissenschaftlich Gebildeten empirische Forschung nicht als Möglichkeit, Wesentliches über Menschen zu entdecken.

Über 20 Jahre hinweg ist niemandem auch nur der geringste Nachteil durch das Einspeichern in die Datei von „Metropolit" entstanden. Die Vertraulichkeit der Datenspeicherung ist niemals verletzt worden. Die als Enthüllung vorgestellte Story von „Dagens Nyheter" erregte jedoch die Öffentlichkeit, obgleich ein Recht nicht verletzt wurde. „Metropolit" wurde gestoppt und die Daten sollen auf dreißig Jahre unter Verschluß bleiben, weil ein Prinzip verletzt wurde – das der elektronischen Unsichtbarkeit. Analoges gilt auch für die Bundesrepublik. Es sind nicht Fälle von Mißbrauch, die für einen Datenschutz angeführt werden, der Sozialforschung zunehmend behindert. Es geht um ein Prinzip. Und wenn es um Prinzipien geht, dann findet keine Güterabwägung mehr statt.

Das Gegenteil war gemeint, als 1972 in der Bundesrepublik wesentlich Sozialforscher Datenschutz forderten. Durch die Fortschritte in der Datenverarbeitung wurde es möglich, bisher getrennt gespeicherte Daten zu Personenprofilen zusammenzufügen. Darin wurde ein Problem gesehen, weil in einer hoch differenzierten Gesellschaft ein großer Teil des Bewegungsspielraums von Personen daraus folge, daß jeweils nur ein Segment von ihm bekannt sei, ein für die jeweilige Rollensituation relevantes Segment. Werde es jetzt möglich, getrennte Daten zusammenzuführen, so werde eine Transparenz wie in einer Dorfgesellschaft hergestellt – aber nur einseitig zugunsten von Institutionen. Wenn sich dies

anhört wie Passagen aus der Begründung des Urteils des Bundesverfassungsgerichts (BVG) zum Volkszählungsgesetz, so ist dies kein Zufall![3])

Datenschutz war damals konzipiert worden als Schutz vor Wissen bei Vollzugsbehörden aus Quellen, die nichts mit dem regulären Geschäftsbereich der jeweiligen Vollzugsbehörde zu tun hatten. Nicht der Schutz des Bürgers gegen Aufzeichnungen in Datenbanken, wohl aber der Transport der Daten sollte unter Kontrolle gestellt werden mit dem Ziel, das Zusatzrisiko (!) durch EDV auszugleichen. Es sollte kein neues Persönlichkeitsrecht als Teil einer Systemveränderung geschaffen werden.[4])

Daraus wurde teilweise im Bundesdatenschutzgesetz (BDSG) etwas anderes und soll nach dem Willen der Datenschützer noch etwas ganz anderes werden. „Informationserwartungen und Verhaltensweisen, die vor den Datenschutzgesetzen auf keinen Widerspruch gestoßen sind, müssen deshalb keineswegs auch weiter hingenommen werden." – so der hessische Datenschützer Prof. Spiros Simitis.[5]) Simitis gibt zu, daß es Umstände rechtfertigen können, auf die Einwilligung des „Betroffenen" zur Datenverarbeitung zu verzichten. Wo dies zuträfe, müsse der so entfallene Schutz durch den „Betroffenen" selber ersetzt werden durch zusätzliche Kontroll- und Sicherungsbedingungen. Vor Datensicherung und Kontrolle bei der Analyse von Dateien sind also im Prinzip auch nach Simitis Ausgleichsmöglichkeiten für das im allgemeinen geltende Erfordernis einer Einwilligung des in der Datei Abgebildeten. „In allen übrigen Fällen (dem Normalfall des wissenschaftlichen Vorgehens) gilt dagegen für die wissenschaftliche Forschung nichts anderes wie für jede andere Verarbeitung: Der Datenschutz ist Zäsur im Umgang mit personenbezogenen Angaben."[6]) Obwohl die Wissenschaft im Grundgesetz gegenüber anderen Tätigkeitsfeldern privilegiert ist (Artikel 5 GG), heißt es bei Simitis apodiktisch „ . . . wie für jede andere Verarbeitung". Das bedeutet: Für die Datenschützer gilt eine Privilegierung der Wissenschaft nicht. Die einzige hier hingenommene Privilegierung ist die der im Grundgesetz schlechter gestellten Presse. Und ein Zweites ist an diesem oben zitierten Satz bemerkenswert: Der Datenschutz soll Zäsur sein. Die vorerwähnte, ursprüngliche Zielsetzung bei der Entwicklung des Datenschutzes, nämlich die Neutralisierung eines Zusatzrisikos, ist mithin für die Datenschützer heute ersetzt durch Datenschutz als Mittel einer Gesellschaftsveränderung.

2 Die Schriftlichkeit der Einwilligung als Prinzip

Die Umfrageforschung, aber auch viele Arten psychologischer Forschung, sind besonders betroffen von der Forderung, daß Datenerhebung nur zulässig sein soll, wenn der Befragte vorher schriftlich sein Einverständnis erklärte. Bekanntlich ist es den Sozialforschern

[3]) Terminus und Begriff der „informationellen Selbstbestimmung" sind die Schöpfung Ernst Bendas. Aber die Vorstellung, nicht nur das einzelne Datum sei gegebenenfalls schutzwürdig, sondern insbesondere die Aufspaltung des Wissens über eine Person geht auf die frühen Diskussionen zurück. Sie beginnen mit einer Tagung der Naumann-Stiftung 1972.

[4]) Siehe hierzu auch Erwin K. Scheuch, Die Weiterentwicklung des Datenschutzes als Problem der Sozialforschung, in: Max Kaase et al. (Hrsg.), Datenzugang und Datenschutz, Frankfurt 1980.

[5]) Datenschutz und wissenschaftliche Forschung, in: Jan Peter Waehler (Hrsg.), Deutsch-polnisches Kolloquium über Wirtschaftsrecht und das Recht des Persönlichkeitsschutzes, Tübingen 1985, S. 121.

[6]) A. a. O.

gelungen, in Absprache mit den Datenschützern eine Suspendierung dieses Grundsatzes zu vereinbaren, wenn bestimmte Kautelen beachtet sind. Es handelt sich aber um eine Suspendierung und nicht um ein Aufgeben dieser Forderung, wie sie insbesondere vom hessischen Datenschützer Simitis vertreten wird.

Der Ursprung dieser Forderung ist inzwischen in der Bundesrepublik vergessen worden. Ausgangspunkt waren in den USA Vorfälle in der medizinischen Forschung. Beispielsweise war Strafgefangenen in den Südstaaten der USA, die von Syphilis befallen wurden, in der Hälfte der Fälle eine wirksame Substanz als Gegenmittel verabreicht worden, in der anderen Hälfte der Fälle jedoch ein Placebo. Die letzteren Gefangenen siechten dann jämmerlich dahin. Ein weiterer, die Öffentlichkeit sehr erregender Vorfall waren Versuche mit alten Menschen. Den Insassen eines Altersheims wurden z. T. Krebszellen injiziert, zum anderen Teil dagegen ein neutrales Mittel. Den alten Menschen war erklärt worden, daß sie sich an einem medizinischen Experiment beteiligten, aber nicht gesagt worden, welche gesundheitlichen Risiken damit für sie verbunden waren. Hier griffen dann amerikanische Gerichte ein: Sie verlangten, daß bei Experimenten den zustimmenden Personen die für sie möglichen Konsequenzen in einer Weise erklärt wurden, die auch von der teilnehmenden Person tatsächlich verstanden wurde. Die Formel „informed consent" bedeutet nach amerikanischer Rechtsprechung, die Pflicht des Forschers nachzuweisen, daß der Teilnehmer an einem Experiment die damit verbundenen Risiken versteht.

Ursprünglich war bei der Diskussion über Datenschutz an das Prinzip der schriftlichen Einwilligung gar nicht gedacht worden. Es ist wohl allein auf Simitis zurückzuführen, der durch die Forderung nach Schriftlichkeit beim Einverständnis der Teilnahme an einer Befragung oder an einem Experiment, bei den „Betroffenen" eine Hemmschwelle aktivieren will. Im Alltag sind wir alle vorsichtiger, wenn uns etwas Schriftliches abverlangt wird, und eben diese Zusatzvorsicht will Simitis aktivieren. Es ist dies eine Vorsicht, die mit dem eigentlichen Inhalt einer Untersuchung überhaupt nichts zu tun haben muß – es ist gewissermaßen die Aktivierung eines Reserve-Mißtrauens.

Es gibt hierfür bei Simitis eine offizielle und eine informelle Begründung. Die offizielle Begründung ist: „Die Datenschutzgesetze weigern sich, in ihm (dem Betroffenen) nur das Verarbeitungsobjekt zu sehen."[7]) Nun sei einmal davon abgesehen, daß hier das Gesetz zu einer handelnden Person gemacht wird, also der Fehler der Reeffizierung begangen wird. In diesem Zusammenhang ist allein wichtig die Vorstellung, einem Menschen werde ein Stück seiner Würde genommen, wenn er nicht genau kontrollieren könne, was mit an ihm beobachteten Eigenschaften in der Analyse geschieht. „Daten, die sich auf seine Person beziehen, sollen nicht ohne seine Kenntnis und Entscheidung verarbeitet werden." Und wenngleich dann viele Mitmenschen bereit wären, Zwecken der Forschung allgemein zuzustimmen, sollen sie durch die Forderung nach Schriftlichkeit daran gehindert werden.

Gegenüber Herrn Simitis wurde von uns eingewandt, daß eine Verweigerung häufig nichts zu tun haben werde mit dem jeweiligen Zweck einer wissenschaftlichen Untersuchung, sondern nur allgemeine Ängstlichkeit vor schriftlicher Festlegung ausdrücke. Hier werde

[7]) A. a. O., S. 105.

also Forschung behindert aus einem Umstand, der mit Forschung selber gar nichts zu tun habe. Mithin werde also durch die Datenschützer nicht nur das im Grundgesetz ausgedrückte Privileg der Wissenschaft ignoriert, sondern die Wissenschaft durch etwas behindert, was mit ihren Tätigkeiten eigentlich nichts zu tun habe. Darauf antwortet dann Simitis privat: „Er wolle nicht, daß es so viele Dateien über Menschen gäbe."

Nun ist die Art der Datenhaltung bei der Wissenschaft völlig anderer Natur als bei Behörden. Wissenschaft hat an Personen lediglich ein kategoriales Interesse und nicht an dem, was eine Person zur Person macht. Ganz anders steht es mit der Datenhaltung bei Behörden, insbesondere wenn es sich um Vollzugsbehörden handelt. Hier interessiert gerade das, was die einzelne Person zu einem Fall für die Behörde werden läßt. Und dies ist – aus verständlichen Gründen – dann auch der Grund gewesen, warum im Datenschutzgesetz das Prinzip der vorherigen schriftlichen Einwilligung nicht durchweg gilt. In Datenschutzgesetzen wie in denen des Landes Nordrhein-Westfalen oder Rheinland-Pfalz ist die vorherige schriftliche Einwilligung nicht erforderlich, wenn bei der Verarbeitung keine „schutzwürdigen Belange" des Betroffenen berührt werden. Mit einer solchen Klausel kann die Wissenschaft hervorragend leben, denn – einige medizinische und psychiatrische Forschungen einmal ausgenommen – schutzwürdige Belange eines Betroffenen werden in dieser kategorial orientierten Forschung ja nicht berührt. Simitis wendet sich entschieden gegen eine solche Formel. Und er wendet sich auch gegen die weitere Ausnahme, die dann gegeben ist, wenn entweder das öffentliche Interesse überwiegt oder die Einholung eines Einverständnisses unzumutbar sei. Auch mit diesen Formeln kann die Wissenschaft hervorragend leben. Unzumutbar ist das schriftliche Einverständnis, wenn dadurch der Forschungszweck zunichte wird. Und öffentliches Interesse überwiegt durchweg bei Forschung.

„Schutzwürdige Belange", „öffentliches Interesse" und „Zumutbarkeit" sind Formeln, die eine Güterabwägung erfordern. Eben eine solche Abwägung will Simitis prinzipiell nicht.

Hierfür gibt wiederum Simitis zwei verschiedene Begründungen. Einmal argumentiert er, daß jede Güterabwägung beim Verzicht auf Schriftlichkeit die Verwaltung zum Richter über Methoden und Ziele der Forschung machte. Das Argument gilt dann nicht, wenn Forschung wirklich voll institutionalisiert ist, also der Verweis auf Forschung ausreicht, um das Vorgehen zu rechtfertigen. Tatsächlich kann heute nicht mehr durchweg davon ausgegangen werden, daß Forschung einen solchen Institutionalisierungsgrad hat – aber vorherrschend trifft dies zu. Zumindest würde sich nach einigen Jahren und einigen Verwaltungsgerichtsverfahren durchsetzen, daß Forschung bei der Güterabwägung ebenso privilegiert zu behandeln ist wie die Verwaltung bei der Durchführung ihrer administrativen Aufgaben. Das Hauptargument für Simitis ist jedoch, daß er der Verwaltung keine Abwägungen zuordnen will, die er alleine den „Betroffenen" vorbehalten will. Der Grund hierfür ist ein Mißtrauen in die Eignung der Verwaltung, nach pflichtgemäßem Ermessen zu entscheiden.

Gegenüber den Sozialforschern wenden viele Datenschützer ein, daß die Schriftlichkeit keine Beeinträchtigung der Forschung zur Folge habe. Diese Auffassung ist aufgrund der

[8]) A. a. O., insbesondere Abschnitt 5: Die Verdrängung des Betroffenen.

gesamten empirisch begründeten Literatur über das Interview als Methode unhaltbar.[9]) Inzwischen gibt es aber auch zusätzlich methodische Erfahrungen mit der Wirkung eines schriftlichen Einverständnisses.

Die methodisch eindrucksvollste Untersuchung ist eine Erhebung von Eleanor Singer in Zusammenarbeit mit dem National Opinion Research Center (NORC) der Universität Chicago.[10]) Die Umfrage sollte die Wirkung der verschiedenen Möglichkeiten einer „informierten Einwilligung" („informed consent") sowohl auf die Bereitschaft zur Beteiligung am Interview wie auch auf die Qualität der Antworten messen. Drei Faktoren wurden geprüft: (1) Schriftlichkeit, (2) Länge der Information, (3) Ausmaß der Zusicherung von Vertraulichkeit. Diese Bedingungen wurden nach dem Prinzip des lateinischen Quadrats in der experimentellen Forschung variiert, womit sich für 2084 ausgewählte Personen pro Zelle zwischen 115 und 116 Fälle ergaben.

Versuchsanordnung zur Prüfung, welche Wirkung verschiedene Formen „informierter Teilnahme" haben

	Ausführliche Erklärung des Untersuchungszwecks			Kurze, vage Erklärung des Untersuchungszwecks		
	Absolute Vertraulichkeit	Vertrauliche Behandlung	Keine Zusicherung	Absolute Vertraulichkeit	Vertrauliche Behandlung	Keine Zusicherung
Unterschrift vor Beginn des Interviews	115	115	115	115	115	116
Unterschrift nach Beendigung	115	115	116	115	116	115
Keine Unterschrift	115	116	116	116	115	116

Inhaltliches Thema des Interviews war überwiegend Freizeitverhalten. Es wurden aber bewußt „schwierige" Fragen hinzugefügt zu den Themen Alkoholgenuß, Marihuanakonsum, sexuelles Verhalten und geistige Gesundheit.

Die stärkste negative Wirkung auf die Bereitschaft, sich an der Umfrage zu beteiligen, hatte die Forderung nach schriftlichem Einverständnis. Dabei war es von untergeordneter Wirkung, ob die Unterschrift vorher zu leisten war oder im Anschluß an das Interview. Während allgemein die Ausschöpfungsrate der Stichprobe 71% der ausgewählten Personen betrug, sank sie bei der Forderung nach schriftlichem Einverständnis auf 64% bzw. 65%. Die Länge der Erklärung und die Zusicherung der Vertraulichkeit war von geringer Bedeutung für die Teilnahme am Interview insgesamt. Es ließen sich allerdings Wirkungen bei der Bereitschaft beobachten, die „schwierigen" Fragen zu beantworten. Nur für diese Fragen hatte die Zusicherung absoluter Vertraulichkeit eine stimulierende Wirkung.

[9]) Vgl. Hartmut Esser, Soziale Regelmäßigkeiten des Befragtenverhaltens, Meisenheim am Glan 1975.
[10]) Eleanor Singer, Informed Consent – Consequences for Response Rate and Response Quality in Social Surveys, in: American Sociological Review, Bd. 43, 1978, S. 144–162.

Allgemein blieb aber die Wirkung der verschiedenen Versuchsbedingungen auf die Qualität eines Interviews, nachdem einmal eingewilligt worden war, von zu vernachlässigender Bedeutung.

Wie zu erwarten, wirkte die Forderung nach schriftlichem Einverständnis unterschiedlich bei Untergruppen der Bevölkerung. Die Verweigerungsrate nach der Forderung, das Einverständnis schriftlich zu erklären, stieg insbesondere an bei Befragten im Alter von 65 Jahren und mehr und bei Befragten mit lediglich Grundschulbildung. Hier lag die Verweigerung bei nahezu 12 %! Dies kann als Bestätigung dessen gedeutet werden, was weiter oben über die Wirkung der Schriftlichkeit gesagt wurde: Hierdurch wird ein diffuses Vorschuß-Mißtrauen gegen Festlegung aktiviert. Dadurch erhält man weniger Auskunft über die Untergruppen der Bevölkerung, die einem sozialpolitisch-sozialreformerisch Motivierten besonders am Herzen liegen müßten.

Aus der Bundesrepublik ist ein besonders drastischer Fall der negativen Wirkung von Schriftform auf die Bereitschaft zur Teilnahme an sozialwissenschaftlichen Forschungen zu berichten: Die Umfrage „Deutsche in der Sowjet-Gesellschaft".[11]) Mit Mitteln der Stiftung Volkswagenwerk sollte da erforscht werden, welches Schicksal der Aussiedlung bei 20000 Aussiedlern zwischen 1979 und 1983 voranging und was darauf folgte. Zusätzlich sollte die Untersuchung Rückschlüsse auf die Sozialstruktur der Sowjetunion erlauben. Aus einer zentralen Kartei des Deutschen Roten Kreuz (DRK) wurde eine Stichprobe von 2538 Aussiedlern gezogen. Auf Insistieren des Bundesinnenministeriums wurden die zu Befragenden nicht nur vorher angeschrieben, sondern auch aufgefordert, postalisch ihr Einverständnis zur Befragung mitzuteilen. Die Ausschöpfung dieser Stichprobe betrug lediglich 16%, womit die angestrebte Repräsentativität verloren war. Eine anschließende Untersuchung bei den phantastisch hohen Ausfällen ergab, daß der wichtigste Faktor bei den zu Befragenden ein allgemeines Mißtrauen war – nicht gegen eine wissenschaftliche Befragung, sondern eben gegen jede Form einer schriftlichen Festlegung gegenüber Fremden. Die Forderung nach postalisch mitzuteilendem Einverständnis löste bei den potentiell Befragten Ängste aus, die andere Objekte hatten als das, was mit der Forderung nach Schriftlichkeit beim Datenschutz gemeint ist. Die Regelung verfehlt also den gemeinten Zweck und beeinträchtigt zugleich eine Forschung.

Max Kaase verweist beim Hearing am 21. April 1986 auf Probleme, die sich mit der angemessenen Form der Einwilligung bei Telefonumfragen ergeben. Wie soll hier die schriftliche Form der Einwilligung erreicht werden? Dabei ist das Telefoninterview von wachsender Bedeutung für die Umfrageforschung. Regelungen, welche diese Entwicklung nicht berücksichtigen, wirken also ausgesprochen innovationshemmend auf die Forschung.

Problematisch ist auch, ob die Forderung nach völliger Information des Befragten über den Befragungszweck das Ziel erreicht, was mit dieser Forderung angestrebt wird. Sehr viele Zwecke der Untersuchung lassen sich überhaupt nicht vernünftig erklären, wenn der Partner nicht ebenfalls sozialwissenschaftlich ausgebildet ist. Und in manchen Fällen wird

[11]) Mitteilung des Umfrage-Instituts GETAS an den Auftraggeber Osteuropa-Institut im März 1986.

der Untersuchungszweck zerstört, wenn er dem Befragten mitgeteilt wird. Ein Beispiel: Soll ich die Bereitschaft, NPD zu wählen als Folge einer bestimmten Persönlichkeitsstruktur (etwa als autoritäre Persönlichkeit), vorher mitteilen? Es ist offensichtlich, daß ich dann auf diese Weise nicht klären kann, ob die NPD für problematische Persönlichkeiten besonders attraktiv ist. Häufig ordnen heute Gerichte an, daß bei einem Streit über die Verwechslungsgefahr eines Gutes die Ergebnisse einer Umfrage entscheiden sollen. In solchen Fällen wird nicht einmal dem Interviewer der Zweck der Erhebung mitgeteilt, damit er nicht befangen ist; wieviel stärker müßte eine solche Erklärung die Fähigkeit des Befragten verringern, unbefangen zu antworten.

Es ist auch nicht einzusehen, welches Rechtsgut hier zu schützen wäre. Das Interesse der Forscher ist kategorial. Der einzelne Befragte ist eine austauschbare Person, die lediglich in diesem gegebenen Falle der Zufall zum Partner macht. Jenseits dieser Zufälligkeit besteht kein Interesse. Die Datenverarbeitung erfolgt selbstverständlich anonym. Warum hier überhaupt das Datenschutzgesetz anwenden? Das ist nur möglich bei einer sogenannten „offensiven" Interpretation der Datenschutzgesetze. Die Sozialforscher hatten gegenüber den Datenschützern dargelegt, daß bei einer normalen Umfrage überhaupt keine personenbezogenen Dateien entstehen. Ein Personenbezug des Interviews bleibt lediglich bestehen, bis die Tatsache der ordnungsgemäßen Ausführung kontrolliert werden konnte. Vor dem Beginn der Datenverarbeitung wird jedoch in den normalen Fällen der Personenbezug gelöscht. Also wäre bei normaler Interpretation der Datenschutzgesetze überhaupt keine Datei im Sinne dieser Gesetze entstanden. Die Forderung, auch auf diesen Normalfall der Umfrageforschung Datenschutzgesetze anzuwenden, erscheint von der Sozialforschung aus als Interpretationswillkür einiger Datenschützer, die in erster und letzter Instanz über ihre Deutung befinden.

3 Die Personenbeziehbarkeit als Prinzip

Das BDSG kennt den Terminus „Personenbeziehbarkeit" nicht. Hier heißt es „personenbezogen". Das ist ein objektiv zu kennzeichnender Sachverhalt, nicht jedoch das, was mit dem Wort „Personenbeziehbarkeit" ausgedrückt wird. Hier handelt es sich um einen Begriff, der bei der Anwendung der Datenschutzgesetze durch die Datenschützer in die Diskussion eingeführt wurde.

Bekanntlich gelten die Restriktionen des BDSG nur für solche Daten, die personenbezogen sind. Entfällt der Personenbezug, dann entfällt an sich auch die Anwendung des BDSG. Mit dem Begriff „Personenbeziehbarkeit" wird die Anwendung des BDSG über den ursprünglich bei der Gesetzesformulierung erörterten Anwendungsbereich hinaus erweitert. Und zwar wird jetzt das Datenschutzgesetz auch auf solche in nicht-personenbezogener Form vorliegende Daten anwendbar, bei denen durch Dritte – auch einschließlich unerlaubter Handlungen – nachträglich eine Entanonymisierung erfolgt oder erfolgen könnte. Wann dies vorliegen kann, ist eine fachliche Frage, zu der aber auch eine Güterabwägung hinzutreten muß.

Hier ein Beispiel, wie durch Kombination von Sortiermerkmalen die Zahl der Fälle, die die gleiche Kombination von Eigenschaften hat, immer kleiner wird:

1. Filterschritt	Geschlecht	männlich
2. Filterschritt	Alter	35–39 Jahre
3. Filterschritt	Berufstätigkeit	selbst voll berufstätig
4. Filterschritt	Berufsgruppe	Freier Beruf
5. Filterschritt	Schulabschluß	Hochschule
6. Filterschritt	Persönliches Netto-Einkommen	über 1 200 DM
7. Filterschritt	Netto-Einkommen der Familie	über 2 000 DM
8. Filterschritt	Besitzgüter	Wochenendhaus
9. Filterschritt	Kfz-Besitz	ja
10. Filterschritt	Konfession	evangelisch
11. Filterschritt	Selbsteinstufung in soziale Schicht	Gehobene Schicht
12. Filterschritt	Wohnsitz	...
13. Filterschritt	Wohnortgröße	Gemeinde bis 2 000 Einwohner

Theoretisch kann überhaupt nie ausgeschlossen werden, daß eine Entanonymisierung erfolgt, wenn eine größere Zahl von Deskriptoren aufgezeichnet wird. Es ist dies ähnlich wie bei einer Geheimschrift: Auch bei dieser verbleibt ein Restrisiko, daß ein für sicher gehaltener Code doch noch entschlüsselt werden könnte.

Andererseits erfordert eine Entanonymisierung einen erheblichen Aufwand und verlangt gute Kenntnisse in der Datenverarbeitung. Ein auf Bändern gespeichertes Material ist besser gesichert als das gleiche Material, wenn es auf Karteikarten verzeichnet wäre.

Nun ist durchaus zuzugeben, daß die Entanonymisierung für die empirische Sozialforschung ein Problembereich ist. Werden spezielle Populationen untersucht und nicht ein Querschnitt für die Gesamtbevölkerung befragt, dann kann sich aus der Umgrenztheit des Personenkreises die Möglichkeit einer Entanonymisierung sehr erhöhen. Am Zentralarchiv für empirische Sozialforschung in Köln versuchte Rolf Uher eine Entanonymisierung eines Datensatzes für „Politische Elite". Das waren im vorliegenden Falle alle Kandidaten für das Berliner Abgeordnetenhaus bzw. die Berliner Bezirksversammlungen von 1981 an. Von insgesamt 1910 Kandidaten beantworteten 875 den Fragebogen. Eine Entanonymisierung alleine auf der Grundlage der anonym gespeicherten Daten war nicht möglich. Aber über diesen Personenkreis existierten weitere Register, in denen u. a. Namen, Adresse, Geburtsdatum, Parteimitgliedschaft, Geschlecht und Beruf verzeichnet waren. Das Risiko der Entanonymisierung variierte jetzt unter Verwendung von Zusatzwissen bei diesem sehr speziellen Personenkreis zwischen einer Chance von 1 aus 2 bis 1 aus 10.[12]) Ähnliche Probleme ergeben sich etwa bei der Befragung von Drogenabhängigen, wo der Forscher sich das Zusatzwissen aus Gesundheits- und Sozialämtern beschaffen könnte. Aber für übliche Umfragen kann das Restrisiko einer Entanonymisierung als praktisch gleich Null bezeichnet werden.

[12]) Rolf Uher, Attempt of a Deanonymization, vervielfältigtes Manuskript für die IFDO-IASSIST-Konferenz in Amsterdam, Mai 1985.

Inzwischen gibt es eine ganze Reihe von empirischen Forschungen und Modellversuchen, um die Gefährdung von auf verschiedene Weise gespeicherten anonymisierten Daten in Wahrscheinlichkeiten auszudrücken.[13]) Dies ist insbesondere notwendig, wenn das Statistische Bundesamt (StBA) in Zukunft Datenbänder für Sozialforschung zusammenstellen will. Selbstverständlich würde dies in anonymer Form geschehen, aber das StBA muß erklären können, wie sicher die so an die Forschung weitergegebenen Datensätze gegen eine unerlaubte Entanonymisierung sind.[14])

Juristen neigen dazu, eine Regelung von einem denkmöglichen Fall her zu werten und nicht von einem wahrscheinlichen. So ist also die bloße Denkmöglichkeit der Entanonymisierung für die Datenschützer ein Anlaß zu restriktiven Überlegungen. Die European Science Foundation hatte als Faustregel vorgeschlagen, Daten dann als gegen Entanonymisierung gesichert anzusehen, wenn „die Re-Identifizierung einen unverhältnismäßigen Aufwand an Zeit, Kosten und Personal erfordert".[15]) Wie Simitis dazu richtig anmerkt, findet sich damit die European Science Foundation mit einem Restrisiko für die Betroffenen ab. Wieso auch nicht? Dieses Restrisiko ist bei vernünftigen Formen der Datensicherung gewiß viel geringer als das Risiko, Opfer eines anderen Delikts oder eines Unfalls zu werden. Zudem kann als gesichert davon ausgegangen werden, daß in der ganz großen Mehrheit aller Fälle es auch nicht die geringste Motivation für einen Forscher gibt, eine Entanonymisierung seiner Fälle zu versuchen. Sein Interesse am Material ist ja eben kategorialer Art und nicht das Interesse eines Verwaltungsbeamten am Einzelfall.

Eine Entanonymisierung ist – bis auf extreme Sonderfälle – allein aus dem Material heraus nicht möglich. Sie wird dann zum Problem, wenn Zusatzwissen vorliegt. Das wird von Simitis auch richtig dargestellt. „Die Relativität aller Anonymisierungen weitet, anders ausgedrückt, den Anwendungsbereich der Datenschutzgesetze um ein Vielfaches aus."[16]) Die Lösung dieses Problems ist die sogenannte funktionale Trennung. Damit ist gemeint, daß die Analyse der Daten und das Wissen, das zur Identifizierung von Einzelpersonen führen könnte, getrennt bleiben muß – durch entsprechende organisatorische und technische Verfahren der Datensicherung. So wird dann bei näherer Betrachtung die Personenbeziehbarkeit weniger eine Frage des Datenschutzes als ein Problem der Datensicherung. Und hier ist ohne Zweifel noch einiges verbesserungwürdig.[17])

[13]) Siehe hierzu als bisher wichtigste deutschsprachige Veröffentlichung Gerhard Paaß und Uwe Wauschkuhn, Datenzugang, Datenschutz und Anonymisierung – Analysepotential und Identifizierbarkeit von anonymisierten Individualdaten. München 1985. Vgl. auch Wolfgang Gorn, Datenschutz und Datensicherung bei Btx, in: net-special, April 85. Zwischen 1980 und 1983 wurden dem Bundeskriminalamt 37 Fälle von Computerkriminalität bekannt, davon in 20 Fällen Computerspionage; Johann Kubica, Computerkriminalität – Versuch einer Systematisierung, in: Schimmelpfeng-REVIEW, Nr. 36, 1985, S. 49–52.

