Die faktische Anonymität von Mikrodaten: Ergebnisse und Konsequenzen eines Forschungsprojektes

Wirth, Heike

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Die faktische Anonymität von Mikrodaten: Ergebnisse und Konsequenzen eines Forschungsprojektes

Von Heike Wirth

Erhebungen der amtlichen Statistik stellen für die Untersuchung vieler Forschungsfragen seit langem eine außerordentlich wichtige und umfangreiche Datenressource dar. In den letzten Jahrzehnten hat sich das Nutzungsbedürfnis bezüglich dieser Daten jedoch nachhaltig geändert. Die Weiterentwicklung und Verfeinerung statistischer Analyseverfahren mit hohem Erkenntniswert und die verbesserten Möglichkeiten der Datenverarbeitung erlauben nicht nur eine stärkere Nutzung von Massendaten, sondern setzen vielfach auch die Verwendung von Individualdaten voraus. Der hieraus resultierende, zunehmende Bedarf an Individualdaten der amtlichen Statistik konnte allerdings nicht annähernd befriedigt werden, da nach dem Bundesstatistikgesetz von 1980 Individualdaten nur übermittelt werden durften, wenn sie absolut anonym waren. Spezifisch auf wissenschaftliche Nutzungsbedürfnisse ausgerichtet, wurde daher im neuen Bundesstatistikgesetz (1987) das Konzept der faktischen Anonymität eingeführt. Das Anonymisierungsprojekt hatte das Ziel, Empfehlungen für die konkrete Umsetzung der faktische Anonymität zu entwickeln.

1. Hintergrund des Forschungsprojektes

Zumindest seit Beginn der siebziger Jahre besteht ein von seiten der empirischen Sozialforschung vielfach artikulierter Bedarf nach Übermittlung anonymisierter Mikrodaten[1)] der amtlichen Statistik (vgl. hierzu u.a. Brennecke/Schneider 1977; Kaase et al. 1980; Zapf 1985; Müller 1982; Müller/Hauser 1987; Krupp/Preißl 1989). Bislang waren diesem Anliegen der empirischen Sozialwissenschaften durch den Gesetzgeber allerdings enge Grenzen gesetzt. So wurde im Bundesstatistikgesetz von 1980 (Paragraph 11 Abs.5) zwar erstmals geregelt, daß anonymisierte Einzelangaben, sofern diese Angaben dem Befragten nicht mehr zuordenbar sind (absolute Anonymität), von den statistischen Ämtern übermittelt werden dürfen. Diese Vorgabe war in der Praxis jedoch mit erheblichen Problemen behaftet. Denn zum einen wurde von Experten schon zum damaligen Zeitpunkt darauf hingewiesen, daß die Deanonymisierung eines Einzeldatensatzes - selbst bei weitreichenden datenorientierten Schutzvorkehrungen - nie mit absoluter Sicherheit ausgeschlossen werden kann (Brennecke 1980; Schlörer 1980). Zum anderen lagen keine Erkenntnisse über das Ausmaß tatsächlicher Reidentifikationsrisiken vor (Scheuch 1980), die eine theoretisch und empirisch abgesicherte Risikoabschätzung bei Übermittlung von anonymisierten Individualangaben ermöglicht hätten (Hamacher 1980).

Die Beschlußempfehlung des Innenausschusses des Bundestages, nach welcher vor einer Übermittlung anonymisierter Einzelangaben lediglich sichergestellt werden sollte, daß eine potentielle Reidentifikation nach Kenntnissen der statistischen Ämter zweifelsfrei ausgeschlossen werden kann (Südfeld 1987), war angesichts dieser allgemeinen Unsicherheit wenig fruchtbar. Im Sinne einer höchstmöglichen Risikoausschließung verhielten sich die statistischen Ämter bei Anforderungen von Individualdaten mit umfangreichem Merkmalskatalog äußerst zurückhaltend, bzw. die Datenweitergabe war mit solch weitreichenden datenmodifizierenden Anonymisierungsmaßnahmen verbunden, daß die Nutzungsbedürfnisse der Wissenschaft nicht befriedigt und das wissenschaftliche Potential der Daten auch nicht annähernd ausgeschöpft werden konnte.

Die im Konzept der absoluten bzw. zweifelsfreien Anonymität implizit enthaltene Befürchtung einer mißbräuchlichen Verwendung von anonymisierten Individualdaten durch die Wissenschaft beruht nicht auf empirischen Erfahrungswerten (Scheuch 1980, 1987), sondern vielmehr auf einem allgemeinen Spannungsverhältnis zwischen Datenschutz und Forschungsfreiheit. Nach Kaase et al. (1980:283) ist diese Konfliktlinie dadurch gekennzeichnet, daß mit dem umfassenden Informationsbedarf der Forschung unter Umständen ein Eingriff in die persönliche Integrität von Individuen einhergehen kann, umgekehrt jedoch auch auf Datenschutzargumente zurückgegriffen wird, um unbequeme Forschungsvorhaben vom Datenzugang auszuschließen.

Diese im internationalen Vergleich restriktive Handhabung der Datenweitergabe zuungunsten der Wissenschaft wurde vom Gesetzgeber erst relativiert, als das Bundesverfassungsgericht im sogenannten Volkszählungsurteil den Nutzungsbedarf von amtlichen (anonymisierten) Mikrodaten durch die Wissenschaft nicht nur ausdrücklich anerkannte, sondern zugleich betonte, daß das Spannungsverhältnis zwischen Datenschutz und Forschungsinteressen nicht einseitig zuungunsten der Wissenschaft gelöst werden dürfe.

In der Novellierung des Bundesstatistikgesetzes (BStatG) von 1987 wurde deshalb eine - in Anlehnung an eine von seiten der Forschung schon sehr früh vorgetragene Forderung (Mohler/Kaase 1980:108) - spezifische Wissenschaftsklausel eingeführt (Paragraph 16 Abs.6, BStatG), die an den Begriff der faktischen Anonymität anknüpft, wie er bereits durch die European Science Foundation definiert wurde. Danach dürfen "für die Durchführung wissenschaftlicher Vorhaben (...) vom Statistischen Bundesamt und den statistischen Ämtern der Länder Einzelangaben an Hochschulen oder sonstige Einrichtungen mit der Aufgabe unabhängiger Forschung übermittelt werden, wenn die Einzelangaben nur mit einem

unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft zugeordnet werden können (...)" (Dorer/Mainusch/Tubies 1988:87).

Das Konzept der faktischen Anonymität ist ein wesentlicher Fortschritt gegenüber der alten Regelung. Denn zum einen ermöglicht die Beschränkung auf den Wissenschaftskontext eine Präzisierung der für einen Reidentifikationserfolg maßgeblichen Randbedingungen, was insbesondere bei der Berücksichtigung von Deanonymisierungsmotiven und Zusatzwissen von großer Bedeutung ist. Zum anderen wird das Mißtrauen gegenüber einer mißbräuchlichen Datenverwendung insofern relativiert, als man von einem rational kalkulierenden Datenangreifer ausgeht, der zwischen dem durch einen Reidentifikationsversuch erzielten Nutzen und den hierfür anfallenden Kosten abwägt.

Eine unmittelbare Umsetzung dieser neuen Regelung in die Weitergabepraxis war allerdings nicht möglich, weil bislang keine Kenntnisse zu den zentralen Problemen der faktischen Anonymität vorlagen. Die in der Vergangenheit durchgeführten Untersuchungen zu Reidentifikationsrisiken orientieren sich an der damaligen Gesetzeslage, bei welcher der Empfängerkreis von (absolut) anonymisierten Mikrodaten nicht näher definiert war. Da hierbei hypothetisch von einem allumfassenden Angriffsszenario und unbegrenzten Mitteln ausgegangen werden mußte, fehlen Untersuchungen, die Reidentifikationsrisiken unter dem Kosten-Nutzen-Aspekt analysieren ebenso wie Untersuchungen, die sich spezifisch auf die im Wissenschaftskontext vorhandenen Randbedingungen konzentrieren.

In Anknüpfung an eine von der amtlichen Statistik schon 1986 aufgegriffene Initiative[2] wurde daher in Kooperation zwischen dem Statistischen Bundesamt, der Universität Mannheim und ZUMA ein Forschungsprojekt[3] durchgeführt, um das Konzept der faktischen Anonymität zu operationalisieren. Ziel war, eine konsensfähige Lösung für die zukünftige Weitergabepraxis zu entwickeln.

Die zentralen Arbeiten des Anonymisierungsprojektes bezogen sich auf die Analyse des potentiellen Reidentifikationsrisikos von amtlichen Mikrodaten und auf die hierbei anfallenden Kosten. Hierfür wurden zunächst die Randbedingungen des Wissenschaftsszenarios untersucht, die für einen Reidentifikationsversuch von Bedeutung sein könnten. Auf dieser Grundlage wurden fünf Angriffsszenarien entwickelt, für die das Reidentifikationsrisiko sowie der damit verbundene Aufwand eingehend überprüft wurde.

Die allgemeinen Fragen zum Problembereich "faktische Anonymität" stellen sich natürlich für alle Arten von Mikrodaten. Aufgrund der vielfältigen

Formen von Mikrodaten ist es allerdings nicht möglich, allgemeine Antworten zu erwarten, die für alle Datenarten gleichermaßen zutreffen. Die Analysen beschränkten sich daher ausschließlich auf amtliche Mikrodaten und ihre Nutzung durch die Wissenschaft. Auch für Daten der amtlichen Statistik stellt sich das Problem unterschiedlich für verschiedene Datentypen. Betriebs- und Wirtschaftsdaten werfen andere Probleme auf als Personen- und Haushaltsdaten (Südfeld 1987; Krupp/Preißl 1989). Vor diesem Hintergrund wurde eine weitere Eingrenzung auf Personen- und Haushaltsdaten und hier wiederum ausschließlich auf den Mikrozensus und die Einkommens- und Verbrauchsstichprobe (EVS) vorgenommen.

In diesem Beitrag werden einige der zentralen Projektergebnisse aufgegriffen und die hieraus abgeleiteten Weitergabeempfehlungen dargestellt.

2. Voraussetzungen und Methoden einer Reidentifikation

2.1 Voraussetzungen einer Reidentifikation

Anonyme Daten unterscheiden sich von personenbezogenen Daten insofern, als sie keine direkten Identifikatoren, wie beispielsweise Namen und Anschriften, enthalten. Das Fehlen direkter Identifikatoren schließt allerdings nicht aus, daß für anonyme Daten nicht nachträglich ein Personenbezug rekonstruiert, d.h. eine Reidentifikation vorgenommen werden kann.[4]

Eine notwendige Voraussetzung hierbei ist, daß ein Angreifer über sogenanntes Zusatzwissen (auch als Identifikationsfile bezeichnet), d.h. personenbezogene Informationen verfügt, das zum einen gemeinsame Merkmale (Überschneidungsmerkmale) mit dem anonymen Mikrodatenfile aufweist und das sich zum anderen - zumindest in Teilen - auf die gleichen Personen wie das Mikrodatenfile bezieht. Ein Angreifer könnte dann versuchen, durch einen Abgleich der Überschneidungsmerkmale auf Identität oder sehr große Ähnlichkeit jene Datensätze im Mikrodaten- und Identifikationsfile zu ermitteln, die von ein und derselben Person stammen. Eine Reidentifikation wäre dann gegeben, wenn es anhand der Überschneidungsmerkmale möglich ist, für einen Datensatz des Mikrodatenfiles eine eins-zu-eins Relation zu einem Datensatz des Zusatzwissens herzustellen und wenn sichergestellt ist, daß sich diese Datensätze auf ein und dieselbe Person beziehen.

Das reale Reidentifikationsrisiko ist im wesentlichen durch drei Eigenschaften der Datenbasis beeinflußt (Paaß/Wauschkuhn 1985):

(1) Informationsgehalt der Überschneidungsmerkmale:
Eine Reidentifikation setzt die Einzigartigkeit von Ausprägungskombinationen voraus. Je höher die Anzahl der Überschneidungsmerkmale, der Differenzierungsgrad der Ausprägungen sowie ihrer Verteilungen ist, desto deutlicher sind die Datensätze im Merkmalsraum voneinander abgegrenzt und desto eher werden eins-zu-eins Zuordnungen möglich sein (für formale Abschätzungen des Informationsgehalts von Überschneidungsmerkmalen vgl. Müller et al. 1991:101f.)

(2) Stichprobeneigenschaften der Daten:
Eine Reidentifikation ist nur dann möglich, wenn eine Person in Mikrodatenfile und Zusatzwissen erfaßt ist. Die Stichprobeneigenschaft stellt daher eine prinzipielle Schranke für die Erfolgswahrscheinlichkeiten von Deanonymisierungsversuchen dar. Wird eine beliebige in der Grundgesamtheit enthaltene Person in einer amtlichen Stichprobe gesucht, entspricht die maximale Erfolgswahrscheinlichkeit dem Auswahlsatz dieser Stichprobe. Ist das Zusatzwissen ebenfalls nur eine Stichprobe - und beide Stichproben sind voneinander unabhängig - ergibt sich die Wahrscheinlichkeit, daß eine beliebige Person in beiden Datenfiles enthalten ist, durch die Multiplikation der Auswahlsätze.
Ein wesentlicher Unsicherheitsfaktor bei Reidentifikationsversuchen ist hierbei durch mögliche statistische Doppelgänger in der Grundgesamtheit gegeben, d.h. durch Personen, die identische Ausprägungskombinationen aufweisen. Ist eine Ausprägungskombination in der Grundgesamtheit mehrfach besetzt und stehen für einen Reidentifikationsversuch nur Stichproben zur Verfügung, ist eine Reidentifikation auch bei eins-zu-eins Zuordnungen nicht mehr mit Sicherheit, sondern nur noch mit einer gewissen Wahrscheinlichkeit möglich, da Verwechslungen mit statistischen Doppelgängern nicht ausgeschlossen werden können (für formale Modelle zur Abschätzung der Populationseinzigartigkeit von Ausprägungskombinationen siehe u.a. Marsh et al. 1991; Bethlehem et al. 1990).
Die von Stichproben ausgehende Schutzwirkung besteht dann nicht mehr, wenn ein Angreifer weiß, welche in seinem Zusatzwissen enthaltenen Personen an der Mikrodatenerhebung teilgenommen haben (Teilnahmekenntnis). Wird eine einzigartige Ausprägungskombination im Mikrodatenfile gefunden, die mit der Kombination des gesuchten Falls im Identifikationsfile übereinstimmt, dann kann - unter der Annahme, daß die Daten kompatibel abgebildet sind - ein Angreifer davon ausgehen, daß es sich um den gesuchten Fall handelt, unabhängig davon, ob in der Grundgesamtheit statistische Doppelgänger existieren oder nicht (vgl. Müller et al. 1991:95). Für eine Angriffssituation mit "Teilnahmekenntnis" ist daher ein wesentlich erhöhtes Reidentifikationsrisiko anzunehmen.