[14]) Das amerikanische Bundesamt für Statistik stellt der Forschung zwei Arten von Dateien zur Verfügung: sogenannte Public Use Samples, die allgemein für Wissenschaftler verfügbar sind, und weitere Microdaten auf besonderen Antrag hin. Ähnlich verfahren weitere Bundesbehörden der USA seit langem – wie z. B. das Bundesfinanzamt oder die Bundessozialverwaltung. Siehe US Department of Commerce: Report on Statistical Disclosure and Disclosure-Avoidance Techniques, Statistical Policy Working Paper 2 Washington 1978. Ferner Robert Mugge, Issues in Protecting Confidentiality in National Health Statistics, in: Review of Public Data Use, Nr. 12, S. 289–294. Beim Statistischen Bundesamt in Wiesbaden ist noch keine endgültige Entscheidung gefallen, ob und wie für Sozialforscher in der Bundesrepublik Datenbänder zugänglich gemacht werden.

[15]) Spiros Simitis, a. a. O., S. 120.

[16]) Spiros Simitis, a. a. O., S. 102.

[17]) Siehe hierzu das Interview mit dem Bundesbeauftragten für den Datenschutz, Dr. Reinhold Baumann, in: die computer zeitung, 2. Oktober 85, S. 7 und S. 10.

4 Verhinderung von Forschung durch „offensive" Auslegung der Datenschutzregeln

Ginge es nur um das BDSG, so gäbe es im Normalfall für Sozialforschung keine Probleme. Zu Recht heißt es in einer Stellungnahme des Bundesministeriums für Verkehr: „Wenn die Markt- und Sozialforschungsinstitute bei Einfachbefragungen personenbezogene Daten nur in manuell geführten internen Dateien speichern, gilt folgendes: Der Personenbezug ist aufzuheben (z. B. durch Löschung der Identifizierungsmerkmale), sobald diese Merkmale nicht mehr benötigt werden. Bei dieser Sachlage findet § 3 BDSG im Hinblick auf die Regelung in § 1, Abs. 2, Satz 2 BDSG keine Anwendung. Eine Einwilligung ist nicht erforderlich. Gleiches gilt, wenn die Angaben zunächst mit Personenbezug manuell gespeichert sind, die anschließende automatisierte Verarbeitung jedoch in anonymisierter Form erfolgt."[18])

Unter den Datenschützern gibt es jedoch überwiegend einen Konsens, daß die Datenschutzgesetze „offensiv" auszulegen seien. Und bei dieser offensiven Auslegung wird insbesondere die Forschung zum Experimentierfeld. Beim Datenschutz gibt es das Gegenteil von dem, was sonst in der Verwaltungswissenschaft öfters thematisiert wird: Statt eines Vollzugsdefizits einen „Vollzugsexzeß".

Vollzugsdefizite ergeben sich insbesondere beim Versuch der Ordnung von Sachverhalten, die nicht zu den klassischen Aufgaben von Staaten gehören. Beispiele sind Bestimmungen über die Reinhaltung der Luft oder von Verordnungen über die Nahrungsmittelsicherheit. Die Komplexität moderner Industriegesellschaften widersteht Versuchen, ein-für-allemal einfache Prinzipien durchzusetzen. Regelungen in Industriegesellschaften bedeuten meist in der Praxis, daß eine Güterabwägung zu erfolgen hat – und damit wird oft die Entscheidung defizitär nach dem Maßstab einer Lex Specialis. „Vollzugsexzeß" wird hier als Neologismus vorgeschlagen, um eine Art der Reaktion auf diese Schwierigkeit zu bezeichnen, die bisher in der Literatur nicht zureichend beachtet wurde. Die Datenschützer versuchen nämlich, mehr und absoluter zu regeln, als der Gesetzgeber im Sinne hatte.

Vielleicht die wichtigste offensive Auslegung des BDSG ist die Erweiterung der Verbotsbestimmungen schon auf die Phase der Datenerhebung. Anders als in der eben zitierten Antwort des Bundesministeriums, die unterscheidet zwischen der Phase der Datenerhebung und der Phase der Datenverarbeitung, will Simitis schon die Datenerhebung dann reglementieren, wenn später einmal – auch nach Löschung des Personenbezuges – eine Datenverarbeitung erfolgt. Dies findet überhaupt keinerlei Stütze in der Gesetzgebung und ist bisher bloße Auslegungswillkür einer Sonderbehörde. Es muß allerdings befürchtet werden, daß diese Auslegungswillkür jetzt nachträglich gesetzlich gerechtfertigt werden könnte, sollte die Datenschutzgesetzgebung um einen Wissenschaftsparagraphen derart ergänzt werden, wie er 1986 im Frühjahr zur Diskussion stand. Nach Simitis ist grund-

[18]) Mitteilung zur Anpassung der empirischen Forschung an die Bestimmungen des Bundesdatenschutzgesetzes (BDSG) vom 18. 6. 1980, in: Deutscher Bundestag, 9. Wahlperiode, Drucksache 9/93, Anhang S. 396-397.

sätzlich die Verarbeitung personenbezogener Daten verboten – und zwar ungeachtet der Absichten des Verarbeiters. So in einem Handbuchartikel 1985.[19])

Anders dagegen Ulrich Dammann vom Amt des Bundesbeauftragten für den Datenschutz.[20]) Er argumentiert: „Die Datenübermittlung wird grundsätzlich zugelassen; die Interessen der Betroffenen und das öffentliche Interesse an dem Forschungsprojekt sind zueinander ins Verhältnis zu setzen; dabei ist das nicht auf die Einzelperson gerichtete Erkenntnisinteresse der Wissenschaft zu berücksichtigen!" Ein schutzwürdiges Interesse der Einzelperson ist verständlicherweise da besonders gegeben, wo er als Einzelperson gewärtigen muß, daß aus der Kenntnis seiner Angabe für ihn als Individuum Folgen entstehen. Das ist jedoch bei der kategorialen Orientierung der Wissenschaft grundsätzlich nicht der Fall. Dies will Simitis nicht zur Kenntnis nehmen, und zwar mit der apodiktischen Begründung: Auf die Absichten des Forschers komme es gar nicht an.

Von großer Bedeutung ist auch die Interpretation der Wendung „Forschungszweck" im BDSG. Das Datenschutzgesetz bindet die Erlaubnis zur Verwendung personenbezogener Daten an einen Forschungszweck. Daten sind zu löschen, sobald der Forschungszweck erfüllt ist. Nun wird in manchen Forschungsbereichen Material – auch personenbezogener Art – auf Dauer gelagert. Eines der Grundprinzipien der Wissenschaftslehre ist die Nachprüfbarkeit von Aussagen. Die Neubearbeitung eines Materials durch einen anderen Forscher als den, der die Datei anlegte, ist ein solcher Fall der Nachprüfung. Hinzu kommt die Gewinnung zusätzlicher Erkenntnisse, an die man bei der ursprünglichen Sammlung des Materials nicht dachte. Durchweg braucht man in den Sozialwissenschaften keinen Personenbezug, aber in manchen Fällen und in manchen Disziplinen – wie in der Medizin und der Psychiatrie – muß die Nachprüfung auch schon einmal personenbezogen sein. Das zu Beginn erwähnte Register von Carl Gunnar Jansson ist ein seltenes Beispiel für die Wichtigkeit einer auf Dauer angelegten Kartei.

Eine solche Datei wie die von Jansson soll nach Simitis prinzipiell unzulässig sein. Er begründet diesen Schluß mit einer sehr engen Auslegung der Wendung „Forschungszweck" im BDSG. Im Gesetz sei damit abgestellt auf ein vorweg anzugebendes und eng umgrenztes Forschungsziel. Die Einengung von Forschungszweck auf Forschungsziel ist jedoch von Simitis bloß gegriffen und wird Disziplinen übergestülpt, in denen ein anderes Verständnis von Forschungszweck üblich ist – wie etwa in der Geschichtswissenschaft.

Simitis konzediert denn auch, daß die historische Forschung mit seiner Auslegung von Forschungszweck teilweise unmöglich würde.[21]) Und eine Reihe von Fällen sind bekannt, wo Behörden entsprechend verfuhren. Fiat Datenschutz, pereat Forschung? Welches Rechtsgut wird hier im Fall der historischen Forschung denn geschützt, wenn wir heute für die Enkel-Generation Forschung unmöglich machen?

[19]) Spiros Simitis, a. a. O.
[20]) Ulrich Dammann, Die falsche Front oder Wissenschaft und Datenschutz, in: Deutsche Universitätszeitung, 1981, S. 609.
[21]) Spiros Simitis, a. a. O., S. 111ff.

In Auseinandersetzungen mit Wissenschaftlern, die sich auf Art. 5 des Grundgesetzes berufen und Güterabwägung verlangen, verweist Simitis auf Ausweichmöglichkeiten. Der Datenschutz, so Simitis, sei ungeachtet der Absichten des Nutzers anzuwenden, blind wie Justitia. Inzwischen kann allerdings mit einigen Datenschützern darüber verhandelt werden, ob nicht ihre Bedenken durch eine funktionale Trennung und die Einführung der Rechtsfigur eines Datentreuhänders Rechnung getragen werden kann. An sich ist das nach dem Wortlaut des BDSG nicht notwendig. Es gibt ja Klauseln für erlaubtes, personenbezogenes Verarbeiten – etwa, wenn ein gesetzlicher Auftrag nur auf die betreffende Weise erfüllt werden kann, oder ein berechtigtes Interesse vorliegt. Ein solches berechtigtes, öffentliches Interesse ist nach dem Selbstverständnis der Forschung eigentlich immer dann gegeben, wenn es um wissenschaftliche Erkenntnisse geht. Eine solche Argumentation versucht Simitis abzublocken, indem er beim Datenschutz eine Güterabwägung zwischen Datenschutz und anderen Rechtsgütern grundsätzlich zurückweist.[22]) „Die Konsequenz liegt auf der Hand: Das Forschungsprivileg droht zum Umgehungsvehikel des gesetzlich garantierten Datenschutzes zu werden."[23])

Die Problematik vergrößert sich durch den Vertrauensverlust, den die Wissenschaft bei einflußreichen Kreisen erlitten hat.[24]) Dies kommt besonders deutlich zum Ausdruck in einem Urteil des Landgerichts Frankenthal vom 30. 1. 1985.[25]) Im Jahre 1983 hatte der Sozialhistoriker Prof. Kocka bei mehreren Gemeinde- und Stadtverwaltungen gebeten, die dort geführten Heirats- und Geburtenbücher der Jahrgänge 1927, 1936, 1955 und 1964 für ein Forschungsvorhaben einsehen zu dürfen. In der Begründung erklärte Kocka, daß zwar zwischen dem Datenschutzrecht einerseits und dem Grundrecht der Forschungsfreiheit andererseits ein Spannungsverhältnis bestehen könne. In diesem Falle aber sei nicht beabsichtigt, Namen zu erfassen, zu verarbeiten oder zu veröffentlichen; vielmehr sei ausschließlich an eine anonyme Erhebung gedacht. Der Schutz persönlicher Daten sei also gewährleistet. Rechtliche Grundlage für eine Zustimmung der Behörden sollte § 61 Personenstandsgesetz sein, der Behörden im Rahmen ihrer Zuständigkeit eine solche Einsichtnahme erlaubt; Kocka bezeichnete sich als Angehöriger der Behörde Universität Bielefeld.

Eine der Stadtverwaltungen verweigerte die Einsichtnahme mit der Begründung, daß die Universität Bielefeld zwar eine Behörde sei, sie aber Einsicht nur für den Zweck der Vollzugsverwaltung haben könnte. Hier gehe es jedoch um Forschung. Das Amtsgericht in Frankenthal hatte die Zurückweisung des Antrags von Kocka für rechtens erklärt und das Landgericht bekräftigte diese Rechtsauffassung. Dabei ist die Begründung für die Wissenschaft von besonderer Bedeutung. Die Einsichtnahme in die Heirats- bzw. Geburtenbücher sei nur erlaubt, wenn ein Dritter ein rechtliches Interesse darlegt. Ein solches rechtliches Interesse läge hier jedoch nicht vor, da Prof. Kocka die Information „jedoch für private (!) Forschungszwecke" benötige.

[22]) Spiros Simitis, a. a. O., S. 117.
[23]) Spiros Simitis, a. a. O., S. 97.
[24]) Vgl. hierzu Erwin K. Scheuch, Deprofessionalisierung von Wissenschaft, in: Helmut Jungermann et al. (Hrsg.): Die Analyse der Sozialverträglichkeit für Technologiepolitik, 1985, S. 16.
[25]) Nach: Zeitschrift für Familienrecht, Heft 6, 1985, S. 615–616. Siehe auch den Kommentar zu diesem Urteil von Michael Hartmer, Wissenschaftsfreiheit und Persönlichkeitsrecht, in: Mitteilung des Hochschulverbandes, Heft 6, 1985, S. 322–323.

Dieses Fehlurteil ist inzwischen rechtskräftig geworden und könnte zu einer langen Auseinandersetzung in der Zukunft darüber führen, ob denn nun wissenschaftliche Forschung wieder als das angesehen wird, was sie vor Hunderten von Jahren war: Das Hobby von Amateuren, und nicht die dienstliche Obliegenheit eines dafür staatlich besoldeten Beamten. Von diesem Gerichtsurteil bis hin zum Feuilleton der Frankfurter Allgemeinen Zeitung läßt sich eine Tendenz bei empiriefreien Akademikern beobachten, die Erheblichkeit von empirisch vorgehenden Wissenschaften zu bezweifeln. Nur innerhalb dieses Klimas wird verständlich, daß Simitis formuliert: „Konsequenter Datenschutz ist, von der historischen Forschung aus gesehen, institutionalisierte Geschichtslosigkeit."[26])

Ganz allgemein wird auch die epidemiologische Forschung außerordentlich erschwert, wenn sich die Auslegung des BDSG durch Simitis durchsetzen sollte. Simitis versteht – im Gegensatz zu manchen anderen Datenschützern – durchaus, worum es bei dieser Art von Forschung geht: Man muß hier mit einer offenen Fragestellung arbeiten, prüft also nicht ein ganz bestimmtes, vorweg erwartetes Ergebnis. Nach dem Wortlaut des Datenschutzgesetzes, auch dem des Novellierungsvorschlags für einen Wissenschaftsparagraphen 3a, muß das kein Hindernis bedeuten. Simitis will allerdings erreichen, daß die Wendung „Forschungszweck" in den gesetzlichen Bestimmungen interpretiert wird als „Forschungsvorhaben". Forschungszweck ist ihm im Wortsinne zu wenig restriktiv. „Im Zweifelsfalle dürfte es ohne allzu große Schwierigkeiten gelingen, die jeweiligen Forschungsinteressen als Forschungszweck auszugeben."[27])

Selbst, wo ein Gesetz ausdrücklich nichts anderes als „Erforderlichkeit für die wissenschaftliche Forschung" voraussetzt, wie das Sozialgesetzbuch in § 75, fordert Simitis die Einengung einer Genehmigung lediglich für ein bestimmtes einzelnes Forschungsvorhaben. Simitis erwartet, daß die Datenschützer erreichen können, daß diese breitere Genehmigung für Forschung im Hinblick auf ihren „längst feststehenden Verständniszusammenhang der Datenschutzvorschriften eingebunden wird". Hier wird deutlich, wie sehr die Datenschützer, allen voran Spiros Simitis, rechtsschöpferisch tätig werden. Dies ist dann eine dritte, bisher in der juristischen Literatur nicht thematisierte Art von Rechtsschöpfung: Neben der des Gesetzgebers und des Richters nun auch noch Rechtsschöpfung durch Behörden.

Besonders deutlich wird diese Tendenz zur Rechtsschöpfung bei dem Begriff „Datei". Hier setzt sich ein Datenschützer wie der frühere Bundesbeauftragte Bull schlichtweg über den Sprachgebrauch hinweg. Im BDSG wird abgestellt auf Dateien, damit ein Begriff aus der Datenverarbeitung übernommen. Er wurde als Neologismus in Analogie zur Kartei geprägt, um damit einen Unterschied auszudrücken: Eine Datei ist ein maschinenlesbares Verzeichnis, eine Kartei ist es nicht. Diesen Unterschied willkürt Bull einfach aus der Welt. Er schreibt: „Der Begriff der Datei ist aber nicht technisch, sondern rein organisatorisch zu verstehen. Eine Datei liegt vor, wenn für eine bestimmte Aufgabe Daten unter einheitlichen Kriterien (formularmäßig) auf Datenträgern (auch Karteikarten) zusammengestellt werden."[28])

[26]) Spiros Simitis, a. a. O., S. 104.
[27]) Spiros Simitis, a. a. O., S. 105.
[28]) Hans-Peter Bull, Bundesbeauftragter für Datenschutz (Hrsg.), Was bringt das Datenschutzgesetz?, Bonn 1978, S. 5.

Wäre das so, dann hätten wir das Wort Datei als Neuprägung nie bekommen. Datei war eben als neues Wort geprägt, um die Andersartigkeit der Verarbeitung zu kennzeichnen, die beim Vorliegen technischer Eigenschaften gegeben ist.

Mit der willkürlichen Ausdeutung des Neologismus „Datei" wollen die so offensiv definierenden Datenschützer erreichen, daß ihnen alle Arten von Verzeichnissen unterliegen. Diese Fehlauffassung hat sich zwar noch nicht allgemein durchgesetzt, aber manche Behörden verhalten sich bereits entsprechend diesem Wunsche einiger Datenschützer. So warnte eine Zeitlang der Deutsche Akademische Austauschdienst (DAAD) seine Gutachter, die ihnen zur Begutachtung zugesandten Anträge unterlägen dem Datenschutzgesetz. Der Hinweis wurde allerdings zurückgezogen, nachdem wir den DAAD informierten, einzelne Handakten seien keine Datei im Sinne des BDSG.

In der Praxis wird der Datenschutz zum Behördenschutz, zur Grundlage einer Auskunftsverweigerung. Hierfür ist ein Beispiel die Stellungnahme des Ministers für Wirtschaft, Mittelstand und Technologie des Landes Nordrhein-Westfalen während der Plenarsitzung vom 7. 3. 1986.[29]) Der Erläuterungsband zum Entwurf des Haushaltsplans des Wirtschaftsministers für das Haushaltsjahr 1986 bringt eine Übersicht über die geplanten wissenschaftlichen Untersuchungen des Ministeriums. Bei einigen Positionen wird der Empfänger der Forschungsmittel nicht genannt; stattdessen findet sich der Hinweis „Privatperson". Hierüber begehrte dann eine der Fraktionen nähere Auskunft. Diese wurde vom Wirtschaftsminister verweigert: „ . . . hat der Verzicht auf die Namensnennung Datenschutzgründe." Darauf hingewiesen, daß im Erläuterungsband bei dem Technologieprogramm Wirtschaft durchaus die Namen von Zuwendungsempfängern genannt werden, erwiderte der Minister, daß er in Zukunft eine Namensnennung aus Gründen des Datenschutzes auch hier unterlassen werde. Da wird mithin eine amtliche Veröffentlichung zu einer Datei im Sinne des Datenschutzgesetzes!

Inzwischen neigen Datenschützer dazu, Behinderungen der wissenschaftlichen Forschung mit dem Urteil des BVG vom 15. 12. 1983 zum Volkszählungsgesetz zu begründen. Dabei wird insbesondere eingewandt, der Grundsatz der „informationellen Selbstgestaltung" verbiete eine Datenverarbeitung, die nicht ausdrücklich und für jeden einzelnen Auswertungsschritt von Untersuchten genehmigt wird. Diese Verwendung des Urteils zur Volkszählung ist nicht nur einseitig, sondern widerspricht dem eigentlichen Tenor.[30]) Das Bundesverfassungsgericht (BVG) behandelt nämlich in seinen Begründungen Forschungs- und statistische Daten anders als Verwaltungsdaten. So schreibt es auf S. 50: „Wenn die ökonomische und soziale Entwicklung nicht als unabänderliches Schicksal hingenommen, sondern als permanente Aufgabe verstanden werden soll, bedarf es einer umfassenden kontinuierlichen sowie laufend aktualisierten Information über die wirtschaftlichen, ökologischen und sozialen Zusammenhänge. Erst die Kenntnis der relevanten Daten und die Möglichkeit, die durch sie vermittelten Informationen mit Hilfe der Chancen, die eine automatische Datenverarbeitung bietet . . . zu nutzen, schafft die für eine am Sozialstaats-

[29]) Landtag Nordrhein-Westfalen, Plenarprotokoll 10/18, S. 1290.
[30]) Wir folgen hier den Darlegungen des Münchener Lehrbeauftragten Dr. Schweizer, unveröffentlichtes Manuskript, März 1986.

prinzip orientierte staatliche Politik unentbehrliche Handlungsgrundlage." Auf S. 53 der Begründung heißt es dann, daß aus dem Recht auf informationelle Selbstbestimmung die Notwendigkeit für besondere Vorkehrungen für die Durchführung und Organisation der Datenerhebung und -verarbeitung folge. „Von besonderer Bedeutung für statistische Erhebungen sind wirksame Abschottungsregelungen nach außen." Weiter auf S. 57: „Wird den erörterten Anforderungen in wirksamer Weise Rechnung getragen, ist die Erhebung von Daten zu ausschließlich statistischen Zwecken nach dem derzeitigen Erkenntnis- und Erfahrungsstand verfassungsrechtlich unbedenklich." Damit entspricht das Urteil in seinen Begründungen der immer wieder von Seiten der Sozialforscher vorgetragenen Position: Nicht Datenschutz, sondern Datensicherung sind bei sozialwissenschaftlichen Erhebungen das Problem.

Nach dem Volkszählungsurteil des BVG folgt aus dem Recht zur informationellen Selbstdarstellung prinzipiell die Zustimmung eines „Betroffenen" bei der Benutzung seiner Angaben. Bei nur statistischem Interesse sind relevante Rechtsgüter nicht betroffen. Insbesondere bei der üblichen Umfrageforschung wird ja gar nicht personenbezogen verarbeitet, sondern lediglich in der Phase der Datensammlung vorübergehend zu Kontrollzwecken ein Abgleich von Fragebögen und Kontrollmitteln für die Arbeit des Interviewers vorgenommen. Daraus folgt nach der Begründung des Volkszählungsurteils: Wenn die Daten anonym verarbeitet und gesichert werden, bedarf es keiner „informierten Einwilligung" des „Betroffenen". Zunächst und zuvorderst folgt aus der Begründung des Volkszählungsurteils eine Privilegierung der Forschung als Grundlage für die Steuerung der gesellschaftlichen Entwicklung. Mit einer einseitigen Auswahl von Worten wird aber in der Datenschutzdiskussion bislang ein gegenteiliger Eindruck bewirkt.

5 Anwendungswirrwarr als Folge der Abwertung von Forschung

Mit der „offensiven" Auslegung der Regeln zum Datenschutz ist eine erhebliche Rechtsunsicherheit entstanden. Meist wird gegen exzessive Auslegung – also beim Vollzugsexzeß – keine Verwaltungsgerichtsklage eingelegt, weil Forschung bis zum Ende des Rechtsstreites schon veraltet wäre. So fehlt dann vorläufig in diesem Bereich, was in anderen Anwendungsgebieten von Gesetzen ein Korrektiv ist: ludicatur.

Beispiele für phantasievolle Ausdehnung des Datenschutzes bietet der Datenschützer des Landes Nordrhein-Westfalen, Weyer. Herr Weyer hat die Meinung vertreten, daß die Eintragungen in den privaten Notizbüchern von Polizeibeamten dem Datenschutz unterliegen, so daß er als Landesbeauftragter die Löschung dieser Eintragungen veranlassen kann. Der gleiche Datenschützer hielt es aber für unbedenklich, alle Führerscheininhaber für eine Weile in einer Datei getrennt zu speichern und diesen Personenkreis auf unfallfreies Fahren hin zu beobachten. Damit soll die Basis für Auflagen zur Nachschulung geschaffen werden. Man kann ja durchaus der Ansicht sein, daß der hohe Zweck der Verringerung von Unfallzahlen eine solche Individualbeobachtung auf der Grundlage von Dateien rechtfertigt, aber dann wäre doch wohl auch eine größere Effizienz der Polizei in der Bekämpfung von Verbrechen ein zu fördernder Zweck.

Herr Weyer vertrat auch die Ansicht, die Benutzung des Hausmülls „zum Gewinnen von Daten" sei rechtswidrig, weil aus dem Hausmüll Rückschlüsse auf den Haushalt gezogen werden können. Richtig. Aber seit wann ist das Durchwühlen des Hausmülls schon eine Datei? Es kann höchstens sein, daß irgendwann einmal eine Datei daraus wird, die personenbezogen verarbeitet werden könnte. Von da ab greift der Datenschutz.

In Baden-Württemberg sollte die Verwaltung des Landes sich der modernen Telekommunikationstechniken, einschließlich des Bildschirms, bedienen. Hiervor warnte die Datenschützerin des Landes, Frau Dr. Leuze. Das Landessystemkonzept – das ist der Plan für die Einführung der Telekommunikation in die Amtsstuben – fordere den Datenschutz aufs äußerste heraus, denn es sähe die ganze Verwaltung als Informationseinheit. Insofern ist Frau Leuze völlig zuzustimmen, denn der Datenschutz als Regelungssystem für den Datenaustausch würde nicht greifen, wenn eine Landesregierung sich mit allen nachgeordneten Ämtern als eine einzelne Informationseinheit definierte. Dies ist zweifellos wider den Sinn und den Buchstaben des Datenschutzgesetzes. Allerdings ist nicht einzusehen, warum insgesamt die Erledigung von Vorgängen nicht von amtsübergreifender Kommunikation begleitet sein soll.

In Rheinland-Pfalz, wo der Datenschutz durch eine Kommission des Parlaments durchgeführt wird, wurde eine pädagogische Untersuchung verboten. Bei dieser sollte der Zusammenhang zwischen Schülerleistungen und Intelligenzquotienten von Schülern ermittelt werden. Das Verbot wurde damit begründet, es könne bei einem Einbruch Intelligenz und Leistung eines bestimmten Schülers erfahren werden. Abgesehen von der Frage, ob Einbrecher für dieses Motiv vorstellbar sind, wird hier lediglich über eine mögliche Verletzung der Datensicherung geurteilt, was nach dem Sinn des Gesetzes und dem Spruch des BVG etwas anderes als Datenschutz ist.

Wenn die Datenschützer selber schon so wenig voraussagbar sind in ihrer Art der Ausdeutung von Gesetzen, dann werden nachgeordnete Behörden erst recht in ihrem Verhalten kapriziös. Ein Beispiel dafür ist das Einwohnermeldeamt der Stadt Köln. Eine ehemalige Schülerin, die ein Klassentreffen organisieren wollte, erbat von diesem Amt die Adressen früherer Mitschülerinnen. Von anderen Einwohnermeldeämtern erhielt sie diese auch, nicht jedoch vom Kölner Amt. Die gleiche Stadtverwaltung hält jedoch eine personenbezogene Datei aller SPD-Mitglieder. Und versendet in Abständen an diese aufgrund der städtischen Datei alle Arten von Mitteilungen. Und die gleiche Stadtverwaltung hat es auch vor einigen Jahren erlaubt, daß eine Kölner Partei eine Abgleichung ihres Mitgliederregisters mit dem Register für die abgegebenen Stimmen nach Bezirk unterteilt vorgenommen hatte. Dadurch wäre es möglich gewesen, alle Parteimitglieder zu identifizieren, die am Wahltag ihr Wahlrecht nicht ausgeübt hatten.

Zur Untersuchung der Sozialstruktur einer Industriestadt in Südwestdeutschland wollte ein Wissenschaftler die Einkommensteuerlisten auswerten, die sich in den Finanzakten des Landesarchivs befinden. Dabei interessierten ihn die Berufsangaben und die jeweils zu versteuernden Einkommen – nicht die einzelnen Personen. Die Finanzbeamten des Landes Baden-Württemberg haben aber die Akten unbefristet gesperrt. In der gleichen Stadt wollte ein Doktorand die nationalsozialistische Machtergreifung untersuchen. Hierzu wollte er die

Personalakten bereits verstorbener Beamter einsehen. Die Akteneinsicht wurde verweigert bzw. für einen Tag 120 Jahre nach der Geburt des Betroffenen zugesagt. In beiden Fällen ist die Geringschätzung des wissenschaftlichen Interesses im Vergleich zu anderen schutzwürdigen Sachverhalten offensichtlich.

Dr. Schäfer vom Landeskriminalamt in Bremen wollte prüfen, ob der Parapsychologe Bender zu recht den Doktortitel führt. 1982 wandte er sich schriftlich unter Angabe seines Dienstranges und mit der Postadresse Landeskriminalamt Bremen an die Bibliothek der Universität Freiburg, um sich dort nach einer Dissertation von Dr. Bender zu erkundigen. Diese einfache Verwaltungsfrage löste eine Anfrage des Rektors der Universität an den Senator für Inneres aus. Statt der einfachen Beantwortung durch die Bibliotheksverwaltung wollte der Rektor vom Vorgesetzten des Dr. Schäfer wissen, ob die erbetene Auskunft zur rechtmäßigen Erfüllung der in die Zuständigkeit des Landeskriminalamtes Bremen liegenden Aufgaben erforderlich sei. Dabei bezog er sich auf § 10 des Datenschutzgesetzes aus Baden-Württemberg.