(3) Dateninkompatibilitäten zwischen Mikrodatenfile und Zusatzwissen:
Der oben skizzierte Reidentifikationsprozeß als einfacher Abgleich von Ausprägungskombinationen setzt voraus, daß die gesuchten Datensätze in Mikrodatenfile und Zusatzwissen identisch abgebildet sind (vgl. Block/Olsson 1976; Marsh et al. 1991). Diese Voraussetzung ist unter empirischen Bedingungen nicht immer erfüllt. So ist aus sozialwissenschaftlichen Untersuchungen bekannt, daß es bei jeder Datenerhebung in einem gewissen Grad zu Abweichungen von den "wahren" Werten kommt (Schnell/Hill/Esser 1988).[5] Als mögliche Ursachen sollen hier nur Antwortverzerrungen und einfache Aufbereitungsfehler genannt werden (für eine detaillierte Analyse siehe Müller et al. 1991:114ff.). Werden - wie bei einem Reidentifikationsversuch - Datenbestände aus unterschiedlichen Generierungsprozessen verglichen, können diese originären Datenfehler mit anderen Quellen möglicher Abweichungen kumulieren, welche beispielsweise bedingt sein können durch unterschiedliche Erhebungszeitpunkte, -ziele oder -kontexte oder unterschiedliche Meßinstrumente bei der Datenerhebung.

Auf individueller Ebene können solche Inkompatibilitäten dazu führen, daß Informationen, die sich auf ein und dieselbe Person beziehen, in unterschiedlichen Datenbeständen in abweichender Weise abgebildet sind. Eine Zuordnung auf Basis von Ausprägungsidentität wäre daher nicht möglich. Ebenso ist es vorstellbar, daß Datensätze, die ursprünglich nur sehr ähnliche Merkmalsausprägungen aufweisen, aufgrund von Inkompatibilitäten nun identisch abgebildet sind, wodurch eine weitere Quelle möglicher Verwechslungen bei Reidentifikationsversuchen entsteht. Unter der Annahme, daß bei einem realistischen Angriffsszenario sowohl die Stichprobeneigenschaften von Daten wie auch das Auftreten von Dateninkompatibilitäten eine sichere Reidentifikation wesentlich behindern können, stellt sich die Frage, inwieweit die einem Angreifer potentiell zur Verfügung stehenden Reidentifikationstechniken diese Unsicherheitsfaktoren berücksichtigen.

In der Literatur werden hauptsächlich zwei Reidentifikationstechniken diskutiert, die in bezug auf ihre Leistungsfähigkeit und den mit ihnen verbundenen Aufwand jeweils Extrempunkte einer Skala denkbarer Angriffstechniken darstellen. Da wir für die Operationalisierung der faktischen Anonymität auf diese Techniken zurückgegriffen haben, sollen sie im folgenden kurz skizziert werden.[6]

2.2 Reidentifikationstechniken

Am unteren Ende der Skala sind einfache Abgleichtechniken anzusiedeln, die in der Literatur als Hintertreppenidentifikation, Sortier- oder Selektionstechniken (Schlörer 1980; Block/Olsson 1976; Dittrich/Schlörer 1985) bezeichnet werden. Diesen Verfahren liegt ein Prinzip zugrunde, das in der empirischen Sozialforschung als sogenanntes Matching zur Anwendung kommt, wenn beispielsweise mehrere Dateien, die sich auf die gleiche Population beziehen, miteinander verknüpft werden sollen. Hierbei werden die Einzeldatensätze auf der Basis von Schlüsselvariablen in einer eins-zu-eins Relation zusammengefügt. Ein analoges Vorgehen ist auch bei einem Reidentifikationsversuch vorstellbar, wobei die Überschneidungsmerkmale als Schlüsselvariablen eingesetzt werden. Über Suchroutinen können dann jene Datensätze des Zusatzwissens ermittelt werden, die eine identische, einzigartige Ausprägungskombination zu Datensätzen des Mikrodatenfiles aufweisen. Der Vorteil einer solch einfachen Abgleichtechnik ist in der allgemeinen Verfügbarkeit (nahezu jedes Statistikprogramm bietet die Möglichkeit des Matching) und der wenig aufwendigen Umsetzung zu sehen. In einigen Untersuchungen wird diesen einfachen Abgleichtechniken daher direkt (Fischer-Hübner 1986; Brunnstein 1987) oder indirekt (Dalenius 1977, 1986, 1988; Bethlehem/Keller/Pannekoek 1990) ein hohes Gefährdungspotential zugeschrieben.

Eine Reidentifikationstechnik, die als Zuordnungskriterium nur die Ausprägungsidentität in Verbindung mit der Einzigartigkeit von Datensätzen voraussetzt, weist bezüglich der oben aufgezeigten Unsicherheitsfaktoren allerdings beträchtliche Defizite auf. Einerseits können Datensätze, die inkompatibel abgebildet sind, beim einfachen Ausprägungsabgleich nicht zugeordnet werden. Treten Dateninkompatibilitäten auf und ist ein gesuchter Fall hiervon betroffen, wird der Datensatz im Verlauf des Suchprozesses als nicht passend aussortiert. Andererseits ist es möglich, daß Datensätze, die bei korrekter Abbildung nur sehr ähnliche Werte aufweisen würden, aufgrund von Inkompatibilitäten identisch abgebildet sind. Hier würden mit einer einfachen Abgleichtechnik Falschzuordnungen erfolgen. Abweichungen vom "wahren" Wert sind nur dann unproblematisch, wenn sie übereinstimmen und damit die Abbildung wiederum kompatibel wäre. Schließlich kann es, wenn sowohl das Mikrodatenfile wie auch das Identifikationsfile nur als Stichproben zur Verfügung stehen, aufgrund von statistischen Doppelgängern zu Verwechslungen und damit zu Falschzuordnungen kommen. Das Ausmaß von Nicht- und Falschzuordnungen ist durch die einfache Abgleichtechnik nicht kontrollierbar.

Empirische Befunde, die konkrete Aussagen über das von einem einfachen Abgleich ausgehende Gefährdungspotential ermöglicht hätten, lagen bislang

allerdings nicht vor. Insofern das einfache Matching die untere Aufwandsgrenze bei einem Reidentifikationsversuch markiert und daher bei einem Angriffsszenario sehr wahrscheinlich an erster Stelle stehen würde, wurde diese Technik ungeachtet der offensichtlichen Defizite für die Operationalisierung der faktischen Anonymität berücksichtigt.