Die wichtigste Kontroverse, an der die Geringschätzung der Forschung deutlich wird, ist der Kampf um das sogenannte Krebsregister. Nachdem Laboruntersuchungen nicht den erhofften raschen Erkenntniszuwachs bei der Erklärung und der Bekämpfung des Krebses brachten, richten sich jetzt viele Hoffnungen in der Medizin auf die Epidemiologie. Entstehung und Verlauf von Krebserkrankungen sind sicherlich ein Gegenstand, der nur durch die Interaktion mehrerer Faktoren erklärt werden kann. Das Instrument zur Erfassung dieser Faktoren sollte ein Krebsregister sein, in dem alle Fälle gespeichert werden, die als Krebs diagnostiziert sind. Nach der Speicherung wird dann verfolgt, wie sich die Krankheit entwickelt. Durch statistische Kausalanalysen kann dann eingekreist werden, welche Kombination von Umständen von besonderer Bedeutung ist. Insbesondere für die seltenen Fälle von Krebs gibt es für eine solche epidemiologische Massenbetrachtung keine Alternative. Bisher haben die Datenschützer die Anlage eines solchen Registers verhindert. Auf die Begründung der Mediziner, daß es zu einem solchen Forschungsinstrument keine Alternative gäbe, erhielten sie die pauschale Antwort: Dann sollten die Wissenschaftler sich eben etwas einfallen lassen.[31])

Auf diese Art und Weise antwortete auch der Abgeordnete Hirsch auf meinen Appell, bei der Novellierung des Datenschutzgesetzes doch an die Erfordernisse der empirischen Sozialforschung als ebenfalls schutzwürdiges Gut zu denken. Er schrieb mit Datum vom 10. 3. 1986: „Es mag sein, daß sie (diese Grundsätze) insofern Probleme bereiten, als manche liebgewordene bisherige Formen der Datenerhebung und Datenverarbeitung nicht mehr fortgesetzt werden können. Auch die empirische Sozialforschung muß also Methoden kritisch untersuchen, ob ihre bisherigen Erhebungsmethoden gleichsam naturgesetzlich unabänderlich sind." Mit der Wendung „liebgeworden" wird suggeriert, daß der hohe Stellenwert, den persönliche Befragungen der empirischen Sozialforschung haben, ein bloßer Ausdruck der Bequemlichkeit oder der Einfallslosigkeit von Sozialwissenschaftlern sei. Selbstverständlich verfügt die empirische Sozialforschung über eine breite Palette von

[31]) Ruth Leuze, Datenschutz und Krebsregister, in: Das öffentliche Gesundheitswesen, Bd. 43, 1981, S. 583–587.

Forschungstechniken. Wenn dennoch das Interview und andere Formen der persönlichen Befragung quantitativ eindeutig vorherrschen, dann ist das nicht Ausdruck von Unkenntnis über mögliche Alternativen, sondern Folge der Eignung dieses Instruments für die zu ermittelnden Sachen. Angesichts der sehr hohen Kosten gerade der persönlichen Interviews besteht auch eine hohe Motivation, auf andere Verfahren auszuweichen – wenn die Sache es rechtfertigt.[32]) Die Geringschätzung der Wissenschaft wird aber aus der bloßen Unterstellung ersichtlich, die Vorliebe für das Interview könnte doch ein Ausdruck von Gedankenlosigkeit sein.

6 Zum Regelungsbedarf

Zu bezweifeln ist generell, daß ein besonderer Regelungsbedarf bei der Forschung gegeben ist. Aus den USA wird berichtet, daß es dort etwa 200 Wissenschaftler und über 1000 Studenten und Doktoranden gibt, die Public Use Samples nutzen. Bisher ist noch kein einziger Fall auf der ganzen Welt bekannt geworden, daß solche Public Use Samples mißbräuchlich entanonymisiert worden seien.[33]) Und auch aus der Umfrageforschung ist bisher noch kein einziger Fall eines mißbräuchlichen Umgangs mit vertraulichen Daten bekannt geworden, der einen zusätzlichen Regelungsbedarf ergeben hätte.

Von den wenigen mißbräuchlichen Fällen sei erwähnt der Verkauf der Namen, die in einer Elite-Befragung ermittelt wurden, an eine Firma, die Anlagemöglichkeiten vertrieb. Dies war ein klarer Verstoß gegen geltendes Standesrecht und ein Bruch der Vertraulichkeit. Zudem ist fraglich, ob dies ein Fall der Verletzung von Datenschutz war, wenn diese Namen ohnehin aus allgemein zugänglichen Registern zusammengestellt worden waren.

Demgegenüber aber schreibt der Abgeordnete Burkhard Hirsch am 10. 3. 1986 an mich: „Es gibt bemerkenswerte Fälle, in denen die Wissenschaftsklauseln des Datenschutzgesetzes bedenkenlos mißbraucht worden sind. Das wird z. B. in einem Bericht der baden-württembergischen Datenschutzbeauftragten, Frau Dr. Leuze, für das sogenannte ‚Zentralinstitut für seelische Gesundheit' in Mannheim eindrucksvoll dargestellt. Das kann so nicht hingenommen werden." Bis heute habe ich von den Befürwortern einer strengeren Reglementierung der Wissenschaft durch Datenschutz nur diesen einen Fall genannt erhalten. Frau Leuze habe ich dann entsprechend um weitere Auskünfte für diesen von Herrn Hirsch als bedenkenlosen Mißbrauch der Wissenschaftsklausel der Datenschutzgesetze genannten Fall gebeten. Sie verwies auf ihren zweiten Tätigkeitsbericht.[34]) Von verschiedenen Beanstandungen wie derjenigen der Basisdokumentation der psychiatrischen

[32]) Siehe hierzu die Statistik in der jährlichen Dokumentation der empirischen Sozialforschung im deutschsprachigen Bereich, Zentralarchiv für empirische Sozialforschung: Empirische Sozialforschung, München, beginnend mit 1969 jährlich. Eine analytische Erklärung der Vorliebe für verschiedene Arten von Daten findet sich in Erwin K. Scheuch, Die wechselnde Datenbasis der Soziologie, Stuttgart 1977, S. 5–41. Siehe ferner ders., Die Weiterentwicklung des Datenschutzes als Problem der Sozialforschung, in: Max Kaase et al. (Hrsg.), Datenzugang und Datenschutz, Frankfurt 1980, S. 252–275.

[33]) David H. Flaherty. Privacy and Government Data Banks – An International Perspective, London 1979.

[34]) Datenschutz für unsere Bürger, 2. Tätigkeitsbericht der Landesbeauftragten für den Datenschutz, 1981, S. 10–30.

Landeskrankenhäuser in Baden-Württemberg ist diejenige des „Zentralinstituts für seelische Gesundheit" in Mannheim die wichtigste. Soweit Frau Leuze beanstandet, daß die ärztliche Schweigepflicht durch Weitergabe von Daten insbesondere psychiatrischer Art verletzt wurde, soll das hier nicht interessieren; das wird durch das Standesrecht bereits geregelt.

Seit 1975 existiert das Zentralinstitut als ein Institut der Landesregierung. Neben vielfältigen einzelnen Projekten führt das Institut auch das „Psychiatrische Fallregister", in dem inzwischen Daten von 26000 Personen gespeichert sind. Im Prinzip handelt es sich um eine Datei analog derjenigen des Projekts „Metropolit" oder dem prospektiven Krebsregister. Aufgenommen werden fortlaufend Informationen über alle im Stadtkreis Mannheim wohnenden psychisch Kranken, die entweder mit dem Zentralinstitut selbst, seinem Konsiliar-, Nacht- oder Notfalldienst oder mit einer größeren Zahl von psychiatrisch tätigen Institutionen und niedergelassenen Nervenärzten im Raum Mannheim Kontakt haben. Die Datenübermittlung erfolgt auf der Rechtsgrundlage eines Erlasses durch das Sozialministerium des Landes im Jahr 1975.[35]) Zweck des Registers ist, epidemiologische Forschung zu ermöglichen.

Bei der Übermittlung der Daten von seiten der Nervenärzte und kooperierenden Kliniken wird der Name des Falles genannt. Vor der Speicherung der Daten wird jedoch der Name ersetzt durch eine Identifikationsnummer (die „I-Zahl"). Sie setzt sich zusammen aus dem Geburtstag, dem Geschlecht, den Anfangsbuchstaben des Geburtsnamens und dem, was Frau Leuze „Mehrlings-Eigenschaft" nennt. Im Prinzip ist das also eine Zahl, die aufgebaut ist, wie das Allgemeine Personenkennzeichen aufgebaut sein sollte. (Seine Einführung scheiterte damals am öffentlichen Widerstand.) Grundsätzlich gibt das Zentralinstitut diese Dateien nicht an Dritte, sondern erstellt gewünschte Auswertungen in Form von Computerausdrucken.

Frau Leuze beanstandete, daß eine solche Fallsammlung nicht erlaubt sei. Das Landesdatenschutzgesetz erlaube lediglich die Speicherung personenbezogener Daten zu wissenschaftlichen Zwecken, wenn es um ein bestimmtes Forschungsvorhaben geht. Eine epidemiologische Forschung sei jedoch in der Forschungsabsicht unbestimmt. Dies ist die gleiche Argumentation, die bereits im Falle des hessischen Datenschützers Simitis angeführt wurde. Beim BDSG bedeutet sie eine Verschärfung durch bloße Interpretation; im Falle Baden-Württembergs scheint diese restriktive Auslegung erlaubter wissenschaftlicher Zwecke durch das Landesgesetz näher gelegt zu werden. Jedenfalls wird durch Frau Leuze nicht so sehr ein Mißbrauch im Umgang mit personenbezogenen Daten beanstandet als vielmehr eine Methode der Forschung verboten: Epidemiologie. Selbstverständlich kann man darüber rechten, ob bei der Art der Speicherung der Daten dem Erfordernis der Datensicherung ausreichend Rechnung getragen wurde. Aber interessanterweise ist das kein Kritikpunkt.

Ein weiteres beanstandetes Projekt war die Untersuchung „Psychische Erkrankung und soziale Isolation bei älteren Menschen in Mannheim". Seit 1978 führt das Zentralinstitut anhand einer Zufallsstichprobe von über 65jährigen Einwohnern eine Felduntersuchung

[35]) Ruth Leuze, Tätigkeitsbericht, S. 13.

durch. Die Interviewbogen sind anonymisiert, aber es gibt eine getrennt aufbewahrte Namensliste, durch die Personenbezug über die Codenummer des Interviews möglich wird. Ruth Leuze ordnete im Sommer 1981 die Vernichtung der Adressenliste an. Dem wurde von seiten des Zentralinstituts unter Verweis auf den Wunsch einer Nachfolgeuntersuchung widersprochen. „Die vage Möglichkeit, eine Adressenliste könne vielleicht später einmal für eine weitere Untersuchung verwendet werden, rechtfertigt keine Aufbewahrung auf längere Zeit."[36]) Eine solche Vernichtung kann eigentlich nur mit Datensicherheit begründet werden, und daß diese im vorliegenden Fall gefährdet war, ist aus dem Untersuchungsbericht nicht ersichtlich.

Wie eng Ruth Leuze ohne Verständnis für Forschung und bei fehlerhafter Güterabwägung den Rahmen für Forschung ziehen will, geht aus ihrer Reaktion auf den Gesetzentwurf der Landesregierung zur zweiten Änderung des Landesdatenschutzgesetzes hervor.[37]) Ihren besonderen Zorn bringen dabei wieder epidemiologische Untersuchungen hervor. Dabei müssen uns wieder die Beanstandungen, die sich gegen eine von ihr gedeutete Verletzung der ärztlichen Schweigepflicht richten, hier nicht interessieren. Alle die hier mitgeteilten Beanstandungen sind aber Fragen der Datensicherheit, und durch entsprechende Vorgehensweisen – insbesondere durch funktionale Trennung – kann ihnen Rechnung getragen werden.[38])

Das waren also die „bemerkenswerten Fälle, in denen die Wissenschaftsklauseln der Datenschutzgesetze bedenkenlos mißbraucht worden sind" (!) – so Dr. Hirsch in dem Brief vom 10. 3. 1986. Hier wurde insbesondere im Falle des Zentralinstituts in Mannheim nicht mißbraucht, denn es lag eine gesetzliche Ermächtigung vor, über deren Angemessenheit man dann allerdings miteinander streiten kann. Prinzipiell handelt es sich aber bei diesen „bemerkenswerten Fällen" eines „bedenkenlosen Mißbrauchs" der Datenschutzgesetze um den Versuch, auch von Frau Dr. Leuze, die epidemiologische Forschung unmöglich zu machen und Wiederholungsbefragungen über längere Zeit zu verbieten.

Wenn dies Fälle von Mißbrauch sind, dann aus der Sicht der Forschung eher von seiten der Datenschützer, welche die Gesetze in „offensiver" Form auslegen. Wie sehr sie sich selbst als erste und letzte Instanz bei Entscheidungen über Datenschutz in Konkurrenz zu anderen Rechtsgütern verstehen, geht nicht zuletzt aus dem dritten Tätigkeitsbericht von Frau Dr. Leuze hervor, wenn der Ton beachtet wird, in dem sie die bevorstehende zweite Novellierung des Landesdatenschutzgesetzes kommentiert.

Es geht nicht um Regelungsbedarf, wohl aber um die Gefahr, daß ganze Forschungsprobleme nicht mehr untersuchbar werden, würde der Wissenschaftsparagraph 3a in der jetzigen Formulierung des Vorschlags für eine Gesetzesnovelle wirklich rechtens.

Ein Beispiel mag dies verdeutlichen. Im Zusammenhang mit einer Standardumfrage stellten wir vor Jahren fest, daß die Mehrheit der Bürger im Umgang mit Behörden auf Hilfe ange-

[36]) A. a. O., S. 21.
[37]) Datenschutz für unsere Bürger, 3. Tätigkeitsbericht der Landesbeauftragten für den Datenschutz, 1982, insbesondere S. 21 ff.
[38]) A. a. O., S. 32–37, insbesondere S. 33–34.

wiesen sind.[39]) Dem gingen wir noch vor Einrichtung der Datenschutzbehörden und entsprechend den damals nicht problematisierten Usancen nach, indem wir in einer zweiten Untersuchung eine Stichprobe von 100 Behördenstellen zogen.[40]) Eine Behördenstelle war für uns jede Amtsstube, in der ein Kontakt zwischen Sachbearbeiter und Klienten erfolgte. Mit Zustimmung der Stadtverwaltung Köln und in jedem Falle des Klienten (mündlich!) protokollierten wir den Verlauf der Gespräche. Nach Abschluß der jeweiligen Behördenkontakte wurden im zweiten Abschnitt der Erhebung die Klienten befragt. Dabei wollten wir auch wissen, wer im Falle der Hilfsbedürftigkeit mit Erfolg um Hilfe gebeten werden konnte. In über 60% der Behördenkontakte war Hilfe durch oder über Bekannte notwendig. Dabei waren das in einem Drittel der Fälle selbst Beamte, die über Netzwerke der Bekanntschaft vermittelt wurden.

In diesem Herbst wollen wir diesen Netzwerken weiter nachgehen. Nicht zuletzt werden wir angeregt durch das Projekt von Granovetter „The Strength of Weak Ties", nach dem in den USA der größte Teil erfolgreicher Arbeitssuche durch Netzwerke persönlicher Bekanntschaft erfolgte.[41]) Insbesondere weniger qualifizierte Arbeitskräfte waren Teil von Netzwerken der Bekanntschaft („closed networks"), die über einen homogenen Personenkreis nicht hinausführten. Quantitativ wissen wir aus der Bundesrepublik über die Art der Netzwerke nichts. Wir haben unsere eigene Untersuchung vorbereitet durch eine Befragung in diesem Frühjahr[42]), in der Personen nach der Art der Menschen gefragt werden, die ihren direkten und indirekten Bekanntenkreis ausmachen. Darauf aufbauend wollen wir jetzt im Herbst 1986 in einem sogenannten Schneeballverfahren unter anderem prüfen, wie eine solche persönliche Bekanntschaft als Teil des Netzwerks von beiden Seiten aussieht. Bei einem Schneeball-System werden die Befragten gebeten, Namen und Adresse der Person zu nennen, die nach ihrer Meinung als nächste zu befragen ist. Es versteht sich, daß bei der späteren Analyse die Vernetzung erhalten bleiben muß, was durch Identifikationsnummern nach der dann erfolgten Vernichtung der Namen geschieht. Warum soll in Zukunft eine solche Forschung unmöglich sein? Ist nicht die Kenntnis solcher Sachverhalte ein Wissen, wie es das BVG in seinem Volkszählungsurteil für einen Sozialstaat ausdrücklich fordert! Wo besteht hier die Gefahr, daß ein schutzwürdiges Rechtsgut verletzt würde?

Inzwischen sind in sehr vielen Ländern Datenschutzgesetze erlassen worden.

Gesetze in Kraft:

Bundesrepublik Deutschland	Luxemburg
Dänemark	Neuseeland
Frankreich	Norwegen
Großbritannien	Österreich
Island	Schweden
Israel	Vereinigte Staaten
Kanada	von Amerika

[39]) Erwin K. Scheuch/Wolfgang Bick/Paul J. Müller, Das Formular – Ausdruck und Vehikel der Bürokratisierung unseres Alltags, Taunusstein 1980.
[40]) Erwin K. Scheuch/P. J. Müller, Mikrowelten – Die Bedeutung der Vernetzung von Mikrowelten für die Vermittlung zwischen Alltag und Institution, Forschungsbericht, Köln 1986.
[41]) M. S. Granovetter, The Strength of Weak Ties, in: American Journal of Sociology, Vol. 78, Nr. 6, S. 1360–1380.
[42]) Es handelt sich um den Allbus 1986.

Für 1985/86/87 erwartete Gesetze:

Australien	Italien
Belgien	Niederlande
Finnland	Portugal
Griechenland	Spanien

Länder, in denen am Datenschutzrecht gearbeitet wird, die Verabschiedung eines Gesetzes ist jedoch für später als 1987 zu erwarten:

Brasilien	Schweiz
Irland	Türkei
Japan	

In keinem dieser Länder ist bekannt geworden, daß der Gesetzgeber einen besonderen Handlungsbedarf für die spezielle Reglementierung der Forschung sieht. Eine Reihe von Gesetzen gehen eingehend auf die Bedürfnisse der Forschung ein und fordern nicht viel mehr als eine besondere Vertraulichkeit und Sorgfalt im Umgang mit Daten. Dieser Selbstverständlichkeit würde bei uns in der Bundesrepublik auch kein verantwortlicher Forscher widersprechen.

„Ein Fall von Mißbrauch ist nicht bekannt" ist die Überschrift eines Berichts über Schweden – einem Land mit einer längeren Tradition der Beaufsichtigung von Forschung im Hinblick auf „Sozialverträglichkeit".[43]) Der Artikel war noch geschrieben bevor das Projekt „Metropolit" von Janson zu einem Medienereignis hochstilisiert wurde – bekanntlich hatten die Datenschützer selbst an der Datensammlung nichts auszusetzen. Aber auch nach dem Projekt „Metropolit" ist das Klima gegenüber der Datenforschung nicht entfernt mit dem in der Bundesrepublik vergleichbar. Was ist denn bei uns so anders, daß die Gesetze so anders sein müssen?

7 Die Notwendigkeit der Güterabwägung bei der Anwendung von Datenschutz

Die Medien verfahren bei uns nicht sehr viel anders als dies im Zusammenhang mit dem Projekt „Metropolit" für Schweden berichtet wurde. So wählte das Wochenblatt „Die Zeit" für einen Aufsatz über den Bericht des bayerischen Datenschutzbeauftragten Stollreiter die Überschrift „Kinder lautlos erfaßt" – ganz analog zur Überschrift, die eine heimliche Erfassung von Daten beim Projekt „Metropolit" in Stockholm behauptete.[44]) Mit besonderer Ausführlichkeit wurde in dem Bericht die Datenweitergabe für polizeiliche Zwecke dargestellt mit dem Tenor: Wir werden alle „lautlos" überwacht. Aber ungeachtet solcher Präsentationen, will sich bei der Bevölkerung kein Gefühl der Dringlichkeit einstellen, daß von der Verschärfung des Datenschutzes die Freiheitlichkeit dieser Republik abhänge. Es gibt keine

[43]) Sten Johanson, Ein Fall von Mißbrauch ist nicht bekannt, in: Das Parlament, 29. April 1986, S. 14.
[44]) Heide Meisel, Kinder lautlos erfaßt, in: Die Zeit, 14. Februar 1986.

Parallele zwischen der weltanschaulichen Intensität von Argumentationen unter den Berufspolitikern in Bonn und den Einstellungen in der Bevölkerung.

1985 befragte die Gesellschaft für Konsum- und Marktforschung, Nürnberg, einen repräsentativen Querschnitt der Bundesrepublik nach den Hauptquellen für Sorgen. In der offen gestellten Frage kam unter den mit nennenswerter Häufigkeit erwähnten Ängstlichkeiten der Datenschutz nicht vor. Das gilt auch für alle früheren Jahre bis 1979, als diese Frage zuerst routinemäßig gestellt wurde. Dabei schwankt die Nennung für andere Themen enorm zwischen 37 und 0 (Sicherung der Energieversorgung) oder 9 bis 41 (Umweltschutz). Datenschutz war nie ein wichtiges Thema.[45]) Das gleiche gilt auch auf europäischer Ebene für den ganzen Zeitraum zwischen 1973 und 1983.[46])

Speziell zum Thema Datenschutz wurde in einer Erhebung im Jahr 1985 gefragt. 64% der Befragten sahen beim Datenschutz keine Probleme. Eigene schlechte Erfahrungen mit mangelhaftem Datenschutz oder Datenmißbrauch wurden von etwa 6% der Bevölkerung genannt. „Berücksichtigt man hierbei, daß Befragte oft dazu neigen, eigene Erfahrungen vorzugeben, wenn sie mit ‚Problem-Themen' konfrontiert sind, so muß diese Zahl als sehr gering eingestuft werden."[47]) Offensichtlich liegt hier eine Eigendynamik bei denjenigen vor, die sich selbst den Datenschutz als wichtiges Thema aussuchten.

Voraussetzung dafür ist wohl auch, daß zwei Gesetzgebungen bei uns fehlen, so daß der Datenschutz als Möglichkeit benutzt werden kann, um diese Gesetzeslücke zu schließen. Es fehlt einmal an einem allgemeinen Gesetz über Persönlichkeitsschutz, sowie an einem weiteren Gesetz über das Recht, sich selbst zu informieren. Der Wortlaut der Datenschutzgesetze eignet sich nicht sonderlich, um insbesondere die erste der beiden Gesetzeslücken auszufüllen. So kommt es dann zu sehr erfinderischen Deutungen mit dem Ziel einer möglichst extensiven Anwendung von Begriffen oder sogar deren Umfunktionierung.

Das Fehlen einer Regelung für das Recht zur Information wird allerdings bei der Erörterung über Datenschutz in der Bundesrepublik durchweg ausgeblendet. In den Vereinigten Staaten war dagegen der „Freedom of Information Act" noch viel mehr Gegenstand der öffentlichen Aufmerksamkeit als die „Data Protection Laws".

Nur in einer Hinsicht ist eine gesetzliche Festlegung des Rechts auf Information erfolgt und zum Teil des Datenschutzes geworden. Nach dem BDSG kann ein Betroffener beantragen, Auskunft über die gespeicherte Information zu erhalten (§ 4 BDSG). Diese Pflicht zur Erteilung von Information betrifft nicht nur Behörden, sondern alle speichernden Stellen. Und diesem Auskunftsrecht für den Betroffenen entspricht ein Berichtigungsanspruch selbst für solche Fehler, die der Speichernde für unbedeutend hält. Die Freiheit der Information ist bei uns also nur als Recht zur Information über gespeicherte Daten der eigenen Person geregelt. Dies ist jedoch ein viel zu enger Begriff von „Freedom of Information".

[45]) GfK Marktforschung, September 1985.
[46]) Amt für amtliche Veröffentlichungen der Europäischen Gemeinschaften, Die Europäer über sich selbst – Zehn Jahre Euro-Barometer 1973–1983, Luxemburg 1983, S. 38f.
[47]) Vgl. Isaac W. Eberstein und Manfred Küchler, Trotz detaillierter Fragen: Privatsphäre bleibt geschützt, in: Das Parlament, 26. April 1986, S. 15.

Ein Aufarbeiten der amerikanischen Entwicklung würde es erleichtern, den Datenschutz als nur ein Element unter anderen für eine Informationspolitik zu verstehen. Ist dies nicht so der Fall, dann ist de facto bei uns die Informationspolitik reduziert auf eines ihrer Elemente. Für die Informationspolitik ist aber ein System von Zielen und Maßnahmen anzusprechen, die in Abwägung bei Konflikten regeln, wo Transparenz den Vorrang vor Schutzwürdigkeit und wo Schutzwürdigkeit den Vorrang vor Transparenz hätte. Denn ohne Transparenz ist offensichtlich ein Gemeinwesen nicht als Demokratie organisierbar, ohne Schutz aber ein Staat nicht als eine humane Gesellschaft. Die Forschung muß hier als Teil des Rechtes auf Information verstanden werden – als Information der Gesellschaft über sich selbst. Wie das BVG in seinem Urteil zur Volkszählung erklärte: Ohne Information ist ein reformorientierter Sozialstaat nicht möglich.

Erwin Südfeld
Statistisches Bundesamt, Wiesbaden

Anonymisierungsstandards und generelle Abwicklungsregelungen für Anforderungen nach anonymisierten Einzelangaben im Statistischen Bundesamt

Einleitung

In den vorstehenden Aufsätzen in diesem Band wird das Spannungsfeld ausgeleuchtet, in dem sich die Amtliche Statistik bei der Entscheidung über die Weitergabe anonymisierter Einzelangaben zu bewegen hat. In dem Beitrag von Müller und Hauser werden die Anforderungen der Wissenschaft an anonymisiertes statistisches Einzelmaterial formuliert. In den Beiträgen von Paaß, Kühn und Kirschner die technischen Möglichkeiten und insbesondere die Grenzen einer Anonymisierung statistischen Einzelmaterials diskutiert. Im folgenden sollen nun die Anonymisierungsstandards und generelle Abwicklungsregelungen für Anforderungen nach anonymisierten Einzelangaben im Statistischen Bundesamt (StBA) dargestellt werden.

1 Rechtliche Rahmenbedingungen des § 11, insbesondere Abs. 5 Bundesstatistikgesetz (BStatG)

Ausgangspunkt aller Betrachtungen zum Problem der Weitergabe von anonymisierten Einzelangaben hat für einen Bundesstatistiker das geltende Recht zu sein, das hinsichtlich der Geheimhaltung von Einzelangaben im § 11 BStatG kodifiziert ist.[1]) Nach dem Grundsatz der Geheimhaltung in § 11 Abs. 1 BStatG sind Einzelangaben über persönliche und sachliche Verhältnisse, die für eine Bundesstatistik gemacht werden, von den mit der Durchführung von Bundesstatistiken Betrauten geheimzuhalten. Die statistische Geheimhaltung steht in untrennbarem Zusammenhang mit der statistischen Auskunftsverpflichtung und dient dazu,

- den einzelnen vor der Offenlegung seiner persönlichen und sachlichen Verhältnisse zu schützen,
- das Vertrauensverhältnis zwischen den Befragten und den Statistischen Ämtern zu erhalten und
- die Zuverlässigkeit der gemachten Angaben und die Berichtswilligkeit der Befragten zu gewährleisten.

[1]) Vgl. Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz) vom 14. März 1980, BGBl. I S. 289 und insbesondere die Begründung zum Gesetz, Bundestagsdrucksache Nr. 8/2517 vom 26. Januar 1979.

Es gibt Ausnahmen von diesem Geheimhaltungsgrundsatz, z. B. wenn der Betroffene einer Weitergabe zustimmt (§ 11 Abs. 1) oder wenn eine einzelstatistische Rechtsgrundlage die Weiterleitung – bei genauer Angabe des Empfängerkreises und der Art der Verwendung der Einzelangaben – zuläßt (§ 11 Abs. 3). Allerdings hat für den letzteren Ausnahmefall das Bundesverfassungsgericht in seinem Urteil zum Volkszählungsgesetz 1983 vom 15. Dezember 1983 sehr enge Grenzen gesetzt.[2])

Eine generelle Ausnahme vom Geheimhaltungsgrundsatz enthält der § 11 BStatG in seinem Abs. 5. Danach dürfen Einzelangaben, die so anonymisiert werden, daß sie Auskunftspflichtigen oder Betroffenen nicht mehr zuzuordnen sind, vom StBA und von den Statistischen Landesämtern übermittelt werden. Diese Bestimmung ist im Zuge der parlamentarischen Beratungen des Entwurfs eines BStatG 1980 eingefügt worden. Sie ist auf eine Beschlußempfehlung des Innenausschusses zurückzuführen, der damals die Auffassung vertreten hat, daß die Möglichkeit einer Deanonymisierung bereits anonymisierter Einzelangaben absolut nicht ausgeschlossen werden könne. Andererseits dürfe aus diesem Tatbestand nicht geschlußfolgert werden, daß eine Übermittlung anonymisierter Daten in jedem Fall ausgeschlossen sei. Vielmehr müsse vor Übermittlung anonymisierter Daten sichergestellt sein, daß die Möglichkeit der Deanonymisierung der nach den Kenntnissen der Statistischen Ämter ausreichend anonymisierten Einzelangaben zweifelsfrei ausgeschlossen ist.[3]) Der Wirtschaftsausschuß hat dazu angemerkt, daß er für wirtschaftsstatistische Daten keine hinreichende Anonymisierungsmöglichkeit sieht.

2 Elemente einer verwaltungspraktischen Umsetzung

Mit dieser rechtlichen Regelung war 1980 eine neue Situation geschaffen worden, auf die sich die Statistischen Ämter einzustellen hatten. Zur Erörterung der damit verbundenen Fragen haben die Leiter der Statistischen Ämter damals einen Gesprächskreis „Anonymisierung" eingerichtet, der sich auf gemeinsame Empfehlungen zur Anonymisierung statistischer Einzelangaben verständigt hat. Damit wurde dem Petitum des Innenausschusses, daß sich die Statistischen Ämter auf ein einheitliches Vorgehen einigen, in einem ersten Schritt nachgekommen. Vor dem Hintergrund dieser Empfehlungen des Gesprächskreises wurden die Anforderungen an die Anonymisierung statistischer Einzelangaben im StBA konkretisiert und Verfahrensregelungen für die Behandlung von Anforderungen nach anonymisierten Einzelangaben im StBA getroffen.

2.1 Die Anforderungen an die Anonymisierung statistischer Einzelangaben

Bei der Konkretisierung der Anforderungen wurden die Ergebnisse der wissenschaftlichen Diskussion zur Anonymisierung personenbezogener Einzelangaben berücksichtigt. Namentlich die Arbeiten des U. S. Federal Committee on Statistical Methodology und die

[2]) Vgl. Entscheidungen des Bundesverfassungsgerichts, Amtliche Sammlung, 65. Band, S. 65ff.
[3]) Vgl. Bundestags-Drucksache 8/3413, S. 13f.