Komplexere Techniken versuchen, die aufgezeigten Probleme durch Fehlerabschätzungen und Wahrscheinlichkeitsberechnungen über das Auftreten von statistischen Doppelgängern zu lösen. Als das - auch international - leistungsfähigste Verfahren gilt das von der Gesellschaft für Mathematik und Datenverarbeitung im Rahmen des AIMIPH-Projektes entwickelte Verfahren zur Abschätzung von Reidentifikationsrisiken (Paaß/Wauschkuhn 1985). Dieser auf der Diskriminanzanalyse und Dichteschätzung beruhende Algorithmus ermittelt die Wahrscheinlichkeiten von spezifischen Ausprägungskombinationen und setzt sie zu der Wahrscheinlichkeit von Datenfehlern - über die bestimmte Annahmen getroffen werden - in Beziehung. Auf diese Weise kann für jeden Zieldatensatz des Zusatzwissens derjenige Datensatz des Mikrodatenfiles ermittelt werden, der in seiner Ausprägungskombination die höchste Übereinstimmung aufweist. Zugleich wird die Wahrscheinlichkeit berechnet, mit welcher diese Zuordnung korrekt ist. Diese ist um so höher, je weniger ähnliche Datensätze es im Mikrodatenfile gibt. Liegt die ermittelte Wahrscheinlichkeit über einer festzulegenden Sicherheitsschwelle (z.B. 0.9), so ist davon auszugehen, daß sich die zugeordneten Datensätze auf eine spezifische Person beziehen und damit eine Reidentifikation vorliegt. Im Gegensatz zu der einfachen Abgleichtechnik liefert diese Technik damit konkrete Entscheidungskriterien, wann eine Zuordnung von Datensätzen als korrekt angesehen werden kann und wann nicht.

Im Rahmen des AIMIPH-Projektes wurde die diskriminanzanalytische Methode einer empirischen Überprüfung unterzogen (Paaß/Wauschkuhn 1985). Hierfür standen ein Mikrodatenfile und ein hieraus erzeugtes - mit Fehlern überlagertes - Identifikationsfile zur Verfügung. Das Verfahren hat bei diesem Material unter spezifischen Randbedingungen vergleichsweise hohe Reidentifikationsquoten gezeigt. Aufgrund dieser Ergebnisse war davon auszugehen, daß von diesem Verfahren ein sehr hohes Gefährdungspotential ausgeht, weshalb auch diese Methode im Anonymisierungsprojekt einer empirischen Überprüfung - allerdings unter Verwendung von realen Mikrodaten und realem Zusatzwissen - unterzogen wurde (Bender 1990; Müller 1991; Bender/Blien/Müller 1990a,b).

3. Randbedingungen eines Deanonymisierungsversuchs im wissenschaftlichen Kontext: Deanonymisierungsmotive und Zusatzwissen

3.1 Deanonymisierungsmotive

Ob, abgesehen von der technischen Machbarkeit, überhaupt Deanonymisierungsversuche zu erwarten sind, hängt ab von dem Nutzen, den sich ein Angreifer von einer Reidentifikation verspricht. Um den potentiellen Nutzen deanonymisierter Daten im Wissenschaftskontext zu bestimmen, wurde geprüft, welche Logik der wissenschaftlichen Datennutzung zugrunde liegt und welche Motive sich hieraus für Deanonymisierungsversuche ergeben könnten (vgl. Müller et al. 1991:132ff.).
Diese Analyse führte zu dem Befund, der auch schon vom Bundesverfassungsgericht angeführt wurde: "(...) der Wissenschaftler ist regelmäßig nicht an der einzelnen Person interessiert, sondern an dem Individuum als Träger bestimmter Merkmale." (Neue Juristische Wochenschrift 1984:428, vgl. auch Scheuch 1980; Hamacher 1980; Zapf 1985). Plausible Motive sind aufgrund der beruflichen Interessenlage eines empirisch arbeitenden Sozialwissenschaftlers kaum überzeugend rekonstruierbar. Nach der vorliegenden Analyse erscheint es allenfalls in Grenzfällen vorstellbar, daß deanonymisierte Daten in einer Zwischenphase des Forschungsprozesses in Verbindung mit einer eigenen Erhebung von Nutzen sein könnten. Etwa um eine Auswahlbasis für eine eigene Stichprobenerhebung zu gewinnen oder um eine eigene Datenbasis mit den im Mikrodatenfile enthaltenen Informationen zu ergänzen. Diesen beruflich motivierten Angriffsszenarien ist gemeinsam, daß das Interesse nicht auf die Reidentifikation einiger weniger, sondern einer Vielzahl von Einzeldatensätzen gerichtet ist. Um Reidentifikationen in größerem Maßstab durchführen zu können, müßte - bedingt durch den relativ geringen Auswahlsatz der hier betrachteten Mikrodatenfiles - entweder eine sehr leistungsfähige Reidentifikationstechnik oder sehr umfangreiches Zusatzwissen zur Verfügung stehen.[7]

Es ist sicherlich diskussionswürdig, inwieweit im Zusammenhang einer Datenweitergabe für Forschungsvorhaben auch wissenschaftsfremde Deanonymisierungsmotive zu berücksichtigen sind. Will man jedoch hypothetisch mögliche - wenn auch sehr unwahrscheinliche - Fälle von Datenmißbrauch im Wissenschaftskontext berücksichtigen, so sind berufsfremde Motive ebenfalls in die Analyse einzubeziehen.

Im Unterschied zu beruflichen bietet sich bei außerberuflichen ein weites Spektrum hypothetisch denkbarer Motive an (Knoche 1989; Müller et al. 1991:151ff.). Diese reichen von persönlicher Neugier über ökonomisch

motivierte Deanonymisierungsversuche (beispielsweise der Verkauf von Information an Adressenhändler) bis hin zu eindeutig krimineller Motivation, bei welcher deanonymisierte Daten etwa für Erpressungsversuche herangezogen werden könnten. Da einzelne dieser Motive auf die Reidentifikation einer oder nur weniger Person(en) abzielen, könnte hier nicht nur der Aufwand für die Beschaffung von Zusatzwissen gering sein, sondern bereits eine einfache Abgleichtechnik die gewünschten Erfolge erbringen.

3.2 Zusatzwissen innerhalb des Wissenschaftsbereichs

Die Reidentifikation einer Person ist nur dann möglich, wenn ein Datenangreifer über entsprechendes Zusatzwissen verfügt. Allgemein läßt sich das erlangbare Zusatzwissen kaum abschließend abgrenzen (Burkert 1979, 1980). Die Weitergabebeschränkung faktisch anonymer Daten auf den wissenschaftlichen Kontext beinhaltet jedoch, daß die Risikoabschätzung nicht vor dem Hintergrund eines beliebig zur Verfügung stehenden Zusatzwissens erfolgen muß, sondern spezifisch auf den Datenempfängerkreis ausgerichtet sein kann.