Überlegungen, die in der DFG-Kolloquienreihe „Datenzugang und Datenschutz" insbesondere von Brennecke angestellt worden sind, waren eine wesentliche Grundlage für die Erarbeitung der Anforderungen.[4])

Diese Anforderungen sind als Empfehlungen formuliert, die bei der Übermittlung anonymisierter Einzelangaben durch das StBA in jedem Einzelfall geprüft werden sollen; ausdrücklich wird jedoch betont, daß darüber hinaus im konkreten Einzelfall einer Anonymisierung zusätzliche Maßnahmen notwendig werden können, d. h., die besonderen Belange des speziellen statistischen Materials, aus dem der jeweilige zu anonymisierende Einzeldatenbestand stammt, müssen entsprechend gewürdigt werden. Insoweit stellen die Anforderungen nur einen allgemeinen Bezugsrahmen bereit, der im einzelnen bei jedem Anonymisierungsfall weiter zu konkretisieren ist.

Für die Anonymisierung werden die nachstehenden Empfehlungen ausgesprochen:

- Der zu übermittelnde Datenbestand soll nur noch eine Stichprobe aus der jeweiligen Statistik sein.

  Begründung:
  Wird aus dem Ursprungsdatenbestand einer Statistik eine Stichprobe gezogen, so wird dadurch das Deanonymisierungsrisiko auch für die im Material verbleibenden Datensätze herabgesetzt, weil ein potentieller Angreifer nicht sicher sein kann, daß ein spezieller Einzeldatensatz enthalten ist, auch wenn er weiß, daß der Betroffene zur Statistik herangezogen wurde. Die Erfolgswahrscheinlichkeit eines Deanonymisierungsangriffs wird herabgesetzt.

- Der zu übermittelnde Datenbestand soll ein bestimmtes Mindestalter aufweisen. In der Regel sollen die Angaben durch eine neue Erhebung bereits überholt sein.

  Begründung:
  Je älter die Einzelangaben werden, desto geringer wird einerseits das Reidentifikationsrisiko, weil die Verfügbarkeit kompatiblen Zusatzwissens immer unwahrscheinlicher wird und das Interesse eines potentiellen Angreifers geringer wird. Zudem dürften sich im Zeitablauf die schutzwürdigen Belange des Betroffenen eher vermindern.

- Die Anordnung der Datensätze im zu übermittelnden Datenbestand muß systemfrei sein.

  Begründung:
  Werden die Einzeldatensätze systematisch angeordnet (z. B. nach räumlichen Gesichtspunkten) weitergegeben, erhöht sich das Reidentifikationsrisiko für alle

[4]) Vgl. Report on Statistical Disclosure and Disclosure-Avoidance Techniques, Statistical Policy Working Paper 2, Prepared by Subcommittee on Disclosure-Avoidance Techniques, Federal Committee on Statistical Methodology; U. S. Department of Commerce, Washington 1978 und Brennecke, Ralph Kriterien zur Operationalisierung der faktischen Anonymisierung, in: Kaase, Max, u. a. (Hrsg.) Datenschutz und Datenzugang; Konsequenzen für die Forschung, Königstein 1980, S. 158ff.

Datensätze, wenn der potentielle Angreifer das Organisationsprinzip erkannt hat oder es ihm bekannt ist.

- Direkte Identifikationsmerkmale (Name und Anschrift, Telefonnummer, Personen- und sonstige Kennummern) dürfen im Datenbestand nicht enthalten sein.

  Begründung:
  Anonymisierte Einzelangaben dürfen keine direkten Identifikationsmerkmale enthalten, da anderweitig direkt personenbezogene Daten vorliegen.

- Regionalangaben sollten nur als Typisierungsangaben im Datenbestand belassen werden, allerdings nur insoweit, als dadurch keine Regionaltypen mit weniger als 500000 Einwohnern identifiziert werden.

  Begründung:
  Regionalangaben sind (fast) immer als Zusatzwissen vorhanden. Je kleiner die – regional – identifizierte Population ist, desto größer ist demnach das Reidentifikationsrisiko.

- Jede Ausprägung eines einzelnen Merkmals soll mindestens fünffach besetzt sein.

  Begründung:
  In den Daten sollen keine Merkmalsausprägungen auftreten, die – für sich genommen – eine Population von weniger als fünf Betroffenen identifizieren.

- Sensible Merkmale (z. B. Daten zur Gesundheit, zur Einkommens- und Vermögenslage u. a. m.) sollen allenfalls klassifiziert übermittelt werden.

  Begründung:
  Für sensible Merkmale lohnt sich ein höherer Reidentifikationsaufwand. Außerdem wird bei einer Reidentifikation solcher Merkmale besonders stark in das informationelle Selbstbestimmungsrecht eingegriffen. Bei diesen Merkmalen ist deshalb besonderer Anonymisierungsaufwand erforderlich.

- Merkmale, über die sehr einfach Zusatzinformationen zu erhalten sind (z. B. Geburtsdatum u. a. m.), sollen ebenfalls nur klassifiziert übermittelt werden.

  Begründung:
  Über Merkmale, die häufiger im Zusatzwissen auftreten, steigt die Reidentifikationsmöglichkeit eines Einzeldatenbestandes. Bei diesen Merkmalen ist deshalb ein besonderer Anonymisierungsaufwand erforderlich.

- Die Kombinationen sensibler Merkmale und von Merkmalen, über die sehr einfach Zusatzinformationen zu erhalten sind, sollen mindestens dreifach besetzt sein.

  Begründung:
  Die Kombination von sensiblen und solchen Merkmalen, über die leicht Zusatzwissen zu erhalten ist, d. h. von Merkmalen, die einem besonderen Reidentifikationsrisiko unterliegen, müssen auf Mehrfachbesetzung überprüft werden. In diesen Kombinationen einmalig oder zweimalig auftretende Datensätze dürfen nicht im Einzeldatenbestand verbleiben.

Die vorstehend genannten Empfehlungen sind in der Zwischenzeit – im Rahmen einer Testphase – auf verschiedene Anonymisierungsprojekte angewandt worden. Dabei hat sich insbesondere gezeigt, daß die Offenkundigkeit bzw. Sensibilität eines Merkmals nicht generell – abstrahiert vom jeweiligen Merkmalszusammenhang – sondern nur im Einzelfall und unter Zugrundelegung des Wissens der Fachabteilung festgelegt werden kann. Durch die Verfahrensregelungen zur Behandlung von Anforderungen anonymisierter Einzelangaben im StBA ist zudem sichergestellt, daß diese Klassifikation der Merkmale nicht nur von der jeweils betroffenen Fachabteilung verantwortet, sondern ihr in einem „breiten Konsens" von allen betroffenen Abteilungen im StBA zugestimmt wird.

2.2 Technische Verfahren der Anonymisierung

Mit den Anforderungen ist zwar der Bezugsrahmen für die Anonymisierung statistischer Einzelangaben im StBA geschaffen worden; noch nicht festgelegt sind damit aber im einzelnen das technische Verfahren bzw. die möglichen Verfahrensvarianten, nach denen im StBA die Anonymisierung abgewickelt wird. Auch zur Frage der Anwendung verschiedener technischer Anonymisierungsverfahren hat die genannte Testphase feste Regelungen erbracht, die – kurz zusammengefaßt – dargestellt werden sollen.

Vorbedingung für alle entwickelten technischen Verfahren ist die Einhaltung der vorab dargestellten Anforderungen an die Anonymisierung. Andererseits ist darauf zu achten, daß die technischen Verfahren der Anonymisierung DV-mäßig abgewickelt und Personal sowie Maschinen in möglichst rationeller Weise eingesetzt werden können. Auf der Basis der Anforderungen wurden in der Testphase einige technische Verfahrensvarianten für unterschiedliche Bereiche der Anonymisierung von Einzelangaben entwickelt. Dabei ist zu unterscheiden, ob es sich um die Anonymisierung von

- Einzelangaben in Tabellen,
- von Einzelangaben in Form von Kurzdatensätzen oder
- von Einzelangaben in Form von Langdatensätzen

handelt.

2.2.1 Anonymisierung von Einzelangaben in Tabellen

In maschinell erstellten Tabellen mit Aggregatdaten können in einzelnen Feldern unter bestimmten Bedingungen Einzelfälle auftreten. Diese sind prinzipiell genauso zu behandeln wie Einzeldatensätze, die zur Weitergabe anstehen, d. h. ihre Anonymität muß gewährleistet sein.[5])

[5]) Vgl. Kühn, J. u. a. Zur technischen Weiterentwicklung des Statistischen Informationssystems, in: Wirtschaft und Statistik, 12/1984, S. 981 ff.

Im Statistischen Informationssystem des Bundes (STATIS-BUND) ist für Fallzahlentabellen ein entsprechendes DV-Programm enthalten, durch das die Tabellen mit Zufallsfehlern überlagert werden. Dazu werden alle Tabellenfelder mit annähernd normal verteilten Zufallszahlen bei Mittelwert 0 überlagert. Negative Tabellenfelder werden bei diesem Verfahren automatisch verhindert, es kann aber vorkommen, daß ein Tabellenwert mit echtem Wert null nach der Überlagerung einen Wert größer null hat. Das Programm ist derart konstruiert, daß die Gesamtmasse der Tabelle „im Mittel" erhalten bleibt. Außerdem ist gewährleistet, daß die tabellierten Verteilungen in ihrer Struktur unverändert bleiben. Allerdings können die schwach besetzten Felder, die jedoch statistisch auch nicht besonders aussagefähig sind, hohe relative Abweichungen aufweisen. Das Verfahren erlaubt es, den Überlagerungsfehler abzuschätzen und ggf. den Konsumenten der Tabelle als Qualitätsindikator mitzuliefern.

Die Anforderungen an die Anonymisierung werden – soweit sie in diesem Anwendungsbereich relevant sind – bei dem Verfahren in jedem Fall eingehalten, da es sich bei der Überlagerung der Tabellenfelder mit Zufallsfehlern nicht mehr um echte, sondern um „künstlich geschaffene" Einzelangaben handelt, sofern solche nach Überlagerung der Felder mit Zufallsfehlern überhaupt noch auftreten.

2.2.2 Anonymisierung von Einzelangaben bei Anforderung von Kurzdatensätzen

Werden von einem Konsumenten Einzeldatensätze unter explizitem Bezug auf den § 11 Abs. 5 BStatG angefordert, so wird im StBA bei solchen Datenanforderungen mit wenig umfangreichem Merkmalskatalog so verfahren, wie Kühn dies in seinem Beitrag zu diesem Band bereits dargestellt hat. Die Einzelheiten dieses Verfahrens brauchen hier nicht wiederholt zu werden.

Die Anforderungen an die Anonymisierung können in diesem Fall als Leitlinie für die Erstellung des Kurzbandsatzes und die Sortierfolge der Merkmale bzw. die Suche nach Strategien der „geeigneten" Zusammenführung herangezogen werden. Die Einhaltung der Empfehlungen ist verfahrensmäßig sichergestellt durch die zufällige Aggregation benachbarter Datensätze.

Gemäß dem Verfahren gibt es keine einmalig auftretenden Einzeldatensätze; jeder Datensatz ist mindestens dreimal vertreten.

2.2.3 Anonymisierung von Einzelangaben bei Anforderung von Langdatensätzen

Werden von einem Konsumenten Einzeldatensätze mit umfangreichem Merkmalskatalog angefordert, führt das oben beschriebene Verfahren in aller Regel zu unbefriedigenden Ergebnissen, weil bei großem Merkmalskatalog die zufällige Aggregation zu so großen Informationsverlusten führt, daß der Konsument mit den weniger wichtigen Merkmalen kaum mehr Auswertungen vornehmen kann, da sie in hohem Maße verfälscht werden.

Andererseits bietet gerade die Weitergabe von anonymisierten Einzelangaben bei Langdatensätzen eine Vielzahl von Ansatzpunkten zur Reidentifikation, wie die Experimente im Rahmen des Forschungsvorhabens der Gesellschaft für Mathematik und Datenverarbeitung (GMD) „Datenzugang, Datenschutz und Anonymisierung; Analysepotential und Identifizierbarkeit von anonymisierten Individualdaten" eindrucksvoll bestätigt haben.[6]) Paaß hat darüber ja in diesem Band ausführlich berichtet.

Gleichzeitig haben die Ergebnisse dieses Forschungsvorhabens gezeigt, daß – bei der gegenwärtigen Rechtslage – die Statistischen Ämter verpflichtet sind, bei der Weitergabe von anonymisierten Einzelangaben in Langdatensätzen zusätzliche Sicherungen vorzusehen. Diese Elemente sind im einzelnen:

1. Strikte Einhaltung des Prinzips der Einzelfallbearbeitung bei der Anforderung anonymisierter Einzelangaben in Langdatensätzen;

2. Umfassende Anonymisierungsmaßnahmen nach Maßgabe der Anforderungen an die Anonymisierung einschl. Reidentifikationsexperimenten gemäß Szenariotechnik;

3. Vertragliche Bindung des Empfängers anonymisierter Einzelangaben;

4. Eingehendes Abstimmungsverfahren im StBA vor der Übermittlung der anonymisierten Einzelangaben an den Empfänger.

ad 1.: Grundsatz der Einzelfallverarbeitung

Bei den gegenwärtigen rechtlichen Rahmenbedingungen sieht das StBA keine Möglichkeit, Public Use Samples für wissenschaftliche Zwecke aus verschiedenen Statistiken anzubieten, vielmehr reagieren wir nur auf einzelne Anforderungen von wissenschaftlicher Seite nach anonymisierten Einzelangaben. Diese Position ist wie folgt zu begründen. Public Use Samples, auch im eingeschränkten Sinne für wissenschaftliche Zwecke, machen nur dann Sinn, wenn der ganz überwiegende Teil der Merkmale einer Statistik im Datensatz erhalten werden kann. Bei Statistiken wie der Einkommens- und Verbrauchsstichprobe (EVS) oder dem Mikrozensus handelt es sich jedoch um jeweils mehrere Dutzend bzw. sogar hundert Eingabefelder in der Datensatzbeschreibung. Nach den Reidentifikationsexperimenten, die von Herrn Paaß durchgeführt worden sind, bieten solche Langdatensätze eine Vielzahl von Ansatzpunkten zur Reidentifikation. Die exakt berechneten Reidentifikationsraten, die zwar von Szenario zu Szenario je nach unterstelltem Zusatzwissen und sonstigen Ausgangsbedingungen schwanken, sind jedoch bei entsprechendem Zusatzwissen immer signifikant von null verschieden, so daß eine vollständige Weitergabe solcher Daten unter den gegebenen rechtlichen Rahmenbedingungen ausscheidet. Derzeit können wir das Problem nur so lösen, daß wir - möglichst viele - Eingabefelder, die für den jeweiligen Wissenschaftler nicht so von Interesse sind, aus dem Datensatz herausnehmen, um das Reidentifikationspotential zu reduzieren. Dies führt nahezu zwangsläufig zu einer Individual-

[6]) Vgl. Paaß, G. und Wauschkuhn, U., Datenzugang, Datenschutz und Anonymisierung; Analysepotential und Identifizierbarkeit von anonymisierten Individualdaten, München 1984.

behandlung, da für fast jede wissenschaftliche Problemstellung der unabdingbar erforderliche Datensatz unterschiedlich ausfällt. So sind jedenfalls unsere Erfahrungen aus den letzten drei, vier Jahren.

ad 2.: Umfassende Anonymisierungsmaßnahmen

Während die Anonymisierung von Einzelangaben in Tabellen und auch bei der Anforderung von Kurzdatensätzen nach den beschriebenen Verfahren ohne übermäßig großen Aufwand gesichert ist, bedarf es bei der Anforderung von Langdatensätzen umfangreicher und – leider – auch sehr aufwendiger Anonymisierungsmaßnahmen: Direkte Identifikatoren und nahezu alle Regionalangaben müssen aus dem ursprünglichen Datensatz herausgenommen werden, bei sensiblen Merkmalen oder solchen, über die einfach Zusatzinformationen gewonnen werden können, müssen umfangreiche Umschlüsselungsmaßnahmen vorgenommen werden, die Datensätze sind in aller Regel umzusortieren, da sie zumeist nicht systemfrei angeordnet sind. Es ist eine Stichprobe aus dem ursprünglichen Material zu ziehen.

Ein sehr aufwendiger Teil der Anonymisierungsmaßnahmen liegt in der Abwicklung von realitätsnahen Szenarien, die bei den Kominationen sensibler und offenkundiger Merkmale ansetzen und zur Elimination einmalig auftretender Datensätze oder zur geregelten bzw. zufälligen Verschiebung von Ausprägungen einzelner Merkmale in Datensätzen führen können. Allerdings sehen wir derzeit keine Möglichkeit, auf diese Reidentifikationsexperimente, die wir allerdings auf Szenarien von der Größenordnung des Adreßszenarios beschränken, im Rahmen der Anonymisierungsmaßnahmen zu verzichten. Nach dem Willen des Gesetzgebers muß eine Deanonymisierung der weiterzuleitenden anonymisierten Einzelangaben nach den Kenntnissen, die in den Statistischen Ämtern über Zusatzwissen vorliegen, subjektiv zweifelsfrei, d. h. entsprechend dem von uns vermuteten Zusatzwissen beim Emfänger ausgeschlossen sein.

ad 3.: Vertragliche Bindung des Empfängers anonymisierter Einzelangaben

Wenngleich von uns diese Reidentifikationsexperimente durchgeführt werden, bei denen ein bestimmtes, allerdings rudimentäres Zusatzwissen beim Empfänger unterstellt wird, können wir nicht mit Sicherheit davon ausgehen, daß unsere subjektiv, nach bestem Wissen gesetzten Annahmen über zusätzlich verfügbares Wissen und informationstechnologisch gegebene bzw. sich noch entwickelnde Möglichkeiten beim Empfänger zutreffen, denn wir können dieses Zusatzwissen nicht mit letzter Sicherheit abschätzen. Deshalb wird mit den Empfängern ein Vertrag über die Weitergabe anonymisierter Einzelangaben geschlossen, in dem sich der Empfänger anonymisierter Einzelangaben zu bestimmten Verwendungsbeschränkungen und Datenschutzsicherungen verpflichtet. Hierzu zählen:

- Der Datenempfänger hat jede Handlung zu unterlassen, die darauf abzielt oder geeignet ist, die anonymisierten statistischen Einzelangaben zu deanonymisieren;

- der Datenempfänger darf die anonymisierten Einzelangaben keinem Dritten zugänglich machen;

- der Datenempfänger muß die anonymisierten Einzelangaben nach Ablauf einer vertraglich festgelegten Frist löschen;

- bei Verstoß gegen eine dieser vertraglichen Regelungen hat er die anonymisierten Einzelangaben zu löschen und eine vertraglich festgelegte Konventionalstrafe zu zahlen. Außerdem kann ihn das StBA vom Bezug weiterer anonymisierter Einzelangaben ausschließen.

ad 4.: Abstimmungsverfahren im StBA vor Übermittlung der anonymisierten Einzelangaben

Für Anforderungen nach anonymisierten Einzelangaben, die an das StBA gerichtet werden, ist regelmäßig die für die jeweilige Statistik zuständige Fachabteilung federführend. Diese hat die Abteilung Z (Verwaltung: wegen der Kostenkalkulation und der Vertragsgestaltung), die Abteilung I (Allgemeine und zusammenfassende Aufgaben der Statistik: wegen methodischer und juristischer Fragen der Geheimhaltung) sowie die Abteilung II (Mathematik, maschinelle Datenverarbeitung wegen Verfahrenstests, maschinelle Aufbereitung, Lieferung auf Magnetband) einzuschalten.

Für die Freigabe der anonymisierten Einzelangaben ist der jeweils zuständige Abteilungsleiter verantwortlich und federführend. Der Freigabe muß jedoch von den Leitern der genannten Querschnittsabteilungen Z, I und II zugestimmt werden. Bestehen seitens einer der Abteilungen Bedenken, die nicht ausgeräumt werden können, ist die Amtsleitung einzuschalten.

2.3 Probleme der verwaltungspraktischen Umsetzung des § 11 Abs. 5 BStatG

Nach unserer Einschätzung sind die Bedingungen einer Veröffentlichung anonymisierter Einzelangaben in Tabellen sowie die Übermittlung anonymisierter Einzelangaben in Kurzbandsätzen auch unter den gegenwärtigen rechtlichen Rahmenregelungen des § 11 Abs. 5 BStatG zur Zufriedenheit der Konsumenten bei gleichzeitig ausreichendem Schutz für die Auskunftsgebenden bzw. Betroffenen einigermaßen gestaltbar. So sind zumindest unsere praktischen Erfahrungen in den letzten Jahren gewesen. Anders verhält es sich dagegen bei der Übermittlung von anonymisierten Einzelangaben in Langdatensätzen. Den Anforderungen der Wissenschaft konnte mit den bisherigen Regelungen, die gleichwohl eine Erleichterung im Zugang zu anonymisierten Einzelangaben gegenüber dem Stand in den siebziger Jahren darstellen, nicht in vollem Umfang nachgekommen werden. Insbesondere hat das StBA bisher im Rahmen der bestehenden rechtlichen Regelungen keine Möglichkeit gesehen, den Wissenschaftlern umfangreiche und ohne weitere vertragliche Klauseln zugängliche Public Use Samples kostengünstig zur Verfügung zu stellen.

Tatsächlich läßt sich nach allen bisherigen Untersuchungen nur eine begrenzte Anzahl von Merkmalen/Merkmalsausprägungen für Personen und/oder Haushalte so anonymisieren, daß eine Reidentifizierung nach dem Kenntnisstand in den Statistischen Ämtern mit Sicherheit ausgeschlossen ist. Das mögliche Zusatzwissen von Empfängern läßt sich vom

Statistiker nicht ausreichend abschätzen und damit auch nicht begrenzen. Die großen Unsicherheitsmomente bei der Abschätzbarkeit des Zusatzwissens beim jeweiligen Empfänger und die Unzulässigkeit eines Restrisikos in der Einschätzung durch die Statistischen Ämter haben dazu geführt, daß von der in § 11 Abs. 5 BStatG vorgesehenen Möglichkeit in der Vergangenheit nur in begrenztem Umfang und mit einem unverhältnismäßig aufwendigen Verfahren Gebrauch gemacht werden konnte. Um der Wissenschaft dennoch zu helfen, mußte jede Anforderung für sich bearbeitet und versucht werden, wenigstens einen Teil der Informationen für den Forscher zu erhalten. Ausgehend davon, daß selbst dann ein absoluter Schutz vor Deanonymisierungsmöglichkeiten fraglich ist, ist ein Vertrag entwickelt worden, in dem sich der Empfänger anonymisierter Einzelangaben zu bestimmten Verwendungsbeschränkungen und Datenschutzsicherungen verpflichtet.

3 Zur möglichen Weiterentwicklung der Anonymisierung

Unser Eindruck ist, daß die Wissenschaftler diese Verpflichtungen bereitwillig auf sich nehmen, wenn sie nur Zugang zu den von ihnen benötigten Einzelangaben erhalten.

Allerdings stellen die gegenwärtig praktizierten Regelungen weder die Wissenschaftler noch das StBA so recht zufrieden. Eine Änderung dieser Regelung ist jedoch nur im Rahmen einer Weiterentwicklung der gesetzlichen Grundlagen der Anonymisierung möglich.

3.1 Elemente einer Weiterentwicklung der rechtlichen Grundlagen der Anonymisierung

Um den Zugang der Wissenschaft zu anonymisierten Einzelangaben aus Bundesstatistiken weiter zu erleichtern, müßte der Gesetzgeber die Akzeptanz eines Restrisikos der Reidentifikation im Gesetz festschreiben. Dies könnte durch den Übergang von der absoluten Anonymität (nach subjektiver Einschätzung durch die Statistischen Ämter) auf die faktische Anonymität (unter Berücksichtigung eines Restrisikos) geschehen. Ein tragfähiges, gesetzlich zu verankerndes Kriterium könnte die Unverhältnismäßigkeit des Aufwandes sein, der für eine Deanonymisierung erbracht werden muß. In dem derzeit vorliegenden Entwurf eines BStatG ist diese Regelung für die Wissenschaft enthalten.

Die gegenwärtig notwendige vertragliche Vereinbarung zwischen StBA und dem Empfänger anonymisierter Einzelangaben kann auf Dauer nicht befriedigen. Die zur Zeit auf der Basis des bestehenden § 11 Abs. 5 BStatG vertraglich geregelten Verwendungsbeschränkungen und Datensicherungen sollten deshalb bei Akzeptanz der faktischen Anonymität in das BStatG aufgenommen werden. Dazu gehören insbesondere das strafbewehrte Reidentifikationsverbot, die Aufzeichnungspflicht, das Weitergabeverbot und die Beschränkung auf bestimmte Auswertungszwecke. In dem gegenwärtig vorliegenden Entwurf des BStatG sind auch diese Elemente verankert.

3.2 Ansätze der wissenschaftlichen Begleitforschung

Im Forschungsprojekt der GMD sind anhand verschiedener Szenarien (unterschiedlichen Beständen an Zusatzwissen beim potentiellen Empfänger anonymisierter Einzelangaben) technische Größenordnungen für die jeweiligen Reidentifikationsrisiken vermittelt worden.

Nach den vorgenannten Überlegungen zur Weiterentwicklung der rechtlichen Grundlagen und den in verschiedenen anderen Beiträgen in diesem Band angestellten Überlegungen wären die technischen Restrisiken allerdings einer Bewertung zu unterziehen. Die Ansatzpunkte dafür müßten noch weiter wissenschaftlich untersucht werden. Ausgehend von den Kostenüberlegungen einer Deanonymisierung wäre insbesondere der Opportunitätskostenansatz zur Quantifizierung des Risikos einer Deanonymisierung auf seine Tragfähigkeit hin zu überprüfen. Bei diesen Überlegungen ist auch der mögliche Verlust der übermittelnden Stelle bei einem erfolgreichen Deanonymisierungsversuch mit einzubeziehen. Dies soll im Rahmen eines Forschungsvorhabens im StBA geschehen. Externe Wissenschaftler sind zur Mitarbeit herzlich eingeladen.

Für die weitere verwaltungsmäßige Ausgestaltung des § 16 Abs. 4 des Entwurfs des BStatG sind nicht zuletzt Organisations- und prozedurale Fragen von ganz erheblicher Bedeutung. Es sollte deshalb auch geprüft werden, ob nicht ein Beratungsgremium für die Weiterleitung von Mikrodatenfiles für wissenschaftliche Zwecke innerhalb der Statistischen Ämter eingerichtet werden könnte, an dem ggf. auch Vertreter der Wissenschaft und der Auskunftgebenden (evtl. die Datenschutzbeauftragten) beteiligt werden könnten, sobald das neue BStatG verabschiedet sein wird.

Manfred Euler
Statistisches Bundesamt, Wiesbaden

Lieferung anonymisierter Einzelangaben aus der Einkommens- und Verbrauchsstichprobe (EVS)

Meine Aufgabe ist es, die von Herrn Südfeld aufgezeigten Rahmenbedingungen und Verfahren für die Anonymisierung von Einzelangaben im Statistischen Bundesamt (StBA) am Beispiel einer statistischen Erhebung, nämlich der EVS, zu verdeutlichen. Das bedeutet zwangsläufig, daß ich Ihnen zunächst über rechtliche Grundlagen, Erhebungsziele und Erhebungsverfahren der EVS berichten muß, obwohl viele von Ihnen aus Ihrer Tätigkeit damit bereits vertraut sein mögen.

Die EVS ist – neben den sogenannten laufenden Wirtschaftsrechnungen bei drei ausgewählten Haushaltstypen – die einzige amtliche Erhebung in der Bundesrepublik Deutschland, die einen wirklich umfassenden Einblick in die ökonomische und soziale Lage der privaten Haushalte ermöglicht. Ausgehend von der kompletten Erfassung der Einnahmen und Ausgaben über die Ausstattung mit langlebigen Gebrauchsgütern, die Wohnsituation und die Vermögensbestände findet sich hier eine Fülle von Informationen, die für den Politiker, den Beamten in der Verwaltung, den Unternehmer, den Wissenschaftler, für Vertreter von Arbeitgeber- und Arbeitnehmerorganisationen, von Verbraucherverbänden und anderen Vereinigungen von großem Wert sein können. Es ist deshalb nur zu verständlich, wenn von allen genannten Gruppen versucht wird, diese Informationsquelle möglichst detailliert, möglichst eigenzweckbezogen und möglichst umfassend zu nutzen, sei es durch die Verwendung des vorliegenden umfangreichen Quellenmaterials,[1]) durch Sonderauswertungen oder durch die Analyse von Einzeldaten für wissenschaftliche Zwecke.

Die rechtliche Grundlage für die EVS bildet das Gesetz über die Statistik der Wirtschaftsrechnungen privater Haushalte aus dem Jahr 1961,[2]) einem Zeitraum also, als die technischen Voraussetzungen für eine Deanonymisierung kaum vorhanden waren, die datenschutzrechtlichen Probleme noch nicht in heutiger Schärfe und Klarheit gesehen worden sind und niemand auf die Idee kam, durch Aufnahme von Weiterleitungsklauseln die Lieferung von Einzeldaten, insbesondere an den wissenschaftlichen Bereich, im Gesetz selbst zu regeln. Trotz späterer Gesetzesänderungen, die ausschließlich den Erhebungszeitraum betrafen, sind die wichtigsten gesetzlichen Vorschriften nach wie vor gültig. Sie bestimmen über den bereits geschilderten Inhalt der Erhebung hinaus

[1]) Vor allem in der Fachserie 15, Reihe „Einkommens- und Verbrauchsstichproben" (Allein der Umfang des Tabellenteils ist von rund 360 Seiten für die Stichprobe 1962/63 auf rund 1800 Seiten für die Stichprobe 1978 systematisch ausgebaut worden) und in der Zeitschrift „Wirtschaft und Statistik" (z. B. für die Erhebung 1978 21 Beiträge).

2) BGBl. I S.18, ergänzt durch Gesetz zur Änderung des Gesetzes über die Statistik der Wirtschaftsrechnungen privater Haushalte vom 19. Januar 1968 (BGBl. I S.97) und geändert durch 1. Statistikbereinigungsgesetz vom 14. März 1980, Art. 10 (BGBl. I S.294).

- den Erhebungskreis (alle privaten Haushalte), die Erhebungsperiode (ein Jahr) und den Erhebungsabstand (5 Jahre im Regelfall) (§ 1, Nr. 2),

- den Erhebungsumfang (maximal 0,3 % aller Haushalte) (§ 3, Abs. 2),

- die Freiwilligkeit der Auskunftserteilung (§ 4) und

- den Arbeitsschnitt (§ 5); danach obliegt die Erhebung den Statistischen Landesämtern, die Aufbereitung und Auswertung dem StBA.

Von besonderer Bedeutung erscheint die Tatsache, daß der Gesetzgeber bewußt auf eine Auskunftspflicht der Befragten verzichtet hat, wobei dahingestellt bleiben kann, ob diese Entscheidung eine seherische Vorausschau des Rechtes auf informationelle Selbstbestimmung bedeutet oder auf der Einsicht beruht, daß selbst 1960 eine Auskunftspflicht in einem so sensiblen Bereich politisch kaum durchsetzbar wäre. Für die amtliche Statistik sind Erhebungen[3]) auf freiwilliger Basis die Ausnahme geblieben; die Wirtschaftsrechnungen privater Haushalte sind zudem der einzige mir bekannte Fall in der amtlichen Statistik der Bundesrepublik Deutschland, wo die Beteiligten als Anreiz für die Teilnahme und als Entschädigung für die langfristigen Anschreibungen eine – wenn auch geringe – Prämie erhalten.

Wegen der aufgrund der Ergebnisse von Piloterhebungen zu erwartenden geringen Teilnahmebereitschaft der Haushalte wurde auf die Ziehung einer Zufallsstichprobe verzichtet. Die Ergebnisse der Stichprobe werden mit Hilfe des jeweiligen Mikrozensus auf die Grundgesamtheit der repräsentierten privaten Haushalte hochgerechnet.

Die Schilderung der Erhebungsmerkmale dürfte keine Zweifel daran gelassen haben, daß es sich bei einem großen Teil dieser Merkmale um sensible Daten handelt. Trotzdem bedarf es der Differenzierung. Die Angaben über Einnahmen aus Erbschaften, Lotto- und Spielgewinnen sind mit Sicherheit wegen der z. T. beträchtlichen Höhe der Einnahmen sensibler als die Angaben über das Kindergeld, dessen Bezug sich bis zu einem bestimmten Alter der Kinder und innerhalb bestimmter Einkommensgrenzen automatisch aus der Zusammensetzung des Haushalts ergibt. Auch das Einkommen eines Beamten ist bei Kenntnis der Dienststellung und seines Alters unschwer zu ermitteln. Ebenso ist die Gefahr der Deanonymisierung z. B. bei Ausgaben für Hauskauf oder -bau wesentlich höher als bei Ausgaben für Grundnahrungsmittel. Generell läßt sich sagen, daß die Sensibilität der Daten wesentlich von der Häufigkeit, mit der sie in der Stichprobe vertreten sind, und von der absoluten Höhe der Einnahmen-, Ausgaben- oder Bestandsgrößen abhängt.

Bereits bei der ersten EVS im Jahr 1962/63 hat sich herausgestellt, daß sich Haushalte mit besonders hohen Einkommen nicht oder in völlig unzureichendem Maß an der Erhebung beteiligt hatten, so daß für sie keine aussagefähigen Ergebnisse ermittelt werden konnten. Bei allen folgenden Erhebungen, also 1969, 1973, 1978 und 1983, wurden

[3]) Einzelne Fragen ohne Auskunftszwang finden sich dagegen in mehreren Erhebungen, z. B. der Volkszählung 1987 (Telefon) und dem Mikrozensus 1985 (Urlaubs- und Erholungsreisen, Erkrankungen).

deshalb Haushalte, die eine bestimmte Einkommenshöhe überschritten, nicht in die Aufbereitung einbezogen; 1983 lag diese „Abschneidegrenze" bei einem monatlichen Haushaltsnettoeinkommen von 25000 DM. Die Nichterfassung dieser Bevölkerungsgruppe erscheint für die Beurteilung des Deanonymisierungsrisikos von erheblicher Bedeutung, weil Extremfälle, die in besonderem Maße Deanonymisierungsrisiken ausgesetzt sind, von vornherein ausgeschaltet sind. Nur am Rande sei vermerkt, daß wegen erhebungstechnischer Probleme auch Haushalte von Ausländern und die Anstaltsbevölkerung an der EVS nicht beteiligt sind.

Einzelangaben der EVS 1962/63 und 1969 wurden wissenschaftlichen Instituten ohne Namen und Anschrift zur Verfügung gestellt; nach dem damaligen juristischen und technischen Kenntnisstand reichte dies aus, um eine Deanonymisierung weitgehend auszuschließen. Einzelangaben der Erhebungen 1973 und 1978 wurden nach Durchführung der im Detail noch zu beschreibenden, im Grundsätzlichen von Herrn Südfeld bereits erläuterten Anonymisierungsverfahren geliefert.

Die Einzelangaben der EVS sind in vier Materialteilen enthalten, die sich aus den Aufbereitungszyklen zwangsläufig ergeben, nämlich Grundinterview, Schlußinterview, Jahresrechnungen (Jahreseinnahmen und -ausgaben) und Nahrungs- und Genußmittel. Die vier Materialteile lassen sich für jeden Haushalt, der an der ganzen Erhebung teilgenommen hat, über die Registriernummer des Haushalts zusammenführen. Diese siebenstellige Registriernummer, die von den Statistischen Landesämtern vergeben wird, besteht aus zwei Kennziffern für das Land, drei Kennziffern für den Interviewerbezirk und zwei Kennziffern für den Haushalt. Name und Anschrift des Haushalts sind in den Datenträgern nicht enthalten; sie sind nur den Statistischen Landesämtern bekannt und werden nach Abschluß der Arbeiten gelöscht.

Tritt ein Konsument mit der Bitte um Lieferung von Einzelangaben an das StBA heran, so wird zunächst versucht, in einem Vorgespräch zu klären, ob sich seine Wünsche nicht auf andere Weise, z. B. durch Rückgriff auf unveröffentlichtes Tabellenmaterial oder durch Sonderauswertungen, erfüllen lassen. Ist dies nicht der Fall, wird geprüft, ob die Lieferung aggregierter Daten, z. B. die Zusammenfassung der Angaben von drei Haushalten zu einem künstlichen Haushalt, möglich ist, weil hier keine datenschutzrechtlichen Bedenken zu befürchten sind. Erst nach gründlicher Abwägung aller dieser Möglichkeiten werden die Voraussetzungen für die Lieferung anonymisierter Einzelangaben mit dem Konsumenten erörtert. Wissenschaftliche Institute sind allerdings sehr häufig auf die Lieferung von Einzelangaben angewiesen, weil die Anwendung vieler wirtschafts- und sozialwissenschaftlicher Modelle und Verfahren (insbesondere Mikrosimulationsmodelle) ohne Einzelangaben nicht möglich ist und weil wissenschaftliche Forschung mit Hilfe standardisierter Programme kaum betrieben werden kann.

Die Nachfrage nach Einzelangaben ist also im wesentlichen auf den wissenschaftlichen Bereich beschränkt, zumal der Umgang mit einem so umfangreichen und komplizierten Datenmaterial wie dem der EVS ein erhebliches Maß an Fachwissen und leistungsfähige Großrechenanlagen voraussetzt. Für kleinere Institute oder gar einzelne Unternehmen oder Privatpersonen ist in aller Regel die Verwendung von Einzelangaben aus der Stichprobe

weder finanziell noch technisch sinnvoll und möglich, so daß ein überschaubarer Kreis von potentiellen Nachfragern verbleibt.

Herr Südfeld hat bereits darauf hingewiesen, daß Anforderungen von Lieferung von Einzelangaben am konkreten Einzelfall auf ihre Realisierbarkeit hin überprüft werden müssen.

Einige Maßnahmen sind allerdings – unabhängig von den jeweiligen Konsumentenwünschen – in jedem Fall zu treffen. Dazu gehört:

1. Der Ersatz der zuvor erwähnten Registriernummer im Urmaterial durch eine neutrale laufende Nummer. Damit wird gleichzeitig der einzige in der EVS vorhandene direkte[4]) regionale Bezug, nämlich die Signierziffer des Bundeslandes, gelöscht. Selbst wenn der Konsument über Anschriftenlisten mit dem Namen des Haushalts und der Registriernummer verfügen würde, könnte er mit der laufenden Nummer keinen Bezug mit Name und Anschrift des Haushalts herstellen.

2. Die Ziehung einer Unterstichprobe. Damit wird verhindert, daß ein mögliches Zusatzwissen beim Konsumenten über die Beteiligung eines bestimmten Haushalts an der jeweiligen Stichprobe die Deanonymisierung erleichtert. Dieses Zusatzwissen hat bei den Reidentifikationsexperimenten der Gesellschaft für Mathematik und Datenverarbeitung für die Beurteilung des Restrisikos eine wesentliche Rolle gespielt. Selbst wenn ein solches Zusatzwissen grundsätzlich vorhanden ist, ist durch die Ziehung der Unterstichprobe kein Zusatzwissen darüber möglich, ob der betreffende Haushalt auch in der Unterstichprobe enthalten ist.

3. Eine Auswahl und Begrenzung der gewünschten Erhebungsmerkmale. Bei der Anforderung von Langdatensätzen ist eine vollständige Lieferung aller erfaßten Merkmale eines oder mehrerer Erhebungsteile ausgeschlossen, weil das Deanonymisierungsrisiko mit der Zahl der übermittelten Einzelmerkmale wächst.

Die Fragen, die sich im konkreten Einzelfall stellen, werden mit den Konsumenten gründlich vorbesprochen, um zu verhindern, daß die schriftliche Datenanforderung von vornherein unrealisierbare Wünsche oder Problemfälle enthält, die sich durch Klumpung von Erhebungsmerkmalen, gröbere systematische Gliederung u. ä. ausschalten ließen. Dazu ist es notwendig zu wissen, welche Ziele der Konsument mit den gewünschten Daten verfolgt.

In einem der wenigen Fälle, in denen bisher die Lieferung von Einzelangaben aus den Stichproben 1978 und 1973 beantragt und durchgeführt wurde, lag der Schwerpunkt des Interesses des Konsumenten bei der Darstellung des Haushaltseinkommens. Darüber hinaus wurden lediglich die Hauptausgabengruppen für den privaten Verbrauch nach Verwendungszweck und Dauerhaftigkeit (13 bzw. 5 Variable), die Gesamtsumme der Ersparnis, 4 Variable zur Wohnsituation und 4 Variable zur Vermögenslage angefordert, zur Beschreibung des Haushalts 7 Variable über Zahl und Art der Haushaltsmitglieder und 2 allgemeine Variable.

[4]) An indirekten regionalen Angaben ist nur die Gemeindegrößenklasse der Wohngemeinde vorhanden.

Zusätzlich zu den zuvor beschriebenen allgemeinen Deanonymisierungsmaßnahmen wurden folgende Einzelmaßnahmen getroffen:

- Einkommensangaben. Die Verhandlungen zwischen dem Konsumenten und dem StBA zogen sich von Anfang 1983 bis Mitte 1985 hin. Die Lieferung der Einzelangaben erfolgte Ende 1985. Unter dieser Voraussetzung konnte davon ausgegangen werden, daß der Konsument nach Ablauf von sieben Jahren nach Abschluß der Erhebung über kein Zusatzwissen über die Einkommensverhältnisse des Haushalts im Jahr 1978 verfügte und auch nicht mehr beschaffen konnte.[5]) Trotzdem wurden mit Ausnahme der Renten der gesetzlichen Rentenversicherung alle Angaben über die Art der Rente (eigene Rente, Witwen-, Waisenrente) sowie über die Person des Rentenbeziehers gelöscht. Mehrere Renten und Übertragungen mit eigenen Code-Bezeichnungen wurden in einer Position zusammengefaßt. Die Summe des Jahreshaushaltsbrutto- und -nettoeinkommens wurden gerundet.

- Bei den Variablen zur Wohnsituation wurde die Wohnfläche gerundet, Wohnungen mit 5 Räumen und mehr in einer Gruppe nachgewiesen. Als weitere Wohnungsmerkmale wurde lediglich die Ausstattung mit oder ohne Bad/Dusche sowie mit oder ohne Sammelheizung nachgewiesen.

- Für den Nachweis von Vermögenswerten wurde der Haus- und Grundbesitz nach 2 Einheitswertklassen, die erfaßten Geldvermögenswerte ohne Unterscheidung der Vermögensarten nach 11 Größenklassen und die Lebensversicherungssummen nach 3 Größenklassen nachgewiesen.

- Haushalte mit 7 Personen und mehr wurden generell nicht in die Unterstichprobe einbezogen. Haushalte dieser Größe machen nur knapp 1 % aller Haushalte aus und sind deshalb in besonders hohem Maß Deanonymisierungsrisiken ausgesetzt. Das Alter der Haushaltsmitglieder der übrigen Haushalte, das im Urmaterial durch das Geburtsjahr gekennzeichnet war, wurde in 17 Altersgruppen zusammengefaßt, 75jährige und ältere Personen in einer Gruppe nachgewiesen.

- Kombinationen sensibler Daten wurden in Form von Streuungsübersichten nachgewiesen und Einerfälle durch Zusammenfassung von Merkmalen ausgeschaltet.

- Mit dem Konsumenten wurde ein öffentlich-rechtlicher Vertrag abgeschlossen, dessen § 4 folgenden Wortlaut hat:

  „(1) Der Datenempfänger hat jede Handlung zu unterlassen, die darauf abzielt oder geeignet ist, die in der Vertragsdatenbasis enthaltenen anonymisierten statistischen Einzelangaben zu deanonymisieren.

  (2) Werden in der Vertragsdatenbasis enthaltene anonymisierte Einzelangaben deanonymisiert, auch wenn dies nicht durch eine darauf abzielende Handlung geschieht, so hat der Datenträger diese statistischen Einzelangaben gegenüber jeder anderen Person geheimzuhalten . . ."

5) Auch das Bundesverfassungsgericht geht davon aus, daß im Bereich der öffentlichen Forschung in der Regel kaum Zusatzwissen vorhanden ist (Urteil des Bundesverfassungsgerichts, 1 BVR 209/83 u. a. vom 15. 12. 1983, Abschnitt IV, 5).

Ferner hat der zuständige Landesdatenschutzbeauftragte schriftlich bestätigt, daß das Datensicherungskonzept des Konsumenten den gestellten Anforderungen entspricht.

Zusätzlich zu diesen gezielten Einzelmaßnahmen ist auf einige Punkte aufmerksam zu machen, die dem Erhebungsverfahren der EVS systemimmanent sind und – für sich genommen – ganz erheblich deanonymisierungshemmend wirken.

Hier ist an erster Stelle die bereits beschriebene Nichterfassung bestimmter Bevölkerungsgruppen zu erwähnen, insbesondere der – vereinfachend formuliert – „reichen" Haushalte. Zweitens enthalten die Auswertungsbänder Jahreswerte der Einnahmen und Ausgaben, die in der Regel durch Summierung der monatlichen Anschreibungen der Haushalte gebildet werden. Bei bestimmten Ausgaben, insbesondere bei Nahrungs- und Genußmitteln, erfolgen die Anschreibungen nur in einem Monat und werden durch Multiplikation des Monatsergebnisses mit 12 auf eine fiktive Jahressumme hochgerechnet; folglich ist auch für den einzelnen Haushalt nur eine fiktive Größe der Jahresgesamtausgaben zu ermitteln. Selbst bei den Einnahmen- und Ausgabenpositionen, die während des ganzen Erhebungsjahres monatlich zu verbuchen sind, ist nicht festzustellen, wie häufig im Laufe des Jahres Zahlungen erfolgt sind. So kann z. B. eine Jahresausgabe für Lebensversicherungen ebenso aus 12 Monatszahlungen in Höhe von 100 DM resultieren wie aus einer einmaligen Zahlung von 1200 DM; selbst für die beteiligten Haushalte dürfte in vielen Fällen die Jahressumme eine Größe sein, die sich nicht im Bewußtsein der Befragten niederschlägt, so daß sie vermutlich selbst Schwierigkeiten hätten, ihren Haushalt anhand der Jahressummen zu identifizieren.

Ferner unterliegen die Anschreibungen der Haushalte umfangreichen Plausibilitätskontrollen, durch die u. a. offensichtlich falsche oder unvollständige Angaben korrigiert werden; so müssen z. B. in vielen Fällen die Zinsen auf Spar- und Bausparguthaben anhand der Angaben der Haushalte über die vorhandenen Vermögensbestände geschätzt werden, weil es sich dabei um (zunächst) unbare Einkommen handelt, deren Verbuchung häufig vergessen wird. Außerdem werden den Haushalten aus systematischen Gründen fiktive Einnahmen und Ausgaben zugerechnet, so z. B. der Mietwert der Eigentümerwohnung oder die mit Einzelhandelspreisen bewerteten Entnahmen aus dem eigenen Betrieb oder Gegenwerte für Sachleistungen des Arbeitgebers, etwa in Form von Kost und Logis oder von Deputaten.

Schließlich werden die Einkommen aus Unternehmertätigkeit, also die Einkommen aus dem eigenen landwirtschaftlichen oder gewerblichen Unternehmen bzw. aus freiberuflicher Tätigkeit, nur als Restgröße errechnet, nämlich als Saldo zwischen verbuchten privaten Ausgaben (einschl. Ersparnis) und den Einkommen des Haushalts aus anderen Quellen als aus Unternehmertätigkeit. Sie sind weder mit dem steuerlichen noch mit dem betriebswirtschaftlichen Einkommen aus Unternehmertätigkeit identisch und selbst den Haushalten nicht bekannt oder (ohne Besitz der Anschreibungsbücher) von ihnen nachträglich zu bestimmen.

Alle diese Dinge sind mit Sicherheit nicht geeignet, Deanonymisierungsversuche zu erleichtern, denn daß – um den Szenario-Stil zu gebrauchen – der „Angreifer" über entsprechendes Zusatzwissen verfügt, ist so gut wie ausgeschlossen.

Was die Angaben über die Zusammensetzung des Haushalts anbelangt, so wird für die Darstellung der Jahreseinnahmen und -ausgaben von dem Zustand ausgegangen, der im größeren Teil des Erhebungsjahres gegeben war. Wird z. B. ein Kind im Mai geboren, wird der Haushalt als Ehepaar mit einem Kind beschrieben, erfolgt die Geburt im September, als Ehepaar ohne Kind. Auch dies erschwert die Deanonymisierung eines Haushalts erheblich.

Meine Damen und Herren, ich habe Ihnen an einem konkreten Beispiel aufgezeigt, unter welchen Voraussetzungen in einem konkreten Fall Einzelangaben der EVS an einen Konsumenten für wissenschaftliche Forschungszwecke herausgegeben worden sind. Die Dauer des Entscheidungsprozesses mag ein Beweis dafür sein, daß alle Beteiligten sich ihre Aufgabe nicht leichtgemacht haben und davon ausgegangen sind, daß im Vordergrund aller Überlegungen der Schutz des Auskunftsgebenden vor einer Offenlegung oder gar vor einem Mißbrauch seiner Daten stehen muß, gleichgültig, ob er zur Erteilung der Auskünfte verpflichtet war oder nicht. Wir waren der Auffassung, daß dieser Schutz in unserem Fall gewährleistet war. Die Arbeiten der Gesellschaft für Mathematik und Datenverarbeitung (GMD) haben gezeigt, daß die Gefahr der Deanonymisierung bei einem hohen Grad von Fach- und Zusatzwissen beim „Angreifer", das die Möglichkeit der Nutzung von umfangreichen Registern und Datenbänken einschließt, sicherlich nicht unterschätzt werden darf. Schon die Kennzeichnung einiger von der GMD benutzten Szenarien (Staatsanwalt, Steuerfahndung, Kripo) weist aber darauf hin, daß die Zielsetzung dieser Angriffe eine deutlich andere ist als diejenige wissenschaftlicher Institute. Auch die GMD kommt zu dem Ergebnis, daß in diesem Bereich mit einer wesentlich geringeren Gefährdung des Datenschutzes zu rechnen ist.[6]) Darüber hinaus ist wohl auszuschließen, daß ein Institut die mit der Anwendung eines Szenarios erforderlichen Kosten für die Datenverarbeitung aufzubringen vermag.

Der Kosten-Nutzeneffekt ist zudem unter einem anderen Aspekt von Bedeutung, auch wenn dieser Aspekt für die datenschutzrechtliche Beurteilung der Weitergabe von Einzelangaben irrelevant ist. Die Sammlung der Einzelangaben von Haushalten im Rahmen der EVS ist – bei kalkulierten Bundes- und Länderkosten von 28 Millionen DM für die Erhebung 1978 – eine äußerst kostspielige Angelegenheit. Nur ein Teil dieser Angaben kann im Rahmen der amtlichen Auswertungen genutzt werden. Für die Beurteilung vieler Probleme unserer Zeit ist die Berechnung von Durchschnittswerten in Tabellenform, die auch heute noch den breitesten Raum in der Darstellung der statistischen Ämter „für allgemeine Zwecke"[7]) einnimmt, ein nur bedingt geeignetes Mittel. Hier lassen sich jedoch mit wissenschaftlichen Methoden – und das heißt faktisch mittels Analysen aufgrund von Einzelsätzen – neue Erkenntnisse gewinnen, die die Fehlleitung von vielen Millionen DM der öffentlichen Hand und der Unternehmen verhindern und den Einsatz entsprechender Mittel für wirtschafts- und sozialpolitisch sinnvolle Maßnahmen steuern können; ich möchte in diesem Zusammenhang nur an die Notwendigkeit der Quantifizierung und Bekämpfung der „neuen Armut" erinnern.

---

[6]) Siehe Paaß, G., Wauschkuhn, U., „Datenzugang, Datenschutz und Anonymisierung – Analysepotential und Identifizierbarkeit von anonymisierten Individualdaten", Bericht Nr. 148 der Gesellschaft für Mathematik und Datenverarbeitung, München/Wien 1985, S. 86.

[7]) Die Veröffentlichung und Darstellung von Ergebnissen „für allgemeine Zwecke" ist nach § 3, Absatz 1, Satz 1 des Gesetzes über die Statistik für Bundeszwecke vom 14. 3. 1980 (BGBl. I S. 289) eine der Aufgaben des Statistischen Bundesamtes.

Ich hoffe, Ihnen einen Einblick in die Fülle von Problemen und Schwierigkeiten vermittelt zu haben, die in der gegenwärtigen rechtlichen Situation mit der Lieferung von Einzelangaben verbunden sind und die beide, Wissenschaftler und Statistiker, belasten: Die Wissenschaftler, weil sie die gewünschten Angaben entweder überhaupt nicht oder nicht in dem gewünschten Umfang und für den oder in dem gewünschten Zeitraum bekommen; die Statistiker, weil letzte Zweifel an der Richtigkeit der getroffenen Entscheidungen wohl nie völlig auszuschließen sind. Im Absatz 4 des § 16 des Entwurfs eines Gesetzes über die Statistik für Bundeszwecke[8]) ist vorgesehen, daß die statistischen Ämter Einzelangaben für wissenschaftliche Zwecke an Amtsträger oder für den öffentlichen Dienst besonders Verpflichtete in Hochschulen oder sonstigen Einrichtungen mit der Aufgabe unabhängiger wissenschaftlicher Forschung übermitteln dürfen, soweit diese Angaben nur mit einem unverhältnismäßig hohen Aufwand dem Auskunftgebenden oder Betroffenen zugeordnet werden können. Wenn dieser Abs. 4 ohne wesentliche Veränderungen die Gesetzesprozedur passiert, werden viele, wenn auch möglicherweise nicht alle der heutigen Probleme bei der Lieferung von Einzelangaben für wissenschaftliche Zwecke gelöst sein.

[8]) Bundesratsdrucksache 19/86 vom 10. 1. 1986.

Podiumsdiskussion

Moderator: Prof. Dr. Allerbeck
(Universität Frankfurt)

Wir wollen uns in dieser Podiumsdiskussion vor allem den Fragen künftiger Regelungen und zur Praxis der Weitergabe anonymisierter Einzelangaben aus der amtlichen Statistik zuwenden.

Für den Ablauf der Diskussion ist es sinnvoll, zwei Runden zu strukturieren; eine erste Runde beträfe das Fazit, welche Situation in Zukunft nach der Verabschiedung des neuen Statistikgesetzes (vorausgesetzt, daß es verabschiedet wird) anzustreben ist, hinsichtlich der Verfügbarkeit von anonymisierten Mikrodaten. In einer zweiten Runde sollten wir darüber diskutieren, welche Verfahren und Schritte angemessen sind, um zu dem angestrebten Ergebnis zu kommen.

Ich möchte um einen ersten, ganz kurzen Beitrag von Herrn Hohmann als Vertreter des Rechts bitten. Ich habe Ihre Äußerung von gestern so verstanden, daß, sofern dieser Gesetzentwurf in kaum modifizierter Form beschlossen wird, eine Rechtsgrundlage für anonymisierte Mikrodaten gegeben ist, wenn faktische Anonymität – was immer das im einzelnen heißen mag, da sind ja auch die Statistiker mehr gefordert als die Juristen – gewährleistet ist.

Hohmann:
(Referent beim Datenschutzbeauftragten des Landes Hessen)

Das ist richtig. Wenn dieser Entwurf Realität wird, jedenfalls gerade diese spezifische Regelung in § 16 Abs. 4 des Entwurfs des Bundesstatistikgesetzes, dann wird erstmals für wissenschaftliche Nutzer von statistischen Mikrodaten die Barriere, die bisher in § 11 Abs. 5 des geltenden Bundesstatistikgesetzes enthalten war, insoweit gesenkt, als erstmals der vom Bundesverfassungsgericht aufgenommene Begriff der faktischen Anonymisierung legal definiert wird. Ferner wird die aufgrund der bisherigen rechtlichen Gegebenheiten nicht mögliche Übermittlung von Mikrodaten nunmehr in mehrfacher Hinsicht erleichtert.

Prof. Dr. Allerbeck:
(Universität Frankfurt)

Herr Müller, Sie hatten den Datenbedarf der Wissenschaft gestern formuliert. Wenn Sie jetzt aufgrund des Verlaufs des Kolloquiums zurückblicken, wie hat sich dieser Bedarf verändert oder wie ist er – u. U. modifiziert – zu artikulieren?

Prof. Dr. Müller:
(Universität Mannheim)

Ich glaube nicht, daß er sich verändert hat, denn das sind ja grundsätzliche Fragen und Wünsche, die seit langem bestehen. Ich kann vielleicht nur noch einige Differenzierungen anfügen, die sich im Verlaufe dieses Kolloquiums herausgestellt haben. Zunächst

scheint mir, daß die Diskussion und insbesondere auch die Ausführungen von Herrn Hohmann gezeigt haben, daß man unterscheiden muß zwischen

- Datenweitergabe an die Wissenschaft und

- Datenweitergabe an andere Institutionen.

Die Konkretisierung dieses „Wissenschaftsprivilegs" stellt sich jetzt ganz offensichtlich als nächste Aufgabe heraus.

Wenn wir innerhalb dieses Bereiches weiter diskutieren, ist es wichtig zu unterscheiden zwischen dem was Herr Hauser und ich als Wissenschaftsstichprobe gekennzeichnet haben, – für diese Stichproben wären Standardlösungen, wie sie heute morgen von Herrn Hohmann angeregt wurden, anzustreben, und von Fällen, die in besonderer Weise zu behandeln sind. Insbesondere sind dies Daten aus sehr spezifischen Stichproben, die sich auf kleine, auffällige spezifische Populationen beziehen, wie beispielsweise Großunternehmen, Elitepopulationen oder soziale Randgruppen. Diese Fälle müßten eine besondere Behandlung erfahren, ebenso wie die regional tief differenzierten Datenwünsche.

Die Direktive, daß nicht jeder einzelne Datenwunsch gesondert behandelt werden muß, sollte sein: Entwicklung allgemeiner Stichproben für die generelle wissenschaftliche Verwendung, die regional wenig differenziert sind und die bestimmte Anonymisierungsstandards erfüllen. Dabei ist der für wissenschaftliche Zwecke unschädlichste Weg der Anonymisierung in der Regel das Ziehen von Substichproben aus der Gesamtstichprobe; der schädlichste Weg, auf den wir uns als Wissenschaftler nicht einlassen können, ist die Manipulation und Veränderung von Daten.

Einen weiteren Punkt möchte ich noch aufnehmen, der die Frage der Sensibilität von Merkmalen betrifft. In der Diskussion heute morgen ist angeklungen, daß sensible Merkmale wie z. B. Einkommen besonders geschützt werden müssen. Nach meiner Kenntnis ist die Charakterisierung der Sensibilität von Merkmalen nicht eine Frage, die im Datenschutzrecht behandelt wird. Und bei der Anonymisierung geht es nicht um spezifische Merkmale, sondern es geht generell darum, den ganzen Fall nicht zu identifizieren. Auf diesen Punkt könnte Herr Hohmann vielleicht nochmals besonders eingehen.

Was das Bundesstatistikgesetz selbst angeht, sollte man den potentiellen Nutzerkreis klären. Wer ist ein Wissenschaftler? Ich

meine, das Bundesstatistikgesetz sieht eine zu enge Abgrenzung vor. Es muß auf jeden Fall sichergestellt sein, daß nicht nur Beamte mit anonymisierten Einzelangaben arbeiten dürfen; Herr Hohmann hat gestern in seiner Interpretation angedeutet, daß es Möglichkeiten geben würde, auch Personen in den Nutzerkreis mit einzubeziehen, die beispielsweise promovieren oder habilitieren oder als studentische Hilfskräfte an Projekten beteiligt sind. Diese Frage bedarf sicherlich der weiteren Klärung.

Prof. Dr. Allerbeck: (Universität Frankfurt)

Herr Paaß hat uns gestern seine Studie in ihren wesentlichen Punkten vorgetragen, und es ist jetzt in dieser Diskussion, aber vor allem auch durch Herrn Eulers Beitrag deutlich geworden, daß der Begriff des Merkmals durchaus der Problematisierung und der Untersuchung bedarf. Ich wäre sehr daran interessiert, von Ihnen, Herr Paaß, zu erfahren, wie diese Schwierigkeiten verschiedener „Härte" von Merkmalen in unterschiedlichen Datensätzen bei Modellrechnungen angegangen werden könnten. Manche Dinge sind ja sicherlich zufällig, andere vielleicht, wie das Einkommen, das man in der Einkommens- und Verbrauchsstichprobe erklärt oder gegenüber dem Finanzamt erklärt, könnten sich möglicherweise nicht zufälligerweise voneinander unterscheiden. Die Schwierigkeiten, die man sich denken kann, sind zahllos. Gäbe es Möglichkeiten im Rahmen solcher Modellrechnungen, wie Sie sie unternommen haben, diese Dinge aufzugreifen?