Nach den im Anonymisierungsprojekt durchgeführten Analysen sind es im wesentlichen zwei Arten von Informationsquellen, die als Zusatzwissen genutzt werden könnten (vgl. Beckmann 1988):

Für den Bereich von öffentlich oder für einen beschränkten Personenkreis zugänglichen Registern oder privaten Datenquellen wurden insbesondere berufsgruppenspezifische Handbücher als möglicherweise riskant charakterisiert. In derartigen Handbüchern finden sich zwar in der Regel nur relativ wenige Überschneidungsmerkmale zu amtlichen Daten, es handelt sich jedoch zum Teil um detaillierte berufsbezogene Angaben in Verbindung mit Regionalangaben für teilweise deutlich von der Durchschnittsbevölkerung abgegrenzte Subpopulationen. Da in den meisten Handbüchern jeweils eine Vollerfassung der entsprechenden Berufsgruppen angestrebt wird, könnte die Suchrichtung bei einer Reidentifikation umgekehrt werden: Der Angreifer sucht sich eine dieser spezifischen Subpopulation angehörende Zielperson im Mikrodatenfile aus und versucht dieser den entsprechenden Datensatz im Zusatzwissen zuzuordnen (Strategie des Fischzugs). Je vollständiger die jeweilige Subpopulation im Zusatzwissen erfaßt und je deutlicher sie von durchschnittlichen Merkmalsträgern abgegrenzt ist, desto höher sind die Erfolgschancen einer solchen Fischzugsstrategie einzuschätzen. Es ist daher vorstellbar, daß unter Umständen ein erheblicher Teil der relevanten Mikrodatensätze aufgrund der in einem spezifischen Handbuch enthaltenen Informationen reidentifizierbar ist.

Als zweite wichtige Quelle von Zusatzwissen sind die im Kontext der Sozialwissenschaft professionsgemäß zur Verfügung stehenden Daten zu berücksichtigen, die im wesentlichen aus eigenen, freiwilligen Erhebungen stammen. Im Unterschied zu öffentlich zugänglichen Informationsquellen handelt es sich bei sozialwissenschaftlichen Erhebungen in der Regel um Stichproben. Der Stichprobenumfang ist meist relativ klein, kann aber in Einzelfällen auch einige tausend Personen umfassen. Für diesen Personenkreis wird zum Teil ein sehr umfangreicher Merkmalskatalog erfaßt. Die Stichprobeneigenschaft stellt zwar eine generelle Schranke für Reidentifikationsversuche dar, eine massenhafte Deanonymisierung kann nahezu ausgeschlossen werden. Darüber hinaus verfügen gerade bei umfangreichen Erhebungen die Wissenschaftler in der Regel nicht über die Adressen, da die Erhebungen durch professionelle Umfrageinstitute durchgeführt werden und die Daten ohne Adressen an die Wissenschaftler weitergegeben werden. Dennoch liegt die Annahme nahe, daß beim Vorliegen einer eigenen, personenbezogenen Erhebung der Informationsgehalt der Überschneidungsmerkmale so hoch sein könnte, daß dem Versuch, einzelne Personen zu reidentifizieren, gewisse Erfolgschancen zugebilligt werden könnten.

4. Empirische Überprüfung des Unverhältnismäßigkeitskriteriums

4.1 Untersuchungsanlage

Wie aus den bisherigen Ausführungen deutlich geworden sein sollte, ist die Wahrscheinlichkeit einer korrekten Zuordnung unter realen Bedingungen durch das Zusammenwirken einer Reihe unterschiedlicher Faktoren bestimmt. Hierzu zählen sowohl die Eigenschaften von Zusatzwissen und Mikrodatenfile (beispielsweise Umfang und Repräsentativität), der Informationsgehalt der Überschneidungsmerkmale, der Grad der Kompatibilität beziehungsweise Inkompatibilität unterschiedlicher Datenbestände wie auch die zur Verfügung stehenden Reidentifikationstechniken. Schon der Versuch, die Bedeutung der einzelnen Komponenten für das Reidentifikationsrisiko zu bestimmen, erweist sich als schwierig. Die Auswirkung der kombinierten Effekte eindeutig zu präzisieren, muß daher auch bei einer sehr sorgfältigen Analyse als in hohem Maße spekulativ erscheinen.

Die meisten Untersuchungen zur Bestimmung von Reidentifikationsrisiken unterstellen daher in aller Regel das reine Wirken einzelner Faktoren und grenzen über Modellannahmen störende Randfaktoren, wie beispielsweise Dateninkompatibilitäten, aus. So werden in einem Teil der Arbeiten Inkompatibilitäten überhaupt nicht erwähnt. Diesen Arbeiten liegt die

Annahme zugrunde, daß Datensätze mit übereinstimmenden und eindeutigen Ausprägungskombinationen, vor allem wenn sie sich auf die Grundgesamtheit beziehen, ohne Einschränkung reidentifizierbar sind (Fischer-Hübner 1986; Brunnstein 1987; Dalenius 1977, 1986, 1988). In anderen Studien werden Dateninkompatibilitäten als möglicher Störfaktor bei Reidentifikationsversuchen zwar explizit erwähnt. Bei der Abschätzung von Reidentifikationsrisiken wird dann jedoch vereinfachend unterstellt, daß die Daten vollständig kompatibel abgebildet sind (Spruill 1983; Dittrich/Schlörer 1985; Bethlehem et al. 1990). Nur wenige Arbeiten stellen Inkompatibilitäten für die Abschätzung von Reidentifikationsrisiken direkt in Rechnung (Paaß/Wauschkuhn 1985; Skinner et al. 1990; Marsh et al. 1991). Bei letzteren wird jedoch, wie in den anderen angeführten Untersuchungen auch, das für einen Reidentifikationsversuch notwendige Zusatzwissen als externe Randbedingung ausgeklammert.

Im Anonymisierungsprojekt wurde ein im Vergleich zu diesen Untersuchungen fast konträrer Weg gewählt, indem potentielle Reidentifikationsrisiken für reale Mikrodatenfiles unter Verwendung von realem Zusatzwissen ermittelt wurden. Entsprechend dem Konzept der faktischen Anonymität stand hierbei das erzielbare Endergebnis von realistischen Angriffsversuchen im Mittelpunkt des Interesses. Mögliche Störfaktoren wurden nicht ausgeklammert, sondern konnten in der Weise wirksam werden, wie dies bei einem realen Datenangriff gegeben wäre. Wenn es dabei auch nicht möglich war, einzelne Wirkungsfaktoren eindeutig zu isolieren, so konnte jedoch das Unverhältnismäßigkeitskriterium konkretisiert, das heißt der Aufwand realistischer Angriffsversuche ermittelt und mit dem Ertrag im Sinne erfolgreicher Deanonymisierungen verglichen werden.

Hierfür wurden im Hinblick auf die Gefährdungssituationen, die sich aus der Analyse des potentiellen Nutzens von deanonymisierten Daten und des im Wissenschaftskontext potentiell verfügbaren Zusatzwissens ergaben, fünf Angriffssituation spezifiziert (vgl. Müller et al. 1991:235ff.). Für zwei dieser Szenarien wurde das Reidentifikationsrisiko und der hierbei entstehende Aufwand empirisch überprüft. Drei Szenarien wurden unter Annahme plausibler Randbedingungen und den Ergebnissen der empirischen Überprüfung einer argumentativen Analyse unterzogen, bei welcher die Kosten eines Reidentifikationsversuchs den Kosten einer alternativen Beschaffung gleichwertiger Informationen gegenüber gestellt wurden (vgl. Helmcke 1989; Müller et al. 1991:351ff.).