Dr. Paaß: (Gesellschaft für Mathematik und Datenverarbeitung, St. Augustin)

Prinzipiell können wir in unserem Reidentifikationsverfahren die jeweilige Qualität und Struktur der Abweichung zwischen Mikrodatenfile und Zusatzwissen detailliert berücksichtigen, und zwar durch die Möglichkeit, Erhebungsfehler und andere Arten von Abweichungen durch das „Fehlermodell" zu modellieren. Wie ich gestern schon ausführte, müssen Abweichungen nicht immer direkte Erhebungsfehler bei der Erhebung des Mikrodatenfiles sein, sondern können auch „Erhebungsfehler" beim Zusatzwissen sein, worunter man beispielsweise auch unterschiedliche Merkmalsdefinitionen zählen könnte. Wegen der in der Regel ungenügenden Kenntnis der Fehlerstruktur wird man nur selten von systematischen Verzerrungen ausgehen und ein entsprechendes „Fehlermodell" konstruieren; meist wird man unabhängige Fehler einfacher Struktur annehmen. Beides läßt sich in unserem Ansatz inkorporieren. Das Problem, das wir hatten, war, daß keine Informationen über die Fehler im Zusatzwissen vorlagen. Wir mußten dazu relativ willkürlich Zahlen greifen. Über die Fehler in der EVS hatten wir Diskussionen mit Herrn Euler, der uns Zahlen nannte, die uns plausibel erschienen und die wir dann auch verwendet haben.

Ich möchte nochmals zurückkommen auf die Ergebnisse der Studie, denn durch meinen gestrigen Vortrag ist vielleicht ein schiefes Bild über die Höhe des Reidentifikationsrisikos entstanden.

Wir sind von einem Public Use File ausgegangen, also einem File, das ohne Restriktionen freigegeben wird mit dem jedermann machen kann, was er will. Daher mußten wir auch Szenarien untersuchen, in denen der Angreifer mit einem sehr extensiven Zusatzwissen ausgestattet war, beispielsweise im Falle des Steuerfahndungs- und des Staatsanwalts-Szenarios. Wenn wir in dieser Diskussion hingegen den Fall betrachten, daß unter gewissen Kautelen diese Daten ausschließlich an Wissenschaftler übermittelt werden, so ist das Zusatzwissen und damit das Reidentifikationsrisiko vermutlich sehr stark eingeschränkt.

Wir haben beispielsweise in allen „erfolgreichen" Szenarien Einkommensinformationen als Zusatzwissen verwandt, da wir annehmen mußten, daß die Steuerfahndung oder sonstige Behörden naturgemäß Informationen über das Einkommen haben. Diese Einkommensinformationen besitzen wegen ihrer Detailliertheit einen sehr hohen Informationsgehalt. Ich glaube nicht, daß bei wissenschaftlichen Institutionen Datenbanken verfügbar sind, in denen als Zusatzwissen Einkommen oder ähnlich detaillierte Variablen vorhanden sind. Ich glaube, daß im wissenschaftlichen Bereich das Zusatzwissen sehr gering ist, wie es auch vom Bundesverfasssungsgericht festgestellt wurde. Insofern würden sich für die Wissenschaft meiner Ansicht nach nur das Konzern- oder das Adreßverlags-Szenario anbieten, bei denen lediglich wenige kategoriale Merkmale im Zusatzwissen vorhanden waren. In diesen Szenarien ergab die Untersuchung gar kein Reidentifikationsrisiko. Dies möchte ich hier nochmals betonen, um die Ergebnisse der Studie ins rechte Licht zu rücken.

Zindler:
(Statistisches Bundesamt)

Erlauben Sie mir ein paar Bemerkungen zu einigen Punkten, die hier in der Diskussion gesagt worden sind.
Zur Deanonymisierung klang gestern die folgende Aussage verschiedentlich durch:

„Aber es ist doch noch nie etwas passiert! Man könnte direkt einen Preis aussetzen für denjenigen, der an einem konkreten Material irgend etwas deanonymisiert!"

Wir haben bis zum Volkszählungsprozeß diesen Standpunkt auch gehabt, sind aber in der Zwischenzeit gründlich belehrt, daß wir zur Selbstzufriedenheit gar keinen Anlaß haben. Ich brauche Ihnen als

Wissenschaftler nicht weiter auszuführen, daß die Erfahrung niemals negative Beweiskraft hat; die Aussage „Es ist noch nie etwas passiert" bedeutet somit keineswegs, daß nicht schon morgen etwas passieren kann. Man muß auch sehen, daß Sie nicht nur in der gleichen Verantwortung wie wir, sondern auch unter den gleichen Strafbestimmungen stehen werden, wenn Sie – wie es im Entwurf des neuen BStatG angestrebt wird – von uns reichlicher mit statistischen Daten versorgt werden können.

Dann möchte ich noch auf einen zweiten Punkt eingehen, der auch heute schon im Referat von Herrn Südfeld angeklungen ist. Die amtliche Statistik kann gar nicht anders, sie muß sich mit Deanonymisierungsmöglichkeiten beschäftigen und mögliche Deanonymisierungsverfahren und -situationen prüfen; das ist keineswegs eine Spielerei. Wir liefern bekanntlich auch Daten an Außenstehende, die nicht zur Wissenschaft gehören, und da muß das Anonymisierungsverfahren sicher sein. Das kann aber nur aufgrund eingehender Untersuchungen in den verschiedensten Richtungen geprüft werden. Das gilt auch für noch ganz andere Bereiche als gerade Einzelmaterial, für die wir auch Untersuchungen durchführen und Schutzmechanismen entwickeln müssen.

Prof. Dr. Allerbeck: (Universität Frankfurt)

Wir haben uns in dieser abschließenden Diskussion das Ziel gesetzt, auf Fragen der Zukunft einzugehen. Wir stellen fest, daß die Zukunft viele Wurzeln in der Vergangenheit hat, und aus den Wortmeldungen schließe ich, daß sich Verästelungen der Wurzeln in der Vergangenheit sehr rasch ergeben. Wir können diesen nachgehen, aber bitte recht kurz.

Prof. Dr. Hauser: (Universität Frankfurt)

Ich wollte eine Zusatzfrage stellen, die sich aber auch mit der Zukunft beschäftigt, und zwar ganz speziell an Herrn Hohmann.

Angenommen, das neue Bundesstatistikgesetz würde verabschiedet werden, sehen Sie dann Möglichkeiten, nach denen die neuen Regelungen des Datenschutzes auch für die vergangene Periode, in der die strengeren Regelungen galten, angewendet werden, d. h., können für die Wissenschaft noch Wissenschaftsstichproben nachträglich aus den Datenbeständen weitergegeben werden, die während der Gültigkeitsperiode früherer Gesetze erhoben werden, oder wie sind die denkbaren Rückwirkungen eines solchen Gesetzes?

Hohmann:
(Referent beim Datenschutzbeauftragten des Landes Hessen)

Es gibt sicher eine Problematik der Rückwirkung von Gesetzen, nur ich denke, daß im Hinblick auf die Frage der Regelung des Statistikgeheimnisses dieses Problem so gewertet werden kann, daß die Weitergaberegelungen sich auf das statistische Material insgesamt beziehen.

Es ist wohl so, daß eine Rückwirkung, wenn sie sich vor allen Dingen auf belastende Eingriffe bezieht, problematisch ist, d. h., wenn unter Umständen Leistungen oder bestehende Leistungsansprüche zurückgenommen werden, wäre eine Rückwirkung verfassungsrechtlich bedenklich. Aber ich glaube nicht, daß eine Statistikgeheimnisregelung unterscheidet zwischen solchen Daten, die unter dem Rechtszustand des Statistikgesetzes 1953–1976 bzw. 1976–1980 und schließlich des achtziger Gesetzes erhoben wurden. Ich glaube, das ist in der Vergangenheit auch so nicht getan worden, sondern es ist jeweils, soweit ich es sehe, die geltende Geheimhaltungsbestimmung auf alle Datenbestände angewendet worden.

Prof. Dr. Esser:
(Zentrum für Umfragen, Methoden und Analysen (ZUMA), Mannheim)

Ich möchte aus der Perspektive von ZUMA bzw. GESIS (Gesellschaft sozialwissenschaftlicher Infrastruktureinrichtungen) kurz meinen Eindruck wiedergeben. Diesen Eindruck möchte ich in drei Punkte gliedern:

1. Die Erfüllung der Aufgaben als Service-Institution.

2. Die Einhaltung methodischer Standards bei Analysen.

3. Die Durchführung von Methodenforschung im weiteren Sinne.

Nun ein paar kurze Bemerkungen zu jedem Punkt. Bei der Bereitstellung von Serviceleistungen ist an zwei Dinge zu denken. Einerseits an eine Erweiterung der Nutzungsmöglichkeiten von zu wissenschaftlichen Zwecken erstellten Individualdatensätzen wie z. B. den Allbus. Wir sind schon daran interessiert, daß im Prinzip solche Studien an den Universitäten auch in der Ausbildung verwandt werden können. Der zweite Punkt in diesem Zusammenhang ist von Herrn Müller kurz angesprochen worden. Es müßte zunehmend – aus wissenschaftlichen Gründen – Wert darauf gelegt werden, daß auch Regionalisierungen und Kontextzuspielungen zu Individualdaten möglich sind. Daraus ergibt sich eine besondere Problematik, die anläßlich einer Zusammenkunft von interessierten Wissenschaftlern mit Vertretern der Bundesforschungsanstalt für Landeskunde und Raumordnung ausführlich in Bad Homburg diskutiert worden ist. ZUMA hat dort aus institutio-

nellen Erwägungen heraus einen sehr restriktiven Standpunkt vertreten. Ich möchte hier den anderen, den wissenschaftlichen Standpunkt in den Vordergrund setzen, nicht zuletzt, um zu verdeutlichen, wie ambivalent die Positionen naturgemäß in diesem Punkt sein müssen. Hier wird wahrscheinlich als einziger Ausweg eine Art von overseer oder irgendeine Art von Treuhänder-Modell notwendig sein, weil sich hier die Anonymisierungsproblematik ganz besonders verstärkt stellt und alle die Maßnahmen, die in diesem Kolloquium vorgeschlagen worden sind, meiner Meinung bei dieser Problematik nicht angemessen sind.

Beim nächsten Punkt, der Einhaltung methodischer Standards muß ich ganz generell festhalten, daß alle Datenmodifikationen als Anonymisierungsmaßnahmen von diesem Standpunkt her nicht erlaubt werden können. Dazu zählen auch Zusammenfassungen in Kategorien. Das Problem der Ungenauigkeit der Messungen ist ohnehin schon gravierend genug. Die Ungenauigkeit der Messungen wird jedoch durch jede Zusammenfassung und durch jede Datenmodifikation immer nur noch erhöht. Und es ist dringend davor zu warnen, anschließend bestimmte Verfahren für Versuche zur nachträglichen Fehlerkorrektur vorzunehmen. Von daher kann ich all diesen Vorstellungen, die die Lösung des Problems in der Vornahme von Datenmodifikation sehen, in keiner Weise zustimmen.

Im Zusammenhang zum dritten Punkt, Methodenforschung, erscheint es notwendig wenigstens für spezifische Zielsetzungen, die möglicherweise in einer Regelung gesondert ausgewiesen werden können oder müssen, weitergehende Datenzusammenführungen auch möglich werden zu lassen. Gestern ist dieser Sachverhalt kurz im Zusammenhang mit der Panel-Problematik diskutiert worden. Ich bin der Auffassung, daß man keine langfristigen Panels laufen lassen kann, ohne daß die Zuordnung zu den Personen sichergestellt ist.

Dieses ist nicht der einzige Aspekt. Dies sei am Problem der Untersuchung von Ausfällen wie etwa im Rahmen der ALLBUS-Nonresponse-Zusatzstudie aufgezeigt. Es ist aus methodischen und wissenschaftlichen Gründen unerläßlich, daß einmal systematisch untersucht wird, wer nicht an den Befragungen teilnimmt, nicht zuletzt, um später den Schutz der Individuen wieder stärker in den Mittelpunkt treten zu lassen, die nicht teilnehmen wollen. Sobald wir z. B. wissen, daß die üblichen Ausfälle unschädlich sind, könnte man natürlich den Schutz der Individuen viel stärker in den Mittelpunkt aller Überlegungen stellen und wirklich zu „milden" Erhebungsformen z. B. in der amtlichen Statistik greifen. Aber einmal

müssen diese Fragen systematisch untersucht werden. Und dazu benötigt man Informationen darüber, wer definitiv verweigert und welche Merkmale diese Personen aufweisen. Dazu muß es im Prinzip noch möglich sein, z. B. durch Datenzusammenführungen für diesen ganz spezifischen Zweck Informationen zu gewinnen, die außerhalb dieser speziellen Zwecksetzung keinesfalls verfügbar sein dürften.

Der letzte Punkt in diesem Zusammenhang betrifft die bisher überhaupt noch nicht untersuchten Validierungsumfragen. Auch hier stellt sich selbstverständlich sofort das Problem der Datenzusammenführung aus anderen Files. Wie will man denn anders überprüfen, ob jemand valide geantwortet hat als auf diese Weise? Die Datenqualität nicht nur der amtlichen Statistik, sondern auch der weite Bereich der alltäglichen Umfrageforschung hängt natürlich nicht nur mit unsystematischen Fehlern zusammen, sondern auch mit den systematischen Fehlern. Dieser Problembereich bedarf dringend einer gründlichen Untersuchung, die aber nur möglich ist, wenn die Zuordnung von Daten aus sehr unterschiedlichen Quellen – für diesen Zweck – nicht ausgeschlossen ist. Es ist in der Tat zu überlegen, ob man nicht für die unterschiedlichen Interessen und Problemfelder auch zu spezifischen Lösungen des Problems der Anonymisierung und des Datenzugangs kommen muß.

Deininger: (Statistisches Landesamt Baden-Württemberg)

Zum Ende des Kolloquiums frage ich mich natürlich: Was haben die zwei letzten Tage gebracht? Wo stehen wir heute in der Diskussion zwischen Sozialwissenschaftlern und Statistikern?

Meiner Meinung nach müßte als erstes noch definiert werden, was „Statistik" ist und was „Wissenschaft". Nicht unbedingt ist universitäre Forschung allein Wissenschaft. Vielmehr müssen auch Kollegen in den Statistischen Ämtern als Wissenschaftler angesprochen werden. Insofern sollten wir nicht einerseits vom Datenzugang der Wissenschaft und andererseits vom Datenzugang der amtlichen Statistik reden. M. E. muß man diese Dinge klar erkennen und feststellen, daß auch die Wissenschaftler in den Statistischen Ämtern Zugangsbeschränkungen gegenüber Einzeldaten unterworfen sind.

Mein zweiter Punkt bezieht sich auf den Begriff des Zugangs zu Individualdaten oder anonymisierten Einzelangaben. Wir haben in den letzten zwei Tagen nur davon gesprochen, wie man derartige Daten an eine neue Stelle außerhalb der Statistischen Ämter transferieren kann. Aufgrund der Definition „Wissenschaft" für den

Bereich der universitären Forschung lautete die Frage: Wie bekommt man Individualdaten ohne Widerspruch zu Anonymisierungsregelungen in universitäre Rechenzentren?

Wenn es aber darum geht, wissenschaftliche Forschung zu betreiben, müßte man sich prinzipieller überlegen, welche Zugangsmöglichkeiten generell bestehen. Gibt es wirklich nur den Zugang durch Weitergabe der Einzeldatensätze?

Zunächst einmal muß schon aus Gründen der Vollständigkeit erwähnt werden, daß auch im wissenschaftlichen Bereich die Masse der Anforderungen – jedenfalls wie sie auf ein Statistisches Landesamt zukommen – durch Aggregatdaten abgedeckt werden können. Und dies zur vollen Zufriedenheit. Allein das Informationssystem Baden-Württemberg, das laut Gesetz nur Aggregatdaten enthalten darf, wird von der Wissenschaft außerhalb des Statistischen Landesamts etwa 1000mal im Jahr benutzt. Es gibt also einen riesengroßen Datenbedarf, der mit dem bestehenden System ohne Individualdaten abgedeckt wird. Darüber hinaus aber steht die Frage, ob es wirklich nur einen Zugang zu Individualdaten gibt, indem man sich Einzeldatensätze geben läßt und diese irgendwo anders speichert. Gibt es nicht vielleicht auch den Weg, den die Wissenschaftler in den Statistischen Ämtern haben: Über die Rechner in den Statistischen Landesämtern so mit Individualdaten (ggf. on-line) zu arbeiten, daß der einzelne Wissenschaftler zwar Berechnungen durchführen kann, aber nur aggregierte Ergebnisse erhält. Für diese wünschenswert breite Möglichkeit des Datenzugangs müßte von der wissenschaftlichen Seite großes Interesse bestehen. Über diesen Lösungsansatz wurde in den beiden letzten Tagen leider nicht diskutiert. Ich will ihn deshalb in den Raum stellen, weil ich hierin eine Möglichkeit sehe, Maximalforderungen des sachlichen Zugangs mit einem realistischen Ansatz zu begegnen, der viele Datenschutzprobleme vermeidet. So gesehen ist dies eine weitere Anonymisierungsmethode: Der sehr freie Zugang besteht in der Berechtigung, mit auf Individualdaten zugreifende Programme zu arbeiten, aber Individualdaten nicht als Ergebnis zu bekommen. Damit könnte der Bedarf im wissenschaftlichen Bereich möglicherweise nicht in allen Teilen - aber jedenfalls an sehr vielen Stellen abgedeckt werden.

Prof. Dr. Allerbeck: (Universität Frankfurt)

Vielen Dank Herr Deininger. Unser Thema war eigentlich auch nicht, auch wenn es dann immer so erscheinen mag, die generelle Zusammenarbeit von Wissenschaft und Statistik, wobei sich beides natürlich überschneidet, sondern das Thema war sehr bewußt eingeschränkt auf diesen sehr kleinen, in mancher Hinsicht

kritischen Bereich. Damit würde ich sehr gerne zur Frage der Verfahren kommen. Herr Zindler hat in seinen Bemerkungen sehr deutlich gemacht, daß nicht nur das „Restrisiko", sondern das Risiko aller Entscheidungen bei der Herausgabe von Einzelmaterial vor allem bei den Statistikern liegt. Das wird nicht weggenommen durch irgendeine noch so kluge Auslegung oder Interpretation. Die Entscheidung wird von dem Statistischen Amt getroffen, das Daten ausliefert, und die Verantwortung liegt bei ihm und kann nicht weggenommen werden. Wir können aber Überlegungen dazu anstellen, wie die Prozeduren, die dazu führen, daß Daten als „faktisch anonym" betrachtet werden, verbessert werden können. D. h. wie die Abläufe so gestaltet werden können, daß es für alle Beteiligten transparenter, kürzer und eben auch für die Statistiker, die die Verantwortung tragen, tragbarer wird, als dies gegenwärtig ist.

Zur Einleitung dieser Runde würde ich gerne unsere Gäste aus den ausländischen Statistischen Ämtern kurz noch einmal bitten, über die Situation in ihren Ländern zu berichten. Herr Dr. Cox und Herr Denham, Sie haben uns über die Verfahren berichtet, die in Ihren Ämtern benutzt werden, um Mikrodaten für allgemeine oder wissenschaftliche Nutzung freizulegen. Sie verwenden in Ihren beiden Ländern recht verschiedene Modelle. Im britischen Fall gibt es eine allgemeine Politik jeweils für die verschiedenen Datentypen (solche mit freiwilliger Beteiligung und solche mit Auskunftspflicht), während in den USA ein Review Board innerhalb des Bureau of the Census besteht, der zu jedem freizugebenden Datensatz eine Empfehlung abgibt, welche in der Regel befolgt wird. Wenn ich hier eine Bemerkung zu einem anderen Gebiet machen darf: Wir haben in Deutschland eine unglückliche Tradition in der Demokratietheorie, sowohl auf die Vereinigten Staaten als auch auf Großbritannien als ein Verfassungsmodell zu schauen und dann beides zu kombinieren und im Ergebnis etwas recht Unerwünschtes zu erhalten, weil wir das Beste beider Welten kombinieren wollten. Dies ist sicherlich ein Fehler, den wir auf diesem Felde vermeiden sollten. Deswegen ist es wichtig, genau zu betrachten, was in diesem Fall die jeweiligen Verfahren sind in den beiden Ländern, die mehr Erfahrung haben. Ich möchte zuerst Sie fragen, Herr Dr. Cox. Sie haben darauf hingewiesen, daß ein Problem, welches Sie jetzt zu lösen versuchen, eine gewisse „Boardroom Mentality" ist, bei der Entscheidungen getroffen werden, vor allem als kollektives Urteil, was ein Problem sein könnte und was nicht, und wie Sie vorhaben, dieses „Boardroom Decision Making" zu ergänzen um die Hinzunahme wissenschaftlicher Verfahren. Könnten Sie uns etwas mehr darüber berichten?

Dr. Cox:
(Bureau
of the Census,
Washington, D. C.)

The common situation is using a "boardroom approach". The people who do this have done it for a number of years, and have developed familiarity with many different microdata sets. You could see that in 1985, they dealt with over ten microdata files in a year.

What we want to give them in addition is some scientific information and some better statistical sense of what disclosure risk means for different kinds of microdata. We hope to do that by developing general principals on which we can build definitions and guidelines. What we also recognize is that each data set is different. And so we accept that in one sense the review will be the same but also that each data set will be scrutinized in terms of its own characteristics.

For example, one data set may have a corresponding matching data set out in the public domain or in the tax office. That data set obviously presents a different set of problems from one that does not. However, if we can develop the right sort of technical and scientific techniques and make these available to the panel, then they will operate not only in an educated way as a group of experts. My feeling is that, as time goes on, we will learn more about this science and I guess you could say that the science will play an increasing role and qualitative judgments a decreasing role.

Specifically we are looking for some written guidelines of a scientific nature. Mr. Südfeld presented some – what I would consider – guidelines this morning in his paper. We have guidelines for tabulations that require that we do not show in tabulations less than five persons. We would like to get similar quantitative measures and guidelines for microdata.

I don't think that this came out in my presentation: We are not in an environment like the statisticians in the Federal Republic of Germany. Our public at this point is certainly sensitive to privacy issues. However, we have not received alot of pressure from that direction. Most of our pressure is coming from data users, who on the one hand want more data and on the other hand want to make sure that the disclosure avoidance methods to be used do not confuse, distort or destroy the use-functions of the data set. Mr. Südfeld spoke about benefit-cost analysis, which is something that we will be doing as well. However, the initial part of our work will be more in terms of trying to measure what the effects of these disclosure avoidance procedures are on data quality.

Denham:
(Office of Population Censuses and Surveys, London)

I have a number of observations, but I will be brief. What strikes me, after these two days, is the comparatively smaller size of operations surrounding the whole business of microdata in Britain – far fewer than here (in West Germany). First, our data protection law. It does not appear to apply to anonymous data; so there is nothing like Herr Homann's office in Britain. As I tried to point out in my presentation, the law in Britain relating both to the census and to other forms of statistics is much more skeletal and, because it is not written in detail, it gives government servants a little more scope.

I also share Dr. Cox's view that while privacy has been a major issue in Britain – it certainly was around the time of 1971 census – the release of microdata in various forms is not currently a major political issue in Britain. But we obviously take a lot of notice of what happens elsewhere in Europe, what happens with the German census and so on, and we are aware that privacy may become an issue at any time.

The result of this is that we are waiting for a decision on census microdata. But we have gone ahead with release of survey data tapes, which we give to a major university data archive to be used by academics. There are a certain amount of regulations over the use of these tapes, in the sense of the general regulations of the data archive, but these are not restrictive and there certainly are not different treatments for different types of scientists. I think it would be wrong to have selective use of such information. I think it is very important that the data are open to all users and they are also open to other people to see what the users have got and what they are doing with it. I think that the idea of having selected users is, from our point of view, a bad one.

We have little "scientific" base for the design of the survey data tapes and we are not anticipating a particularly "scientific" approach if we release microdata from the census. But we will consider some of the things said here. In Britain its very much a matter of political judgement. We tend, if we approach an issue like the release of census microdata, to ask our ministers directly. In other words, the thoughts of (government) officials tend to be in terms of political consequences. I note that there has not been much reference to West German politicians' views on microdata. We have not got a panel considering the release of microdata. This is an interesting idea, but I think that it is a development to be considered for the future in Britain as we release data tapes for only four major surveys and some ad hoc surveys.

Finally, there were some very interesting ideas in Herr Südfeld's paper, particularly the idea of contracts for microdata users. I use the term "belt and braces", in other words, anonymization is carried out as one protection against the disclosure of information. But also to have contracts with users (which we have already with some census statistics in Britain) is a very useful extra protection for microdata. This is a development for us to consider in Britain.

Prof. Dr. Allerbeck: (Universität Frankfurt)

Wie könnte denn im Kontext der deutschen amtlichen Statistik ein Äquivalent eines „Review Board" aussehen. Es wäre sicherlich nicht ausreichend, einfach eine Institution oder eine Semi-Institution zu schaffen, die die Diskussion in anderer Form fortsetzt, sondern es müßte sicherlich eine Daten- und auch Forschungsgrundlage geben, die möglicherweise an irgendeinem Beispiel, sei es die Einkommens- und Verbrauchsstichprobe oder was auch immer, durchspielt, wie ein solches „Wissenschaftsfile" aussehen könnte und welche Modifikationen vorzunehmen sind, die unschädlich sind für Auswertungszwecke. Dies soll nicht etwas sein, was man einfach so am Tisch entscheiden kann, sondern das man vermutlich mit den relativen Kosten und sonstigen Überlegungen ausprobieren muß, um zu sehen, wie so ein Datensatz aussehen könnte, für den dann ein solcher „Board" festzustellen hätte, ob die faktische Anonymität gewährleistet ist.

Ein Problem, das immer besteht, ist, daß sich auch Bedenken akkumulieren können, und zwar unabhängig davon, wie fundiert sie sind, so daß die Risiken immer größer werden und dazu führen, wie es ein bedeutender deutscher Versicherungsvertreter einmal formuliert hat, „daß Mühlsteine unter Wasser gegen Feuer versichert werden". Ich glaube nicht, daß so etwas eine kluge Prozedur ist, genauso wie es unklug wäre, einfach zu sagen „es ist ja nie etwas passiert, also machen wir einfach so weiter". Diese Einstellung wäre sicherlich auch keine Grundlage. Es ist eine vernünftige Risikoabwägung zu treffen. Das ist mein Verständnis des Begriffs der „faktischen Anonymität", der in die Rechtsgrundlage eingegangen ist, d. h., unvernünftige, absurde, nur denkbare Risiken ohne praktische Bedeutung können in diesem Zusammenhang keine hinderliche Funktion haben. Ich glaube, das wird man auch als Nichtjurist – ich bitte mir gegebenenfalls jetzt zu widersprechen – so ausdrücken können. Es geht darum, eine Entscheidung zu treffen, was „faktische Anonymität" im Einzelfall heißt. Dazu wäre es hilfreich, nicht ein System von kumulierenden Aktenvermerken zu haben, sondern ein Gremium, das zu irgendeinem Zeitpunkt eine Entscheidung trifft und erklärt: Wir können gemeinsam nach Kenntnis aller Probleme sagen, faktische

Anonymität ist gewährleistet durch einen solchen Datensatz, wenn er unter bestimmten Kautelen weitergegeben wird. Bestimmte Randbedingungen wären hinzuzunehmen.

Herr Hohmann, habe ich die rechtliche Lage falsch interpretiert?

Hohmann:
(Referent beim Datenschutzbeauftragten des Landes Hessen)

Nein, Herr Allerbeck, es ist ein Vorschlag, den Sie vorgetragen haben, und zwar den Verfahrensvorschlag, einen „Review Board" bei der amtlichen Statistik zu schaffen, um gemeinsam mit einer Feststellungswirkung, dann auch für amtliche Entscheidungen oder zumindest eine Empfehlungswirkung für amtliche Entscheidungen in der Statistik, eine Datenproliferation aus der amtlichen Statistik an die Wissenschaft in einer generalisierbaren, voraussehbaren und methodisch abgesicherten und darüber hinaus zwischen Wissenschaftssystem und amtlicher Statistik konzentrierten Weise möglich zu machen.

Es ist aber zweifellos so, daß man auch daran weiterarbeiten muß, methodisch begründete Kriterien in materieller Hinsicht für die faktische Anonymisierung weiterzuentwickeln, und zwar jenseits der Verfahrenslösung, so daß ich das eine von dem anderen nicht trennen möchte.

Als zweiten Aspekt, den auch Herr Scheuch gestern erwähnt hat und den heute morgen Herr Zindler noch einmal angesprochen hat, könnte ich mir auch vorstellen, daß neben die faktische Anonymisierung, auf die sich die Tagung eigentlich sehr konzentriert hat, zusätzlich noch komplementäre oder unter Umständen auch funktionaläquivalente, rechtliche, organisatorische und methodische Mechanismen treten, die sich jeweils untereinander auch substituieren können.

Ich könnte mir vorstellen, daß ein dritter Aspekt, der wenig angesprochen worden ist, nämlich der Standard der Datensicherung rein technischer Art, innerhalb des Wissenschaftssystems noch einzubeziehen wäre. Ein Teilnehmer hat das umschrieben mit den Worten: „Wir haben auch durch einen Landesdatenschutzbeauftragten quasi den Datensicherungsstandard einer Wissenschaftsinstitution wenn nicht absegnen lassen, so doch zumindest einmal ansehen lassen, ob hinsichtlich dieses Punktes Bedenken bestehen."

Die Datensicherungsstandards beim Empfänger können dafür Hinweise geben, daß technische und organisatorische Maßnahmen in ausreichender Weise getroffen worden sind, um Daten-

proliferationen unbefugter Art an Dritte zu verhindern. Sie könnten zum Beispiel auch dazu beitragen, Datenlieferungen ohne eine Vielzahl weiterer Maßnahmen als unbedenklich erscheinen zu lassen, so daß auch möglicherweise das Aufwandsargument, was ja ein ganz wesentliches wird im Rahmen der faktischen Anonymisierung, dann auch unter Umständen solche Faktoren, wie die Datensicherheit beim Empfänger, miteinbeziehen kann. Die ganze faktische Anonymisierung beinhaltet im Grunde eine Verhältnismäßigkeitsprüfung sehr komplexer, einander ergänzender und komplementärer Maßnahmen, die, abgehoben auf das einzelne zu übermittelnde Datenmaterial, sehr unterschiedlich sein können. Für eine Verfahrensregelung – wie Sie sie vorschlagen, Herr Allerbeck – spricht, daß es diesen einzelnen für Entscheidungen generalisierbaren Gesichtspunkt, noch nicht gibt. Deshalb spricht vieles dafür, möglichst schnell und möglichst bald eine solche Verfahrenslösung zu entwickeln, die jedenfalls in einer Phase, wo man keine allgemeinen materiellen Kriterien hat, ein Entscheidungsverfahren sehr rationalisieren könnte.