Die Verwendung von realen Datenbeständen bedeutet nicht, daß gesetzeswidrig Deanonymisierungen versucht oder vorgenommen wurden.

Konkrete Aussagen über die Erfolgswahrscheinlichkeit von Reidentifikationsrisiken erfordern jedoch, daß das Verhältnis von korrekten und falschen Zuordnungen ebenso bekannt ist wie das Gesamtpotential der hypothetisch möglichen korrekten Zuordnungen. Zu diesem Zweck wurde ein aufwendiges Doppelt-Blind-Verfahren entwickelt, bei welchem ein Treuhänder zwischen die Adressenbesitzer einerseits und die mit der Durchführung der Experimente betrauten Forscher andererseits geschaltet wurde. Die Ergebnisse der mit anonymen Daten durchgeführten Experimente wurden an den Treuhänder weitergeleitet, der in datenschutzgerechter Weise überprüfte, welche der Zuordnungen korrekt, welche falsch und wie viele korrekte Zuordnungen maximal möglich gewesen wären (vgl. Müller et al. 1991:243f.).

Die empirische Überprüfung potentieller Reidentifikationsrisiken bei unterschiedlichen Angriffsvarianten und mit unterschiedlichen Reidentifikationstechniken erfolgte jeweils am Beispiel des Mikrozensus. Denn zum einen weist der Mikrozensus in der Regel mehr Überschneidungsmerkmale mit anderen Informationsquellen auf als die EVS (Müller et al. 1991:190-210), was auf ein höheres Gefährdungspotential hinweist. Zum anderen ist der Auswahlsatz des Mikrozensus um den Faktor fünf größer als der der EVS, das heißt die Wahrscheinlichkeit, daß eine beliebige Person sowohl im Zusatzwissen als auch im Mikrodatenfile erfaßt wurde, ist für den Mikrozensus größer als für die EVS.

Konkret wurde jeweils der Mikrozensus 1987 von Nordrhein-Westfalen genutzt. Dieser wurde dankenswerterweise vom Landesamt für Statistik und Datenverarbeitung Nordrhein-Westfalen mit den jeweils benötigten Überschneidungsmerkmalen und dem vollen Auswahlsatz (N=169.368) zur Verfügung gestellt. Die Daten waren anonym, das heißt sie enthielten keine personenbezogenen Angaben. Sonstige Anonymisierungsmaßnahmen wurden nicht vorgenommen (Müller et al. 1991:251f.).

Für ein erstes Angriffsszenario wurde - repräsentativ für öffentlich zugängliche Informationsquellen - Kürschners Deutscher Gelehrtenkalender als Zusatzwissen herangezogen. Alle in diesem Handbuch enthaltenen Informationen, die als Überschneidungsmerkmale zum Mikrozensus in Betracht kamen, wurden nach den Konventionen des Mikrozensus kodiert und auf Datenträger übernommen. Da der Gelehrtenkalender die Hochschullehrer nahezu vollständig erfaßt, konnte mit dieser Datenbasis überprüft werden, inwieweit der Versuch einer massenhaften Deanonymisierung realistisch ist.

Als zweite wichtige Quelle von Zusatzwissen wurde exemplarisch für sozialwissenschaftliche Datenbestände eine der umfangreichsten, in den alten Bundesländern durchgeführten, repräsentativen Erhebungen herangezogen.[8] Am Beispiel dieser Studie sollte zum einen das Gefährdungspotential eingegrenzt werden, daß sich aus dem Zugang zu umfangreichen, personenbezogenen Datenbeständen in den Sozialwissenschaften ergeben könnte. Zum anderen können die überwiegend erwerbs-, haushalts- und familienspezifischen Überschneidungsmerkmale zum Mikrozensus, als typische Beispiele für Alltags- beziehungsweise Anschauungswissen über andere Personen (Nachbarn, Bekannte oder Arbeitskollegen) gelten. Auf diese Weise konnte für etwa zweieinhalb Tausend unterschiedlich realistische Fälle geprüft werden, welches Risiko besteht, eine beliebige Person in einem Mikrodatenfile anhand von Alltagswissen zu deanonymisieren.

4.2 Empirische Überprüfung des Gelehrtenszenarios

4.2.1 Datenbasis

Kürschners Deutscher Gelehrtenkalender ist das umfassendste Verzeichnis von Forschern und Gelehrten im deutschsprachigen Raum ist und damit eine nach bisherigem Wissen sehr riskante Quelle von Zusatzwissen. Die Ausgabe von 1987 umfaßt circa 45.000 Gelehrte, bezogen auf Nordrhein-Westfalen sind etwa 8000 Fälle enthalten.

Der Gelehrtenkalender enthält zehn teilweise sehr differenzierte Überschneidungsmerkmale zum Mikrozensus (vgl. Übersicht 1) mit einem sehr hohen Informationsgehalt. Anhand der Merkmale "Beruf" und "Branche" ist es möglich, die Gelehrtenpopulation im Mikrozensus einzugrenzen. Zugleich ist diese Gelehrtenpopulation durch die Angaben "Fachzugehörigkeit" und "Geburtsjahr" stark differenziert. Mit den verfügbaren Regionalinformationen "Bundesraumordnungsregion" und "Gemeindegrößenklasse" ist es (in Verbindung mit dem Bundesland) möglich, im Mikrozensus Regionaleinheiten mit weniger als 200.000 Einwohnern in der Grundgesamtheit einzugrenzen (vgl. Müller et al. 1991:445f.). Es ist davon auszugehen, daß die im Gelehrtenkalender enthaltenen Angaben wenig fehlerbehaftet sind, da den Befragten die Möglichkeit geboten wird, die Angaben vor der Publizierung nochmals zu überprüfen. Um bei der Datenaufbereitung Quellen möglicher Inkompatibilitäten zu minimieren, wurde die Umsetzung der Klartextangaben des Gelehrtenkalenders in die Mikrozensus-Verkodung sowie die maschinelle Aufbereitung von Mitarbeitern des Statistischen Bundesamtes vorgenommen. Da diese Fachkräfte mit der Mikrozensusverkodung vertraut sind, ist von einer wesentlich höheren Vergleichbarkeit der

Angaben auszugehen, als dies der Fall wäre, wenn ein Angreifer eine solche Verschlüsselung vornimmt.

Hierbei zeigte sich, daß eine Verschlüsselung von Klartextangaben - auch wenn sie professionell erfolgt - durchaus problembehaftet sein kann. Nicht alle der im Gelehrtenkalender enthaltenen Angaben ließen sich eindeutig einer bestimmten Kategorie zuordnen. Ein solcher Fall war z.B. dann gegeben, wenn ein Hochschullehrer zugleich Arzt war, den Angaben aber nicht entnommen werden konnte, inwieweit beide Tätigkeiten gleichzeitig verfolgt werden, bzw. bei welcher Angabe es sich um die Haupttätigkeit handelt. Um das höchstmögliche Reidentifikationspotential auszuschöpfen, wurden in solchen Zweifelsfällen Alternativ-Verschlüsselungen vorgenommen. Insgesamt betrafen diese Alternativen die vier Überschneidungsmerkmale "Beruf", "ausgeübte Tätigkeit", "Wirtschaftszweig" und "Fachrichtung des Hochschulabschlusses" (vgl. Übersicht 1). Um diese Alternativvariablen zu berücksichtigen, wurden entsprechend der Variationsmöglichkeiten 2^4 unterschiedliche Identifikationsfiles erzeugt und bei der empirischen Überprüfung berücksichtigt. Für die konkrete Überprüfung des Szenarios wurde sowohl die einfache Abgleichtechnik wie auch die diskriminanzanalytische Methode von Paaß/Wauschkuhn herangezogen.

4.2.2 Ergebnisse

4.2.2.1 Einfache Abgleichtechnik

Das hohe Reidentifikationsrisiko der hier zur Verfügung stehenden Datenbasis kam insbesondere in der Zahl der einzigartigen Ausprägungskombinationen sowohl im Mikrodatenfile wie auch in den unterschiedlichen Identifikationsfiles zum Ausdruck. In Abhängigkeit der berücksichtigten Alternativvariablen, wiesen zwischen 45 und 65 Prozent der 7983 Datensätze im Gelehrtenkalender eine einzigartige Ausprägungskombination auf (vgl. Müller et al. 1991:274). Für die 169.368 Datensätze des Mikrozensus betrug die Einzelfallquote 38,7 Prozent. Berücksichtigt man nur die bei der Verwendung einer einfachen Abgleichtechnik relevante Zielpopulation, d.h. Fälle, deren Merkmalsausprägungen auch im Gelehrtenkalender enthalten sind, stehen im Mikrozensus noch 3099 Datensätze zur Verfügung. Hiervon weisen 79,6 Prozent eine einzigartige Ausprägungskombination auf.

In einer ersten Angriffssituation wurde unterstellt, daß ein Angreifer über einen Massenfischzug versucht, möglichst viele Fälle des Mikrozensus zu reidentifizieren. Hierbei könnte er den hohen Auswahlsatz des Gelehrtenkalenders nützen, indem er, ausgehend von der relevanten Subpopulation im Mikrozensus (n=3099) anstrebt, die zugehörigen Datensätze im Zusatzwissen zu ermitteln.

Übersicht 1: Überschneidungsmerkmale im Gelehrtenszenario

Gelehrtenkalender[9)] N=7983	Anzahl Merkmals-ausprägungen im Gelehrtenkalender
Gemeindegrößenklasse	9
Geschlecht	2
Geburtsjahr	61
Geburtsmonat	2
Wirtschaftszweig1	15
Wirtschaftszweig2	40
Stellung im Beruf	4
Beruf1	26
Beruf2	47
Ausgeübte Tätigkeit1	5
Ausgeübte Tätigkeit2	7
Fachricht. Hochschulabschluß1	71
Fachricht. Hochschulabschluß2	68
Bundesraumordnungsregion	16

Tabelle 1: Ergebnisse der experimentellen Überprüfung des Gelehrtenszenarios (Zuordnungskriterium: identische Ausprägungskombinationen)

	Teilnahmekenntnis: ja	nein
Anzahl der Fälle: Gelehrtenkalender*	7983	53
Mikrozensus	3099	151
Anteil Einzelfälle (gerundet)	80%	80%
eins-zu-eins Zuordnungen	14	9
hiervon korrekt	4	9
mehrdeutige Zuordnungen	15	2
(potentiell korrekt)	(6)	(1)
Wahrscheinlichkeit zu einem Fall des Gelehrtenkalenders den entsprechenden Datensatz im Mikrozensus zu lokalisieren	(4/7983) 0,0005	(9/53) 0,16

* Die Fälle des Gelehrtenkalenders wurden jeweils für beide Situationen in 16 unterschiedlichen Verkodungsvarianten überprüft.

In einer zweiten Angriffssituation wurde die Risikokonstellation insofern verschärft, als unterstellt wurde, daß der Angreifer über Teilnahmekenntnisse verfügt, also weiß, welche in seinem Zusatzwissen enthaltenen Personen an der Mikrozensuserhebung teilgenommen haben. Unter dieser zusätzlichen Randbedingung wäre es möglich, das Identifikationsfile auf jene 53 Fälle zu beschränken, die nach der Überprüfung des Treuhänders in beiden Datenbeständen erfaßt wurden. Analog wurde die Zielpopulation im Mikrozensus eingegrenzt, indem alle Fälle ausgeschlossen wurden, die von den 53 gesuchten Fällen des Gelehrtenkalenders abweichende Ausprägungen aufwiesen. Von den verbleibenden 151 Mikrozensusdatensätzen weisen 80 Prozent eine einzigartige Ausprägungskombination auf.

Wie Tabelle 1 zu entnehmen ist, variiert das Risiko hinsichtlich dieser Szenarien. Besteht keine Teilnahmekenntnis, können 14 der 7983 Datensätze des Gelehrtenkalenders in einer eins-zu-eins Relation 14 Datensätzen des Mikrozensus zugeordnet werden. Das heißt, diese Datensätze weisen im Mikrodatenfile und im Zusatzwissen eine je einzigartige und je identische Ausprägungskombinationen für die zehn Überschneidungsmerkmale auf. Weiteren 15 Datensätzen des Mikrozensus werden 29 Datensätzen des Gelehrtenkalenders in mehrdeutiger Weise zugeordnet.[10] Bei mehrdeutigen Zuordnungen kann ohne zusätzliche Informationen nicht entschieden werden, welche der zugeordneten Datensätze sich auf eine spezifische Person beziehen. Eine Reidentifikation wäre daher auf der Basis der im Gelehrtenkalender enthaltenen Informationen nicht möglich. Aber auch eine eindeutige Zuordnung ist nicht schon per se gleichbedeutend mit einer Reidentifikation. Nach der Überprüfung der Ergebnisse durch den Treuhänder beziehen sich lediglich vier der 14 eindeutigen Zuordnungen auch in der Realität auf ein und dieselbe Person. Die Wahrscheinlichkeit, für eine beliebige im Gelehrtenkalender enthaltene Person den entsprechenden Datensatz im Mikrozensus zu lokalisieren, ist mit 0,0005 daher äußerst gering. Zugleich liegt bei den wenigen zugeordneten Datensätzen die Wahrscheinlichkeit einer falschen Zuordnung mit 0,71 wesentlich höher als die einer korrekten Zuordnung.

Obwohl sich die absolute Zahl der potentiell korrekt zuordenbaren Datensätze nicht verändert, ist für die Angriffssituation "Teilnahmekenntnis" die Wahrscheinlichkeit einer Reidentifikation mit 0,17 höher als in der Situation "ohne Teilnahmekenntnis". Dies ist darauf zurückzuführen, daß sich ein Angreifer hier - wie oben ausgeführt - auf jene 53 Fälle im Zusatzwissen konzentrieren kann, für welche er sicher weiß, daß sie an der Mikrozensuserhebung teilgenommen haben. Auf diese Weise werden störende statistische Doppelgänger aus dem Zusatzwissen ausgegrenzt und ursprünglich mehrdeutige Zuordnungen liegen nun in eindeutiger Weise vor.