Dr. Hamer: (Statistisches Bundesamt)

Es erscheint mir ratsam, zunächst einmal darauf hinzuweisen, daß eine Arbeitsgruppe eingerichtet worden ist, in der die Arbeitsgemeinschaft Sozialwissenschaftlicher Institute und das Statistische Bundesamt sich mit den praktischen Fragen der Anonymisierung weiter beschäftigen werden. Zur Frage, wie man darüber hinaus zu einer gewissen Institutionalisierung kommen kann und welche Aufgaben dabei behandelt werden könnten, meine ich, daß man versuchen sollte, an Hand von sehr konkreten Beispielen oder auch von tatsächlichen Fällen die verschiedenen Anonymisierungsbedingungen zu erläutern, um auf diese Weise eine Art „Kommentar" zu erstellen, in dem u. a. begründet wird, weshalb man zu dieser oder jener Lösung gekommen ist. Dabei sollten meines Erachtens von vornherein auch die Datenschutzbeauftragten einbezogen werden, um einmal im Hinblick auf die Anonymisierung, zum anderen aber auch im Hinblick auf die Datensicherheit bei den Empfängern entsprechende Aussagen und Kommentierungen abgeben zu können. Ein solcher „Kommentar" müßte durch neu aufgetretene Fälle laufend ergänzt werden.

Ich kann mir dagegen nicht vorstellen, wie wir bei unserem zwar zentralen, aber regional aufgefächerten föderalen Statistiksystem zu einer weitergehenden Lösung einer Institutionalisierung kommen können, insbesondere wenn unterschiedliche regionale Anforderungen zu berücksichtigen sind.

Bei diesen Gegebenheiten fällt es mir schwer, mir vorzustellen, wie ein „Board" oder eine ähnliche Einrichtung in praktischer Hinsicht zufriedenstellende Entscheidungen fällen kann. Ob ein solches Gremium grundsätzlich überfordert sein wird, will ich dabei offenlassen. Die von mir erwähnte Herausarbeitung von Leitlinien an Hand sehr konkreter Fälle einschließlich der Kommentierung wäre jedoch auch aus dieser Sicht sehr wichtig und nützlich.

Ich meine, daß man sich aufgrund der heutigen Diskussion die Frage des weiteren Vorgehens noch einmal ganz gründlich durch den Kopf gehen lassen sollte, um sie dann in den zuständigen Gremien weiter zu besprechen.

Deininger: (Statistisches Landesamt Baden-Württemberg)

Verfahrenslösungen für die Anonymisierung sind selbstverständlich eine dringende Sache. Sie müssen aber permanent fortgeschrieben werden. Was heute als Lösung angesehen wird, mag schon morgen als Unsinn bezeichnet werden. Ein „Board" kann natürlich sehr hilfreich sein, aber man muß sich im klaren sein, ob bzw. welche Verantwortung, Funktion und Bedeutung er gegebenenfalls konkret hat.

Aber auch dies ist letztlich kein Allheilmittel. 1983 war bekanntlich eine Volkszählung vorgesehen. Sechs Monate vor dem Zählungsstichtag hatte noch kein Datenschutzbeauftragter etwas gegen die Zählung einzuwenden. Ein „Board" hätte möglicherweise zu diesem Zeitpunkt seine konkrete Zustimmung gegeben. Die Entscheidung aber fiel woanders. In diesem Fall halfen weder „Board" noch Verfahrensregeln.

Noch ein anderer Aspekt zur Frage der Regelungen für die Datenweitergabe. Gestern ist, wenn ich mich richtig erinnere, von Herrn Dr. Fröhner gemahnt worden: „Paßt bitte auf, daß die Erhebungen noch laufen, es gibt eine große Gefahr für den Markt- und Meinungsforschungsbereich." Genau das gilt auch für die gesamte amtliche Statistik und insbesondere im nächsten Jahr für die Volkszählung. Ein „Datenskandal" – was immer dies bedeuten mag (zum Beispiel eine Deanonymisierung von statistischen Daten außerhalb eines statistischen Amtes) – hätte erhebliche Folgen. Dies ist entscheidend. Da hilft dann auch ein Begriff wie „faktische Anonymisierung" und seine Auslegung nicht weiter. Ich will damit zum Ausdruck bringen, daß diese Hauptsorge alle haben müssen, die Wissenschaft genauso wie diejenigen, die Daten sammeln, und die Datenschützer.

Prof. Dr. Hauser: (Universität Frankfurt)

Ich glaube, man sollte bei dieser Überlegung doch die Zeitstruktur genau beachten und vielleicht eine gewisse Prioritätenliste für denkbare Arbeiten zu Grunde legen.

Es ist völlig einsichtig, daß bis zum Zeitpunkt der Volkszählung größtmögliche Vorsicht nötig ist; aber wenn Sie jetzt den Zeitablauf betrachten – frühestens wird das Bundesstatistikgesetz im Sommer verabschiedet, vielleicht auch erst später –, dann würde aus meiner Sicht die nächstliegende Aufgabe sein, sich damit zu beschäftigen, probehalber eine solche Wissenschaftsstichprobe, etwa aus dem Mikrozensus, zu entwickeln. Wenn man noch weitere Tests u. ä. damit verbinden muß, unterschätzt man den Zeitbedarf sicherlich nicht, wenn man von mindestens einem Jahr ausgeht.

Zu dem Zeitpunkt ist die Volkszählung schon vorüber, und wir wissen dann, wie die Dinge gelaufen sind. Diese Vorüberlegungen würden ja nur bedeuten, daß im Statistischen Bundesamt und in den Landesämtern vielleicht im Zusammenhang mit einer Arbeitsgruppe versucht wird, innerhalb des Gesetzes ein Modell für den einfachsten Fall, eben dieser Wissenschaftsstichprobe, zu konstruieren, ohne diese bereits freizugeben. Aber alle nötigen Vorarbeiten, Tests usw., das Abwägen und Entwickeln von einigen Richtlinien, könnten bereits in Bewegung gesetzt werden.

Wenn man mit diesem Prozeß erst nach Abschluß der Volkszählung beginnt, wären 2½ Jahre bis zu seinem Abschluß vergangen. Man darf diesen Prozeß ja in seinem Zeitbedarf nicht unterschätzen. Der Wissenschaft liegen hierfür Erfahrungswerte vor.

Eine gewisse Parallelarbeit wäre in diesem Fall m. E. eine sinnvolle Strategie, und zwar mit der Reihenfolge: Wissenschaftsstichprobe einfacher Art, das Regionalisierungsproblem, das Problem spezieller Gruppen, und zwar auf der Basis einer Stichprobe mit beschränktem Variablenkatalog wie dem Mikrozensus. Hierbei könnte man bereits wichtige Erfahrungen sammeln.

Als nächster Schritt würde sich die Entwicklung einer Wissenschaftsstichprobe auf Basis der EVS 1983 anbieten.

Danach könnte man vielleicht fortschreiten, indem das Problem der völligen Anonymisierung, das auch schon angesprochen wurde und das im Prinzip in dem neuen Bundesstatistikgesetz auch eine Möglichkeit darstellt und sicherlich schwieriger zu lösen ist, aufgegriffen wird. Aber Sie haben zu diesem Problem eine bestimmte Lösung gefunden, und diese Frage scheint mir nicht so oben anzustehen in der Prioritätenskala.

Auch das Problem der besonderen Stichprobe bei einer Treuhandstelle ist erst zu einem späteren Zeitpunkt angehbar, wenn man sicher sein kann, daß das Bundesstatistikgesetz in Kraft getreten ist, und daß tatsächlich bei dem einfacheren Fall der Wissenschaftsstichprobe akzeptable Lösungen in Sicht sind.

Prof. Dr. Allerbeck
Universität Frankfurt

Schlußwort

Vielen Dank, Herr Hauser, das war nahezu unser Schlußwort und hat uns einen wichtigen Gesichtspunkt für die künftige Arbeit mit auf den Weg gegeben. Wir haben keine Zeit zu warten, bis die Volkszählung stattgefunden hat, und bis das neue Statistikgesetz in Kraft getreten ist, um uns dann erst möglicherweise neue Gedanken zu machen. Wir müssen jetzt in Kenntnis all dieser Dinge, die auf uns zukommen, die ersten Schritte unternehmen. Die Arbeitsgemeinschaft Sozialwissenschaftlicher Institute hat sich ja generell zur Maxime gemacht, nicht perfekte Lösungen, die unabhängig von irgendeinem Zeithorizont sind, anzustreben, sondern im überschaubaren zeitlichen Rahmen, das anzugehen, was mit den vorhandenen Mitteln lösbar ist.

Das war eigentlich auch der Geist unserer Besprechungen mit dem Statistischen Bundesamt, seit es eine gemeinsame Arbeitsgruppe gibt. Die konkreten Dinge sollten genau so vorgenommen werden, um zu sehen, was in diesem Bereich machbar ist, und die machbaren Dinge sollen dann auch durchgesetzt werden.

Es ist aber wichtig, auch wenn es nur zu einer Wissenschaftsstichprobe kommt, daß dies nicht in einem Geist der Heimlichkeit geschieht, auch wenn natürlich eine Wissenschaftsstichprobe andere Interessengruppen begehrlich macht, und die dann mit dem Wunsch vorstellig werden, so etwas auch zu bekommen. Skandale, die angesprochen wurden, entstehen immer dann, wenn etwas Unerwartetes, Heimliches plötzlich bekannt wird.

Wenn die Daten der amtlichen Statistik und ihre Verfügbarkeit in moderner Form auf Datenträgern, die ja inzwischen fast in jeder Schule vorhanden sind, zur Verfügung gestellt werden und das Material annonciert ist und jederman das weiß oder wissen könnte, ist das Potential, daß daraus ein Skandal wird, denkbar gering. Deswegen ist Publizität für ein solches Vorhaben nicht etwa schädlich, sondern nützlich.

Erlauben Sie mir am Schluß der Veranstaltung eine persönliche Quintessenz aus dem Symposion. Es scheint mir offensichtlich zu sein, daß es eine einfache einzelne, allseits befriedigende Regelung nicht geben kann, welche den Zugang der Wissenschaft zu den Daten der amtlichen Statistik ermöglicht. Vielmehr wird es hier mehrerer Gesichtspunkte bedürfen, welche nach dem Grundsatz der Verhältnismäßigkeit der Mittel zu gewichten sind. Ein Aspekt pflegt heute stets vernachlässigt zu werden: der der Erfahrung. Erfahrung läßt sich nur schlecht in Form mathematischer Gleichungen abbilden; dies heißt aber nicht, daß auf sie verzichtet werden kann. Institutionen, die der Entscheidungsfindung in diesem Bereich dienen sollen, müssen der Erfahrung hinreichend viel Raum geben. Die technische Entwicklung muß in diesem Zusammenhang stets beachtet werden, insbesondere durch die rasante Entwicklung der Mikroelektronik ändern sich ständig die Voraussetzungen, die zu beachten sind. So bedarf auch die Zusammenarbeit zwischen amtlicher Statistik und der Wissenschaft, wenn es um die Fragen des Datenzugangs geht, stets der Berücksich-

tigung des Stands der Technik. Dies läßt sich nur realisieren, wenn das mit diesem Symposion begonnene Gespräch der Institutionen nicht abreißt. Fraglos bedarf es allgemeiner Regelungen, welche jedoch den Grundsatz der Verhältnismäßigkeit in sich aufnehmen müssen. Ich vermag mir nicht vorzustellen, daß bei der raschen Entwicklung in vielen betroffenen Bereichen im voraus enumerativ und katalogisierend vorgegangen werden kann. Bei allen Regelungen wird letztendlich für jeden einzelnen Datensatz, welcher der Wissenschaft als anonymisierter Individualdatensatz zugänglich gemacht werden soll, die Entscheidung stehen müssen, daß in diesem Fall die Voraussetzungen für die Wahrung der Anonymität der Auskunftsgeber vorliegen.

Die richtige Form von Regelung in diesem diffizilen Bereich hat wohl eine Kombination zu sein: von abstrakten Regelungen, von Restvertrauen und Strafvorschriften, die dann gegebenenfalls – in welcher Mischung auch immer – in die Entscheidung eingehen. Das Symposion hat für eine solche „Mischung" viele wichtige Beiträge bereitgestellt. Jetzt ist es erforderlich, das richtige Mischungsverhältnis zu finden und praktische Erfahrungen mit dem Thema des Symposions zu sammeln: Dem Zugang der Wissenschaft zu den Daten der amtlichen Statistik.

Das rege Interesse an diesem Symposion, das die Teilnehmerzahl weit überstieg, hat deutlich gemacht, daß es sich hierbei nicht um ein esoterisches Thema handelt; vielmehr geht es hier um eine Schlüsselfrage in der Zusammenarbeit der Wissenschaftler in der amtlichen Statistik, den Universitäten und Forschungsinstituten. Diese Zusammenarbeit verspricht Erkenntnisse, deren die moderne Gesellschaft ebenso wie eine aufgeklärte Politik bedarf. Die Gefahr der kollektiven Erblindung besteht nicht nur dann, wenn die Erhebung der notwendigen Daten unterbleibt, sondern auch, wenn der Zugang zu diesen Daten über das von der Vernunft Gebotene hinaus beschränkt wird.

Ich bedanke mich beim Statistischen Bundesamt sehr herzlich für die Gastfreundschaft in den vergangenen Tagen und bis jetzt, und um Nachsicht für die Offenheit in der Ihre Gäste Probleme und Ihre Sicht der Dinge angesprochen haben. Wir haben mit dem Statistischen Bundesamt die Erfahrung gemacht, eigentlich stets vorzüglich kooperiert zu haben, auch als einzelner Wissenschaftler, und heute sind uns die rechtlichen Rahmenbedingungen deutlicher als zuvor geworden, unter denen das Statistische Bundesamt arbeiten muß. Insofern haben die Diskussionen auch von dieser Seite zu einem vertieften Verständnis beigetragen.

Dr. Hamer
Vizepräsident des Statistischen Bundesamtes

Schlußwort

Ich möchte mich seitens der Mitarbeiter des Statistischen Bundesamtes bei allen Teilnehmern herzlich bedanken. Besonders danken möchte ich Ihnen, Herr Professor Allerbeck, für die intensive Vorbereitung der Tagung und für die Diskussionsleitung – und auch dafür, daß Sie viele kontroverse Punkte angesprochen haben, um deutlich zu machen, an welcher Stelle man möglicherweise aneinander vorbeiredet und wo es notwendig ist, daß aufeinander zugegangen wird. Mein Dank geht ebenso an die Referenten, u. a. an die Referenten aus den Vereinigten Staaten und aus Großbritannien. Ich wäre Ihnen, Herr Professor Allerbeck, sehr dankbar, wenn Sie Herrn Professor Kaase, der die Arbeitsgruppe zwischen dem Statistischen Bundesamt und der Arbeitsgruppe Sozialwissenschaftlicher Institute mit ins Leben gerufen hat, noch einmal den Dank des Statistischen Bundesamtes ausrichten würden. Aus meiner persönlichen Sicht möchte ich abschließend vier Punkte erwähnen, die mir zum Verlauf der Tagung besonders wichtig erscheinen.

Erstens hat die Veranstaltung sehr deutlich gemacht, wo das Bedürfnis für anonymisierte Einzelangaben für wissenschaftliche Zwecke und die Notwendigkeit des gemeinsamen Vorgehens gerade auf diesem Gebiet liegen und welche Anforderungen seitens der Wissenschaft gestellt werden.

Der zweite m. E. sehr wichtige Punkt ist, daß wir auf Grund der heutigen Diskussion in der Lage sind festzustellen, daß wir eine gemeinsame Basis gefunden haben, auf der wir uns weiter über den Sachstand informieren können, so daß sichergestellt ist, daß wir von einer einheitlichen Gedankenbasis ausgehen.

Ein dritter Punkt ist, daß wir konkrete Lösungen angesprochen haben. Ich bin Ihnen allen dafür dankbar, daß gerade zum Schluß der Diskussion noch einige wesentliche Gesichtspunkte für das weitere Vorgehen behandelt worden sind.

Ein vierter, außerordentlich wichtiger Punkt scheint mir die Erkenntnis zu sein, die wir gewonnen haben, daß ein enges Zusammengehen zwischen der Wissenschaft und der amtlichen Statistik – unter Einbeziehung des Datenschutzes, der in vieler Hinsicht in diesem Zusammenspiel gefordert ist – in Zukunft unbedingt erforderlich ist.

Ich darf von unserer Seite sagen, daß wir die Arbeiten weiter fördern werden. Im Statistischen Bundesamt freuen wir uns auf die weitere Zusammenarbeit. In diesem Sinne möchte ich mich von Ihnen allen mit herzlichem Dank verabschieden.

Anhang

Rechtsdokumentation zur Entwicklung und zum Stand der Geheimhaltung und Weitergabe von Einzelangaben aus der amtlichen Statistik

Auszug aus dem Gesetz über die Statistik für Bundeszwecke (StatGes) vom 3. September 1953 (Bundesgesetzblatt I S. 1314):

Geheimhaltungspflicht

§ 12[1])

(1) Einzelangaben über persönliche oder sachliche Verhältnisse, die für eine Bundesstatistik gemacht werden, sind, soweit durch Rechtsvorschrift (§ 6) nichts anderes bestimmt ist, von den Auskunftsberechtigten geheimzuhalten. §§ 93, 97, 105 Abs. 1, § 111 Abs. 5 in Verbindung mit § 105 Abs. 1 sowie § 116 Abs. 1 der Abgabenordnung gelten nicht.

(2) Das Statistische Bundesamt, die Statistischen Landesämter und die sonstigen erhebenden Behörden und Stellen sind berechtigt und verpflichtet, den fachlich zuständigen obersten Bundes- und Landesbehörden oder den von ihnen bestimmten Stellen auf Verlangen Einzelangaben auf dem Dienstweg weiterzuleiten, wenn und soweit dies in der die Statistik anordnenden Rechtsvorschrift zugelassen und in den Erhebungsdrucksachen bekanntgegeben worden ist.

(3) Eine Zusammenfassung von Angaben mehrerer Auskunftspflichtiger ist keine Einzelangabe im Sinne dieses Gesetzes.

(4) Veröffentlichungen dürfen keine Einzelangaben im Sinne dieses Gesetzes enthalten.

Auszug aus der Begründung zum obigen Gesetz:

Zu § 12

Absatz 1

Hier wird der Grundsatz festgelegt, daß alle Einzelangaben von allen Auskunftsberechtigten geheimzuhalten sind und insbesondere nicht zu Auskünften und Anzeigen an die Finanzämter benutzt werden dürfen.

[1]) Abs. 1 Satz 2 neu gefaßt durch Art. 52 des Einführungsgesetzes zur Abgabenordnung (EGAO 1977) vom 14. Dezember 1976 (BGBl. I S. 3341).

Absatz 2

Das Interesse des Auskunftsverpflichteten an der Geheimhaltung erstreckt sich aber nicht nur auf das durch Strafvorschriften sanktionierte Verbot der Veröffentlichung oder Bekanntgabe von Einzelangaben, sondern ebenso auf Art und Umfang der Verwertung von Einzelangaben durch die obersten Bundes- und Landesbehörden, für deren Aufgabenbereiche die Statistiken durchgeführt werden. Deshalb bestimmt Absatz 2, daß die Weitergabe von Einzelangaben von der erhebenden Behörde oder Stelle im Wege der dienstlichen Berichterstattung an die fachlich zuständigen obersten Bundes- und Landesbehörden nur zulässig ist, wenn und soweit es den Befragten vorher bekanntgegeben worden ist. Der Rechtsschutz der Befragten erfordert es, daß, wenn schon eine allgemeine statistische Auskunftspflicht begründet wird, die einen Eingriff in die private Rechtssphäre des einzelnen darstellt, dieser auch erfährt, inwieweit und zu welchem Zweck seine Einzelangaben verwertet werden.

Absatz 3

Die Vorschrift gibt eine Definition des Begriffs „Einzelangabe", vor allem um Unsicherheit in der strafrechtlichen Praxis bei der Verfolgung der unbefugten Weitergabe von Einzelangaben zu vermeiden.

Auszug aus dem Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG) vom 14. März 1980 (Bundesgesetzblatt I S. 289):

Geheimhaltung

§ 11

(1) Einzelangaben über persönliche und sächliche Verhältnisse, die für eine Bundesstatistik gemacht werden, sind, soweit durch Rechtsvorschrift nichts anderes bestimmt ist, von den Amtsträgern und für den öffentlichen Dienst besonders Verpflichteten, die mit der Durchführung von Bundesstatistiken betraut sind, geheimzuhalten, es sei denn, daß der Betroffene im Einzelfall in die Übermittlung oder Veröffentlichung der von ihm gemachten Einzelangaben ausdrücklich einwilligt. Die §§ 93, 97, 105 Abs. 1, § 111 Abs. 5 in Verbindung mit § 105 Abs. 1 sowie § 116 Abs. 1 der Abgabenordnung vom 16. März 1976 (BGBl. I S. 613), zuletzt geändert durch Zweites Kapitel Artikel 1 des Gesetzes vom 26. November 1979 (BGBl. I S. 1953), gelten nicht für Personen und Stellen, soweit sie mit der Durchführung von Bundes- und Landesstatistiken betraut sind.

(2) Die Übermittlung von Einzelangaben zwischen den mit der Durchführung einer Bundesstatistik betrauten Personen und Stellen ist zulässig, soweit dies zur Erstellung der Bundesstatistik erforderlich ist.

(3) Das Statistische Bundesamt, die Statistischen Landesämter und die sonstigen erhebenden Stellen und Behörden sind berechtigt und verpflichtet, den fachlich zuständigen obersten Bundes- und Landesbehörden, den von ihnen bestimmten Stellen sowie sonstigen Amtsträgern und für den öffentlichen Dienst besonders Verpflichteten auf Verlangen statistische Einzelangaben zu übermitteln, wenn und soweit diese Übermittlung unter Angabe des Empfängerkreises und der Art des Verwendungszweckes in der die Statistik anordnenden Rechtsvorschrift zugelassen und in den Erhebungsvordrucken bekanntgegeben ist. In dieser Rechtsvorschrift und den Erhebungsvordrucken ist auch anzugeben, ob die Übermittlung mit oder ohne Nennung von Namen oder von Namen und Anschrift zugelassen ist. Aus den Angaben gewonnene Erkenntnisse dürfen nicht zu Maßnahmen gegen den Betroffenen verwendet werden.

(4) Die Geheimhaltungspflicht nach Absatz 1 gilt auch für die Personen, denen nach Absatz 3 Einzelangaben zugeleitet werden.

(5) Einzelangaben, die so anonymisiert werden, daß sie Auskunftspflichtigen oder Betroffenen nicht mehr zuzuordnen sind, dürfen vom Statistischen Bundesamt und von den Statistischen Landesämtern übermittelt werden.

(6) Eine Zusammenfassung von Angaben mehrerer Auskunftspflichtiger ist keine Einzelangabe im Sinne dieses Gesetzes.

(7) Die zur Identifizierung der Auskunftspflichtigen sowie sonstiger Betroffener dienenden Daten, insbesondere Namen und Anschriften, sind zu löschen, wenn ihre Kenntnis für die Erfüllung der Aufgaben auf dem Gebiet der Statistik für Bundeszwecke nicht mehr erforderlich ist. Namen und Anschriften der Auskunftspflichtigen sollen von den übrigen Angaben getrennt und unter besonderem Verschluß gehalten werden.

Auszug aus der Begründung zum obigen Gesetz:

A. Allgemeiner Teil

Anpassung der Geheimhaltungsbestimmungen an neuere Gegebenheiten unter Berücksichtigung der Entwicklungen auf dem Gebiet des Datenschutzes und des Strafrechts.

B. Die einzelnen Bestimmungen

Zu § 11

Die Neufassung des bisher geltenden § 12 soll unter dem unveränderten Grundsatz der Geheimhaltung statistischer Einzelangaben einerseits die Bedingungen für eine ausnahmsweise Weitergabe und Veröffentlichung von Einzelangaben unter Berücksichtigung der gesammelten Erfahrungen klären und präzisieren sowie andererseits den inzwischen eingetretenen Veränderungen auf dem Gebiet des Strafrechts und den allgemeinen Grundsätzen des Bundesdatenschutzgesetzes Rechnung tragen.

Nach wie vor sind grundsätzlich alle Einzelangaben, die für eine Bundesstatistik gemacht werden, von allen mit der Durchführung dieser Statistik betrauten Personen geheimzuhalten. Die Gewährleistung der Geheimhaltung statistischer Einzelangaben dient folgenden Zielen:

- Schutz des einzelnen vor der Offenlegung seiner persönlichen und sachlichen Verhältnisse
- Erhaltung des Vertrauensverhältnisses zwischen den Befragten und den Statistischen Behörden
- Gewährleistung der Zuverlässigkeit der gemachten Angaben und der Berichtswilligkeit der Befragten.

Die statistische Geheimhaltung steht in untrennbarem Zusammenhang mit der statistischen Auskunftspflicht, die es dem Staat erlaubt, tief in die persönliche oder betriebliche Sphäre der Bürger und sonstigen Befragten einzudringen.

Eine Weiterleitung bzw. – nach der Terminologie des Bundesdatenschutzgesetzes – Übermittlung von Einzelangaben an andere Stellen und Personen ist nur im Ausnahmefall zugelassen. Sie darf nur erfolgen, wenn dies in der die jeweilige Statistik anordnenden Rechtsvorschrift ausdrücklich zugelassen und in den Erhebungsvordrucken bekanntgegeben ist. Der einzelgesetzlichen Übung der letzten Jahre folgend sind in der Neufassung des Geheimhaltungsparagraphen die Bedingungen präzisiert worden, unter denen Ausnahmeregelungen von der statistischen Geheimhaltung getroffen werden können. In den jeweiligen Rechtsvorschriften ist anzugeben, welche Daten an welche Empfängerkreise für welche Arten von Verwendungszwecken weitergegeben werden dürfen und in welcher Form dies geschehen soll. Bei der Festlegung der Bedingungen muß ein Ausgleich gefunden werden zwischen den berechtigten Interessen der Befragten am Schutz ihrer Angaben und den berechtigten Interessen der zahlreichen Benutzer an der Auswertung des mit großen Kosten erhobenen statistischen Materials. Außerdem ist auf die Praktikabilität des Verfahrens für die mit der Durchführung der Bundesstatistiken betrauten Statistischen Ämter und sonstigen Stellen zu achten.

Auch alle nicht mit der Durchführung von Bundesstatistiken betrauten Stellen und Personen, denen Einzelangaben aus dem Bereich der Bundesstatistik übermittelt werden, unterliegen der gleichen Geheimhaltungspflicht wie die die Statistik durchführenden Stellen und Personen. Der Kreis der Stellen und Personen, die Einzelangaben bekommen können, ist, der langjährigen einzelgesetzlichen Übung folgend, bei der Neufassung abschließend festgelegt worden. Er ist auf die Bestimmungen über die Strafbarkeit bei Verletzung von Privatgeheimnissen im Strafgesetzbuch abgestellt, so daß jeder Verstoß gegen die Geheimhaltung strafrechtlich belangt werden kann. Übernommen ist auch die Terminologie des Strafgesetzbuches (Amtsträger und für den öffentlichen Dienst besonders Verpflichtete).

Um gelegentlichen Zweifeln zu begegnen, ist bei der Neufassung der Geheimhaltungsbestimmungen auch klargestellt, daß zur Erstellung einer Bundesstatistik Einzelangaben zwischen den mit der Durchführung der Bundesstatistik betrauten Personen und Stellen übermittelt werden dürfen.

Die amtliche Statistik hat mit den Geheimhaltungsbestimmungen und den strengen Bedingungen für Ausnahmeregelungen im Bundesstatistikgesetz und den darauf beruhenden einzelstatistischen Rechtsvorschriften eine langjährige erfolgreiche Praxis auf dem Gebiet des Datenschutzes aufzuweisen. Die Neufassung des Geheimhaltungsparagraphen trägt dem Schutzgedanken des Bundesdatenschutzgesetzes voll Rechnung. Gegenüber den allgemeinen Regeln für die Datenübermittlung in den §§ 10 und 11 BDSG gelten nach den bisherigen gesetzlichen Bestimmungen und der vorliegenden Neufassung der Geheimhaltungsparagraphen im Bereich der Bundesstatistik vor allem in folgender Hinsicht strengere Maßstäbe:

Die Geheimhaltungsbestimmungen des Bundesstatistikgesetzes (BStatG) beziehen sich nicht nur auf natürliche, sondern auch auf juristische Personen.

Nach den Bestimmungen des BStatG sind Einzelangaben grundsätzlich geheimzuhalten; sie dürfen nur im Ausnahmefall weitergeleitet werden und dann auch nur auf Grund einer speziellen Rechtsvorschrift. Das Bundesdatenschutzgesetz geht dagegen davon aus, daß eine Übermittlung von Einzelangaben dann zulässig ist, wenn sie der rechtmäßigen Aufgabenerfüllung oder einem berechtigten Informationsbedürfnis dient und die Persönlichkeitsrechte der Betroffenen nicht unnötig beeinträchtigt (§§ 10, 11 BDSG).

Im Interesse der Erhaltung der Auskunftsbereitschaft der Befragten bleiben in den statistischen Einzelgesetzen – auch wenn Ausnahmen von der Geheimhaltung zugelassen werden – sensible Daten regelmäßig von der Weiterleitung ausgeschlossen. Diese bereits in der Gesetzgebung vorgenommene Einschränkung garantiert die Wahrung der schutzwürdigen Belange der Betroffenen.

Dem Schutz des Bürgers wird ferner dadurch Rechnung getragen, daß in den statistischen Einzelgesetzen in Fällen einer erlaubten Weiterleitung von Einzelangaben häufig nur eine Weiterleitung von Einzelangaben ohne Namen und Anschrift zugelassen ist.

Wenn im Bereich der Statistik Ausnahmen von der Geheimhaltung zugelassen sind, werden bereits seit Jahren in den speziellen Gesetzen der Empfängerkreis und die Art der Verwendung der Angaben, die weitergeleitet werden dürfen, angegeben. Durch die Abstellung auf bestimmte Informationsbedürfnisse werden die allgemeinen Regelungen des BDSG über die Datenübermittlung für den speziellen Bereich der Statistik konkretisiert und präzisiert.

Die Übermittlung nach den Vorschriften dieses Gesetzes muß nach Maßgabe des Verwendungszwecks erforderlich sein:

- bei fachlich zuständigen obersten Bundes- und Landesbehörden zur rechtmäßigen Erfüllung der in ihrer Zuständigkeit liegenden Aufgaben;

- bei den von den fachlich zuständigen obersten Bundes- und Landesbehörden bestimmten öffentlichen und nicht-öffentlichen Stellen zur rechtmäßigen Erfüllung des ihnen erteilten Auftrags;

- bei sonstigen öffentlichen Stellen (z. B. Gemeinden) zur rechtmäßigen Erfüllung der in ihrer Zuständigkeit liegenden Aufgaben;

- bei sonstigen nicht-öffentlichen Stellen muß der Empfänger ein berechtigtes Interesse an der Kenntnis der zu übermittelnden Daten glaubhaft machen; durch die Übermittlung dürfen schutzwürdige Belange des Betroffenen nicht beeinträchtigt werden.

Im übrigen wird bei der Abfassung der statistischen Einzelgesetze darauf zu achten sein, daß sie sich am Bundesdatenschutzgesetz orientieren und insbesondere nicht hinter seinen Schutzbestimmungen zurückbleiben.

Von besonderer Bedeutung sind dabei die Vorschriften des § 14 Abs. 3 BDSG bezüglich der Löschung personenbezogener Daten. Die Aufnahme einer allgemeinen Löschungsvorschrift in das Bundesstatistikgesetz hat sich aus verschiedenen Gründen als nicht zweckmäßig erwiesen. Ausschlaggebend hierfür war insbesondere, daß keine für alle Bundesstatistiken gemeinsam geltenden Kriterien und Voraussetzungen für den Zeitpunkt festgelegt werden können, zu dem personenbezogene Einzelangaben im Sinne des BDSG bzw. Einzelangaben in der weitergehenden Abgrenzung des Bundesstatistikgesetzes – personenbezogene Einzelangaben und Einzelangaben über Unternehmen, Betriebe u. a. Institutionen – gelöscht werden können (§ 14 Abs. 3 Satz 1 BDSG) bzw. zu löschen sind (§ 14 Abs. 3 Satz 2 BDSG). Dies wäre nur durch eine den Rahmen und die Systematik des Bundesstatistikgesetzes durchbrechende Aufzählung der verschiedenen bevölkerungs- und wirtschaftsstatistischen sowie der übrigen statistischen Rechtsgrundlagen möglich, die zudem mit zahlreichen Ausnahmeregelungen zu versehen wäre.

Für jede einzelstatistische Rechtsgrundlage werden daher künftig die nach § 14 Abs. 3 BDSG gebotenen Löschungsregelungen besonders zu prüfen und zu beachten sein.

Zu § 11 Abs. 1

Im Interesse eines möglichst lückenlosen Schutzes des Betroffenen sowie einer praktikablen Handhabung der Geheimhaltungsbestimmungen umfaßt der Begriff der „Einzelangaben" alle für die Bundesstatistik gemachten Einzelangaben über persönliche und sachliche Verhältnisse. Zu den Einzelangaben, die in der Regel geheimzuhalten sind, gehören auch alle nicht einwandfrei anonymisierten Einzelangaben. Die Notwendigkeit eines generellen Geheimhaltungsschutzes auch dieser Einzelangaben ergibt sich aus der beim heutigen Stand der Technik und durch das Vorhandensein zahlreicher Personendateien im öffentlichen und privaten Bereich immer größer werdenden Gefahr nachträglicher Deanonymisierung und damit der Offenbarung von Individualverhältnissen. Im statistischen Dienst können derzeit wegen der Fülle und Differenziertheit des anfallenden und schnell zu verarbeitenden Einzelmaterials meist nur schematische, wenig arbeitsaufwendige Anonymisierungsmethoden (wie z. B. Weglassen des Namens und der Anschrift) angewandt werden, die im allgemeinen keinen hinreichenden Schutz bieten. An der Entwicklung universell anwendbarer und automatisierbarer Verfahren, die die Nichtbestimmbarkeit von Einzelangaben ausreichend sicherstellen, wird gearbeitet, wobei zu berück-

sichtigen ist, daß im Bereich der Wirtschaftsstatistik die Möglichkeit der Deanonymisierung auf Grund spezieller Kenntnisse eher gegeben ist.

Die Nichtbestimmbarkeit von Einzelangaben hängt jedoch nicht nur von einwandfreien Verfahren, sondern auch davon ab, daß ihre einheitliche Anwendung durch alle in § 11 Abs. 3 genannten weitergabeberechtigten Stellen und Behörden in Bund und Ländern gewährleistet ist. Ähnliche Einschränkungen gelten auch für die Prüfung, ob Einzelangaben offenkundig sind. Bei beiden Arten von Einzelangaben ist daher eine Einbeziehung in den Geheimhaltungsschutz unabdingbar, es sei denn, der Gesetzgeber ließe eine Weiterleitung ausdrücklich zu.

Die Strafbarkeit bei einer Verletzung der Geheimhaltung findet ihre Grenze dort, wo nach der in den Statistischen Ämtern vorhandenen Sachkenntnis eine für die Statistik gemachte Angabe nicht mehr einem einzelnen zuzuordnen ist.

An die Stelle der als zu eng und zu unklar empfundenen Formulierung „Auskunftsberechtigter" tritt die Formulierung „Amtsträger und für den öffentlichen Dienst besonders Verpflichtete, die mit der Durchführung von Bundesstatistiken betraut sind". Diese Formulierung stellt klar, daß alle mit der Durchführung von Bundesstatistiken (amtlich) betrauten Personen gemeint sind, und nimmt Bezug auf die Bestimmungen über die Strafbarkeit bei Verletzung von Privatgeheimnissen (§§ 203 bis 205 StGB).

Durch die Abstellung der Geheimhaltungsverpflichteten auf den im Strafgesetzbuch festgelegten Täterkreis wird erreicht, daß der Personenkreis, der zur statistischen Geheimhaltung verpflichtet ist, derselbe ist, der bei Verstoß gegen die Geheimhaltungspflicht strafrechtlich belangt werden kann.

Zum Kreis der mit der Durchführung von Bundesstatistiken betrauten Personen rechnen neben den Angehörigen der Statistischen Ämter u. a. auch Zähler und Interviewer, Beschäftigte der mit der maschinellen Aufbereitung von Bundesstatistiken oder entsprechenden Teilarbeiten beauftragten Landesrechenzentren oder private Firmen sowie Werkvertragspartner in Heimarbeit. Soweit die genannten Personen keine Amtsträger sind, sind sie nach dem Verpflichtungsgesetz förmlich zu verpflichten. Aus Zweckmäßigkeitsgründen ist im BStatG die Möglichkeit eines Verzichts auf die Geheimhaltung durch den Betroffenen ausdrücklich zugelassen.

Die Erweiterung des § 11 Abs. 1 Satz 2 über den Ausschluß der Beistands- und Anzeigepflichten gegenüber den Finanzämtern auf Landesstatistiken ist im Hinblick darauf erfolgt, daß eine unterschiedliche Behandlung der Stellung des Auskunftspflichtigen bei Bundes- und Landesstatistiken nicht gerechtfertigt erscheint. Da es sich bei den in § 12 Abs. 1 Satz 2 bisheriger Fassung außer Anwendung gesetzten Bestimmungen der Abgabenordnung um Bundesrecht handelt, konnten diese Bestimmungen für den Bereich der Landesstatistiken nicht durch Landesrecht außer Kraft gesetzt werden, es bedarf vielmehr einer bundesrechtlichen Regelung.

Zu § 11 Abs. 2

Um gelegentlich aufgetretene Zweifel zu beseitigen, ob die Übermittlung statistischer Einzelangaben zwischen den verschiedenen mit der Durchführung von Bundesstatistiken betrauten Stellen zulässig ist, erscheint die Aufnahme eines neuen Absatzes 2 in § 11 zweckmäßig, der klarstellt, daß innerhalb dieses Kreises Einzelangaben zur Erstellung einer Bundesstatistik weitergeleitet werden können. Dies gilt insbesondere bei Einschaltung von Landesrechenzentren und Privatfirmen (z. B. Service-Unternehmen für Lochkarten), aber auch in den Fällen, in denen die Statistischen Landesämter und die sonstigen mit der Durchführung von Bundesstatistiken beauftragten Stellen dem Statistischen Bundesamt Einzelangaben zur Erfüllung der dem Statistischen Bundesamt nach § 3 Abs. 1 Nr. 1 und 2 und § 13 obliegenden Aufgaben zuleiten.

Zu § 11 Abs. 3

Die bisher in § 12 Abs. 2 enthaltene Ausnahmeregelung von der statistischen Geheimhaltung wird in der Neufassung unter Berücksichtigung der herrschenden Gesetzgebungs- und Verwaltungspraxis präzisiert.

Dies gilt zunächst für den Kreis der möglichen Empfänger von Einzelangaben. Neben die bisher genannten fachlich zuständigen obersten Bundes- und Landesbehörden sowie die von ihnen bestimmten Stellen treten sonstige Amtsträger und für den öffentlichen Dienst besonders Verpflichtete (z. B. in Gemeinden oder in Instituten, die Aufgaben der öffentlichen Verwaltung wahrnehmen oder für eine Behörde oder sonstige Stelle Aufgaben der öffentlichen Verwaltung ausführen). Damit ist der Empfängerkreis abschließend aufgezählt. Auch dieser Kreis ist so abgegrenzt, daß er bei Verstoß gegen die Geheimhaltungspflicht strafrechtlich belangt werden kann. Soweit es sich nicht um Amtsträger handelt, sind die Empfänger nach dem Verpflichtungsgesetz förmlich zu verpflichten. Im Hinblick auf den Schutz des Betroffenen soll künftig die Übermittlung von Einzelangaben nicht mehr wie bisher „auf dem Dienstweg", sondern direkt durch die Statistischen Ämter erfolgen, weil sie mit dem erweiterten Empfängerkreis aus Zweckmäßigkeitsgründen unmittelbar verkehren sollen und müssen.

Ausnahmen von der statistischen Geheimhaltung sind nur durch die Rechtsvorschrift zugelassen, in der die jeweilige Statistik angeordnet wird. Im Interesse eines weitgehenden Schutzes des einzelnen ist es damit nicht der Verwaltung, sondern dem Gesetzgeber überlassen, bei jeder neuen Statistik eine Abwägung zwischen den Interessen der Befragten an der Geheimhaltung ihrer Angaben und den Interessen der Konsumenten an einer weiteren Verwertung des gewonnenen statistischen Einzelmaterials vorzunehmen.

Wie bereits dargelegt, erfordert der Eingriff in die private Rechtssphäre des einzelnen, wie er durch eine allgemeine statistische Auskunftspflicht begründet wird, eine Information des Auskunftspflichtigen, inwieweit und zu welchem Zweck seine persönlichen und sachlichen Angaben verwertet werden. Im Interesse einer besseren Kontrollmöglichkeit durch den Betroffenen und einer größeren Transparenz des statistischen Datenflusses ist es deshalb erforderlich, daß der Kreis der Empfänger statistischer Einzelangaben, die Art der Angaben

und ihre Verwendung in der die Statistik anordnenden Rechtsvorschrift angegeben werden, wie es auch bisher schon regelmäßig in den betreffenden Einzelgesetzen geschehen ist. Dabei sind häufig abgestufte Regelungen für die verschiedenen Empfängerkreise geboten, und zwar sowohl hinsichtlich der weiterleitungsfähigen Angaben als auch hinsichtlich der Art der zugelassenen Verwendungszwecke. Es ist davon auszugehen, daß sensible Daten, insbesondere auf einzelne Unternehmen oder Arbeitsstätten bezogene wirtschaftliche Daten wie Angaben über Kostenstrukturen, aber auch Angaben über Einkommens- und Gesundheitsverhältnisse der Bevölkerung wie bisher regelmäßig von der Übermittlung ausgeschlossen bleiben und daß die Verwendungsbereiche so weit konkretisiert werden, wie es jeweils im Hinblick auf die schutzwürdigen Belange der Befragten und die berechtigten Interessen der Empfänger an der Weiterverwertung der Einzelangaben nötig und möglich ist. Im Interesse der Betroffenen soll in den Einzelrechtsvorschriften auch darüber befunden werden, ob es erforderlich ist, Namen bzw. Namen und Anschrift weiterzuleiten oder nicht. Auch dies wurde in den letzten Jahren bereits in den diesbezüglichen Einzelgesetzen geregelt. Außerdem sind die Ausnahmen von der Geheimhaltung auch in den Erhebungsvordrucken bekanntzugeben.

In den Geheimhaltungsparagraphen wurde ein ausdrücklicher Hinweis darauf aufgenommen, daß aus den Einzelangaben gewonnene Erkenntnisse nicht zu Maßnahmen gegen den Betroffenen verwendet werden dürfen.

Zu § 11 Abs. 4

Der neueingefügte Absatz 4 bestimmt, daß auch alle nicht mit der Durchführung von Bundesstatistiken betrauten Personen, denen zulässigerweise geheimhaltungspflichtige Angaben zugeleitet werden, der Geheimhaltungspflicht nach Absatz 1 unterliegen.

Zu § 11 Abs. 5

Um eine Unsicherheit in der strafrechtlichen Praxis bei der Verfolgung der unbefugten Weitergabe von Einzelangaben zu vermeiden, stellt diese Vorschrift wie bisher klar, daß eine Zusammenfassung von Angaben mehrerer Auskunftspflichtiger keine Einzelangabe im Sinne dieses Gesetzes ist.

Auszug aus Beschlußempfehlung und Bericht des Innenausschusses des Deutschen Bundestages zu § 11 Abs. 5 des obigen Gesetzes (Drucksache 8/3413 vom 20. 11. 1979):

Der Innenausschuß des Deutschen Bundestages hat die Auffassung vertreten, „daß die Möglichkeit einer Deanonymisierung absolut nicht ausgeschlossen werden könne. Andererseits dürfe daraus nicht der Schluß gezogen werden, eine Übermittlung anonymisierter Daten sei in jedem Fall ausgeschlossen. Vielmehr müsse vor Übermittlung anonymisierter Daten sichergestellt sein, daß nach den in den Statistischen Ämtern vorliegenden Kenntnissen die Möglichkeit einer Deanonymisierung der übermittelten – nach Auffassung der Statistischen Ämter ausreichend anonymisierten – Einzelangaben zweifelsfrei ausge-

schlossen wird. Der Ausschuß geht davon aus, daß die Statistischen Ämter sich über den Anonymisierungsgrad bei jeder Einzelstatistik verständigen und ein einheitliches Vorgehen bei der Beurteilung von Übermittlungsbegehren anonymisierter Einzelangaben vereinbaren".

Auszug aus dem Urteil des Bundesverfassungsgerichts zum Volkszählungsgesetz 1983 (VZG 1983) bezüglich der Übermittlung von Einzelangaben an die Wissenschaft (BVerfGE 65, S. 69f.):

5. Demgegenüber verletzt § 9 Abs. 4 VZG 1983 nicht das allgemeine Persönlichkeitsrecht. Diese Vorschrift gestattet für wissenschaftliche Zwecke die Übermittlung bestimmter Einzelangaben an Amtsträger und für den öffentlichen Dienst besonders Verpflichtete. Die Übermittlung hat sich in den Grenzen des für wissenschaftliche Zwecke Erforderlichen zu halten, Name und Anschrift dürfen überhaupt nicht weitergegeben werden. Die Regelung folgt damit der Erkenntnis, daß für die meisten Untersuchungsbereiche ein direkter Personenbezug nicht erforderlich ist; denn der Wissenschaftler ist regelmäßig nicht an der einzelnen Person interessiert, sondern an dem Individuum als Träger bestimmter Merkmale. Da bei den Übermittlungsadressaten des § 9 Abs. 4 VZG 1983 regelmäßig kaum Zusatzwissen vorhanden sein wird, ist nach dem derzeitigen Erkenntnis- und Verfahrensstand nicht davon auszugehen, daß der Schutz des informationellen Selbstbestimmungsrechts bei der Verarbeitung von Daten nach § 9 Abs. 4 VZG 1983 über die durch § 5 BDSG, § 11 Abs. 5 BStatG, § 9 Abs. 5 VZG 1983 und die Kontrolle der Datenschutzbeauftragten des Bundes und der Länder gewährleisteten Sicherungen heraus weitere Vorkehrungen von Verfassungswegen erfordert.

Auszug aus dem Entwurf eines Gesetzes über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG) der Bundesregierung (Drucksache 10/5345 vom 17. 4. 86):

Geheimhaltung

§ 16

(1) Einzelangaben über persönliche und sachliche Verhältnisse, die für eine Bundesstatistik gemacht werden, sind von den Amtsträgern und für den öffentlichen Dienst besonders Verpflichteten, die mit der Durchführung von Bundesstatistiken betraut sind, geheimzuhalten, soweit durch besondere Rechtsvorschrift nichts anderes bestimmt ist. Dies gilt nicht für

1. Einzelangaben, in deren Übermittlung oder Veröffentlichung der Befragte schriftlich eingewilligt hat, sowie für

2. Einzelangaben aus allgemein zugänglichen Quellen, wenn sie sich auf die in § 15 Abs. 1 genannten öffentlichen Stellen beziehen, auch soweit eine Auskunftspflicht aufgrund einer eine Bundesstatistik anordnenden Rechtsvorschrift besteht.

Die §§ 93, 97, 105 Abs. 1, § 111 Abs. 5 in Verbindung mit § 105 Abs. 1 sowie § 116 Abs. 1 der Abgabenordnung vom 16. März 1976 (BGBl. I S. 613), zuletzt geändert durch Artikel 1 des Gesetzes vom 19. Dezember 1985 (BGBl. I S. 2436), gelten nicht für Personen und Stellen, soweit sie mit der Durchführung von Bundes-, Landes- oder Kommunalstatistiken betraut sind.

(2) Die Übermittlung von Einzelangaben zwischen den mit der Durchführung einer Bundesstatistik betrauten Personen und Stellen ist zulässig, soweit dies zur Erstellung der Bundesstatistik erforderlich ist.

(3) Das Statistische Bundesamt darf an die statistischen Ämter der Länder die ihren jeweiligen Erhebungsbereich betreffenden Einzelangaben für Sonderaufbereitungen auf regionaler Ebene übermitteln. Für die Erstellung der Volkswirtschaftlichen Gesamtrechnungen des Bundes und der Länder dürfen sich das Statistische Bundesamt und die statistischen Ämter der Länder untereinander Einzelangaben aus Bundesstatistiken übermitteln.

(4) Einzelangaben dürfen vom Statistischen Bundesamt und den statistischen Ämtern der Länder übermittelt werden, wenn sie so anonymisiert sind, daß sie Auskunftspflichtigen oder Betroffenen nicht zuzuordnen sind. Wenn sie nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft zugeordnet werden können, dürfen sie für wissenschaftliche Zwecke an Amtsträger oder für den öffentlichen Dienst besonders Verpflichtete in Hochschulen oder sonstigen Einrichtungen mit der Aufgabe unabhängiger wissenschaftlicher Forschung übermittelt werden.

(5) Die Übermittlung aufgrund einer besonderen Rechtsvorschrift oder nach Absatz 4 ist nach Inhalt, Empfänger, Datum und Zweck der Weitergabe von Statistischen Ämtern aufzuzeichnen. Die Aufzeichnungen sind mindestens fünf Jahre aufzubewahren.

(6) Die aufgrund einer besonderen Rechtsvorschrift oder nach Absatz 4 Satz 2 übermittelten Einzelangaben dürfen von den Empfängern nur für die Zwecke verwendet werden, für die sie übermittelt wurden. Bei den Empfängern muß durch organisatorische und technische Maßnahmen sichergestellt sein, daß nur Amtsträger oder für den öffentlichen Dienst besonders Verpflichtete Kenntnis von den Einzelangaben erhalten.

(7) Die Geheimhaltungspflicht nach Absatz 1 gilt auch für die Personen, die bei Stellen beschäftigt sind, denen Einzelangaben aufgrund einer besonderen Rechtsvorschrift oder nach Absatz 4 zugeleitet werden.

(8) Eine Zusammenfassung von Angaben mehrerer Befragter gilt nicht als Einzelangabe im Sinne dieses Gesetzes.

Auszug aus der Begründung zum obigen Gesetzentwurf:

A. Allgemeiner Teil

Statistische Geheimhaltung (§ 16)

Die Vorschrift enthält gegenüber dem BStatG von 1980 gravierende Einschränkungen der Möglichkeit, Einzelangaben zu übermitteln. Sie verstärkt damit die statistische Geheimhaltung, die grundlegende Voraussetzung für die Auskunftsbereitschaft und damit für einen möglichst hohen Grad an Genauigkeit und Wahrheitsgehalt der erhobenen Daten ist.

B. Besonderer Teil

Zu § 16 (Geheimhaltung)

Die Geheimhaltung der statistischen Einzelangaben ist seit jeher das Fundament der Bundesstatistik. Ihre Gewährleistung dient, wie bereits in der Begründung zum Bundesstatistikgesetz von 1980 (vgl. BT-Drucksache 8/2517, S. 16) ausgeführt worden ist, folgenden Zielen:

- Schutz des einzelnen vor der Offenlegung seiner persönlichen und sachlichen Verhältnisse;

- Erhaltung des Vertrauensverhältnisses zwischen den Befragten und den statistischen Ämtern;

- Gewährleistung der Zuverlässigkeit der Angaben und der Berichtswilligkeit der Befragten.

Das Bundesverfassungsgericht hat im Volkszählungsurteil die herausragende Bedeutung des Statistikgeheimnisses hervorgehoben. Es betrachtet den Grundsatz, die zu statistischen Zwecken erhobenen Einzelangaben strikt geheimzuhalten, nicht nur als konstitutiv für die Funktionsfähigkeit der Bundesstatistik, sondern auch im Hinblick auf den Schutz des Rechts auf informationelle Selbstbestimmung als unverzichtbar. Auf der Grundlage seiner Rechtssprechung sind die Ausnahmeregelungen gegenüber dem bisherigen Recht (vgl. § 11 Abs. 3 Bundesstatistikgesetz) weitergehenden Restriktionen unterworfen.

Zu Absatz 1

Einzelangaben sind Erklärungen, die von einem Auskunftspflichtigen oder Befragten in Erfüllung seiner statistischen Auskunftspflicht nach § 15 oder – bei Erhebung ohne Auskunftspflicht – freiwillig abgegeben werden. Diese Angaben sind dazu bestimmt, in einer Bundesstatistik, d. h. in der Zusammenfassung von Einzelangaben mehrerer Befragter und damit im statistischen Ergebnis, unterzugehen.

Wie bisher sind nach Nummer 1 nicht geheimhaltungsbedürftig Einzelangaben, wenn der Betroffene in ihre Übermittlung oder Veröffentlichung ausdrücklich eingewilligt hat. Die Schriftform der Einwilligung wurde – der Praxis entsprechend – ausdrücklich in den Gesetzestext aufgenommen. Nicht geheimhaltungsbedürftig sind auch Sachverhalte, die juristische Personen des öffentlichen Rechts, Behörden des Bundes und der Länder sowie Gemeinden und Gemeindeverbände betreffen, wenn sie unmittelbar aus allgemein zugänglichen Quellen von jedermann entnommen werden können (Nummer 2). Darunter fallen insbesondere Einzelangaben, die bereits durch die Presse oder andere Publikationsorgane mitgeteilt wurden. Dabei ist es unerheblich, ob diese Einzelangaben aufgrund einer statistischen Auskunftsverpflichtung abgegeben wurden. Daten öffentlicher Stellen, die nicht auch in allgemein zugänglichen Quellen der Öffentlichkeit zur Verfügung stehen, sowie alle Einzelangaben privater Betroffener, die inhaltsgleich auch allgemein zugänglich sind, sind von der Geheimhaltungspflicht nicht ausgenommen.

Weitere Ausnahmen von der statistischen Geheimhaltung bedürfen – soweit sie nicht durch das Bundesstatistikgesetz selbst geregelt sind (vgl. Absätze 2, 3 und 4) – einer ausdrücklichen Zulassung durch besondere Rechtsvorschrift in einem eine Bundesstatistik anordnenden Bundesgesetz. Ob und inwieweit solche Ausnahmen in Betracht kommen, entscheidet der Bundesgesetzgeber konkret und abschließend beim Erlaß der die Statistik anordnenden Rechtsvorschrift. Er muß hierbei die Auflagen des Volkszählungsurteils des Bundesverfassungsgerichts berücksichtigen, durch die die Bedeutung des Grundsatzes der Trennung von Statistik und Vollzug herausgestellt worden ist. Eine Auswirkung dieses Grundsatzes ist es, daß der Gesetzgeber bei Übermittlungsregelungen an Stellen außerhalb der Statistischen Ämter des Bundes und der Länder berücksichtigen muß, daß es auch dort einer Organisation bedarf, die die Zweckbindung ebenso sichert, wie es innerhalb der statistischen Ämter der Fall ist.

Das Urteil des Bundesverfassungsgerichts zum Volkszählungsgesetz verpflichtet darüber hinaus den Gesetzgeber wie auch den Rechtsanwender dazu, die Übermittlung davon abhängig zu machen, daß der Übermittlungszweck im Einzelfall nicht auf andere, den Betroffenen weniger belastende Art erfüllt werden kann. Die Übermittlung ist danach beispielsweise dann unzulässig, wenn Zusammenfassungen in statistischen Ergebnissen oder anonymisierte Einzelangaben ausreichen, den Informationszweck zu erfüllen.

Zu Absatz 3

Satz 1 räumt den Statistischen Ämtern der Länder den Bedürfnissen der Praxis entsprechend die Befugnis ein, bei Statistiken, die nach der einzelgesetzlichen Regelung vom Statistischen Bundesamt erhoben werden, regionale Sonderaufbereitungen für ihre Erhebungsbereiche vorzunehmen. Das Statistische Bundesamt darf dementsprechend die dafür erforderlichen Einzelangaben zur Verfügung stellen.

Nach Satz 2 dürfen die zur Erstellung der Volkswirtschaftlichen Gesamtrechnungen des Bundes und der Länder erforderlichen Einzelangaben zwischen dem Statistischen Bundesamt und den Statistischen Landesämtern übermittelt werden, um Vollständigkeit und Einheitlichkeit der Berechnung auf Bundes- und Länderebene im Rahmen der bestehenden Arbeitsteilung zu gewährleisten.

Zu Absatz 4

Satz 1 entspricht dem § 11 Abs. 5 des Bundesstatistikgesetzes von 1980. Das mit dieser Vorschrift seinerzeit verfolgte Ziel, der Wissenschaft und anderen Stellen in gewissem Umfang Daten zur eigenen Aufbereitung unter Wahrung des Datenschutzes zur Verfügung zu stellen, hat sich angesichts der fortschreitenden Möglichkeiten der Deanonymisierung nur sehr eingeschränkt verwirklichen lassen. Dieser Entwicklung wird im neuen Bundesstatistikgesetz insoweit begegnet, als nunmehr der Wissenschaft Daten übermittelt werden können, die eine Deanonymisierung zwar nicht mit Sicherheit ausschließen, aber Betroffenen nur zugeordnet werden können, wenn der Datenempfänger einen unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft erbringen müßte. Die gesetzliche Neuregelung knüpft damit an den Begriff der faktischen Anonymität an, wie er durch die European Science Foundation definiert worden ist (vgl. auch 4. Tätigkeitsbericht des Bundesbeauftragten für den Datenschutz, BT-Drucksache 9/1243, S. 50). In der Regel wird faktische Anonymität nur auf der Grundlage von Stichproben aus dem Datenmaterial herstellbar sein (vgl. auch 6. Tätigkeitsbericht des Bundesbeauftragten für den Datenschutz, BT-Drucksache 10/877, S. 60). Wirtschaftsstatistische Daten eignen sich zumindest nicht generell für eine Anonymisierung (Protokoll der Sitzung des Ausschusses für Wirtschaft des Deutschen Bundestages, Arbeitsgruppe „Statistik", vom 17. September 1979, S. 77).

Bei der Übermittlungsregelung für wissenschaftliche Zwecke wurde der Empfängerkreis mit Rücksicht auf das vorhandene Restrisiko einer Deanonymisierung auf Amtsträger und für den öffentlichen Dienst besonders Verpflichtete und damit auf einen Kreis beschränkt, der bei unbefugter Offenbarung strafrechtlich belangt werden kann.

Zu Absatz 5

Die Aufzeichungspflicht für die Statistischen Ämter soll einerseits eine effektive Kontrolle durch die Datenschutzbeauftragten über die Einhaltung der Übermittlungsvorschriften gewährleisten, andererseits dem Betroffenen die Verfolgung seiner Rechte erleichtern, wenn er sich gegen eine Übermittlung seiner Daten wenden will.

Zu Absatz 6 und Absatz 7

Die Vorschriften verstärken als zusätzliche Sicherungsmaßnahmen die Zweckbindung und Geheimhaltung übermittelter Einzelangaben.

Zu Absatz 8

Das in den Statistischen Ämtern des Bundes und der Länder seit jeher praktizierte Verfahren bei der Veröffentlichung statistischer Ergebnisse hat sich in der Praxis bewährt. Hiernach dürfen Veröffentlichungen grundsätzlich keine Angaben über weniger als drei Auskunftspflichtige oder Betroffene enthalten. Bereits im Gesetzgebungsverfahren des Gesetzes über die Statistik für Bundeszwecke von 1953 wurde überprüft, ob die Veröffentlichung statistischer Ergebnisse davon abhängig gemacht werden muß, daß „bei der

Zusammenfassung von Angaben Rückschlüsse auf Einzelangaben nicht möglich sind" (vgl. Kurzprotokoll des Bundestags-Ausschusses für Wirtschaftspolitik zur Sitzung am 24. Juni 1953). Hiervon wurde abgesehen, weil wegen der zum Teil tiefen Gliederung der in Statistischen Ämtern zu verarbeitenden Angaben nicht in jedem Einzelfall mit dem dafür erforderlichen Aufwand an Personal und Kosten in der für aktuelle statistische Ergebnisse zur Verfügung stehenden Zeit festgestellt werden kann, ob ausnahmsweise einmal aus einem statistischen Aggregat Rückschlüsse auf eine Einzelangabe möglich sind.