Hierdurch bedingt, ist es nun für neun der 53 gesuchten Fälle möglich, den entsprechenden Datensatz im Mikrodatenfile in eindeutiger Weise zu lokalisieren. Dennoch wird auch unter der Bedingung "Teilnahmekenntnis" die überaus wichtige Schutzfunktion von Dateninkompatibilitäten deutlich: Selbst wenn die mehrdeutigen Zuordnungen ebenfalls berücksichtigt werden, sind von den 53 in beiden Datenfiles erfaßten Fällen 81 Prozent so gelagert, daß sie in mindestens einem Überschneidungsmerkmal inkompatibel abgebildet sind. Diese Fälle sind daher mit einer einfachen Abgleichtechnik auch dann nicht zuordenbar, wenn beliebig viele zusätzliche Überschneidungsmerkmale zur Verfügung stehen würden (vgl. Müller et al. 1991:54).

Folgerungen für die faktische Anonymität:
Das Bundesstatistikgesetz definiert die faktische Anonymität anhand der Unverhältnismäßigkeit des Aufwandes an Zeit, Kosten und Arbeitskraft, der für die Herstellung eines Personenbezugs notwendig ist. Allerdings ist im Gesetzestext über die angegebene Formulierung hinaus nicht näher konkretisiert, im Vergleich zu welchem Gut der Aufwand abzuwägen ist. Dies ist nur implizit zu entnehmen. Im wesentlichen handelt es sich hierbei um das Verhältnis zwischen dem Wert beziehungsweise dem durch einen Reidentifikationsversuch erzielten Nutzen und den hierfür anfallenden Kosten (vgl. Helmcke 1989, Müller et al. 1991:212ff.). Dieser Nutzen läßt sich objektiv am ehesten im Vergleich zu den Kosten einer alternativen Informationsbeschaffung ermessen. Im folgenden wird deshalb das Unverhältnismäßigkeitskriterium durch einen Vergleich mit alternativen Informationsbeschaffungskosten bestimmt.

Legt man einen solchen Vergleich zugrunde, so war die faktische Anonymität unter der Randbedingung "Massenfischzug" zweifellos gegeben. Allein für die Aufbereitung des Gelehrtenkalenders fielen Kosten in Höhe von circa 44.000 Mark an. Für die Vorbereitung und Durchführung der Experimente einschließlich der Rechenzeit (ohne Testläufe) müssen circa 17.000 Mark veranschlagt werden. Der damit für die Ermittlung der 14 eindeutigen Zuordnungen angefallene Betrag von 61.000 Mark stellt eine untere Grenze dar, da weitere Kosten entstanden wären, um die vier korrekten Zuordnungen zu ermitteln. Im Vergleich dazu würde bei einer Alternativbeschaffung dieser Informationen - etwa durch ein professionelles Umfrageinstitut - im Schnitt zwischen 120 und 150 Mark pro Interview anfallen.[11] Selbst unter der Bedingung, daß die obere Grenze bei 150 Mark angesetzt wird, hätte das Reidentifikationsrisiko damit noch um ein Vielfaches höher liegen können, und die faktische Anonymität wäre dennoch gegeben gewesen.

Eine unmittelbare Ermittlung der unter der Annahme "Teilnahmekenntnis" anfallenden Kosten ist nicht möglich, da wir unser "Teilnehmerwissen" über den Treuhänder bezogen haben. Es ist zwar plausibel, daß ein Angreifer eventuell für einige wenige Personen weiß, daß sie an einer amtlichen Erhebung teilgenommen haben, aber sehr unwahrscheinlich, daß dieses Wissen beispielsweise für alle in einem Handbuch enthaltenen Fälle besteht. Wenn wir dennoch von dieser äußerst unrealistischen Annahme ausgehen und stark vereinfachend unterstellen, daß keine zusätzlichen Recherchekosten anfallen, kann der Betrag von 61.000 Mark proportional umgerechnet werden. Ausgehend von den 7983 Fällen des Gelehrtenkalenders hätte der Aufwand für die Überprüfung eines einzelnen Falles etwa 7,60 Mark betragen. Für die Überprüfung der 53 gesuchten Fälle und die Ermittlung der neun korrekten Zuordnungen wären dann Kosten in Höhe von etwas mehr als 400 Mark angefallen. Legt man wiederum die obigen Interviewkosten mit 150 Mark pro Interview zugrunde, hätte die Alternativbeschaffung der gewünschten Informationen etwa 1350 Mark gekostet, so daß das Unverhältnismäßigkeitskriterium nicht erfüllt gewesen wäre. Wie erwähnt, ist bei dieser Kalkulation vorausgesetzt, daß ein Angreifer - bezogen auf ganz Nordrhein-Westfalen - weiß, welche der im Gelehrtenkalender erfaßten Personen auch an der Mikrozensuserhebung teilgenommen haben und sich diese Informationen nicht erst mühsam beschaffen muß. Unabhängig hiervon zeigt sich jedoch, daß man bei einer deutlich von den durchschnittlichen Merkmalsträgern abgrenzbaren Subpopulation, wie beispielsweise den "Gelehrten", unter der Randbedingung "Teilnahmekenntnis" von einem erhöhten Gefährdungspotential ausgehen muß und dies bei entsprechenden Schutzvorkehrungen zu berücksichtigen ist.

4.1.2 Diskriminanzanalytische Methode nach Paaß/Wauschkuhn

Bei gleicher Datenkonstellation wurde im weiteren überprüft, ob und inwieweit die Nutzung einer Reidentifikationstechnik, welche Dateninkompatibilitäten und statistische Doppelgänger in der Grundgesamtheit berücksichtigt, die Erfolgschancen einer Reidentifikation erhöht (Blien/Müller 1991).

Die Verwendung dieser Methode setzt die Spezifizierung eines Fehlerprozesses und damit Informationen über Strukturen und Ausmaß möglicher Dateninkompatibilitäten voraus. Derartige Kenntnisse liegen in den Sozialwissenschaften bislang allerdings nur bruchstückhaft vor (vgl. u.a. Esser 1986; Schnell/Hill/Esser 1988). Es gibt Untersuchungen zu Veränderungsraten im Zeitablauf, der Reliabilität von Erhebungsinstrumenten und der Häufigkeit von Erhebungsfehlern (Koch 1986; Porst/Zeifang 1987; Schwarz/Hippler/Strack 1988). Jüngere Untersuchungen beschäftigen sich auch mit der Wirkung unterschiedlicher Erhebungskontexte auf das

Antwortverhalten (Schwarz/Hippler/Noelle-Neumann 1989). Eine abgeschlossene Theorie liegt bislang nicht vor.

Die sich hierdurch ergebenden Probleme bei der Festlegung von Fehlerprozessen begründen zwar generelle Zweifel an der Einsatzfähigkeit dieses Verfahrens. Ein zentrales Ergebnis der AIMIPH-Studie bestand jedoch darin, daß das Verfahren relativ robust auf Fehlspezifikationen (in bezug auf die Höhe der Fehler) reagiert (Paaß/Wauschkuhn 1985:186; Paaß 1987). Daher wurde analog zum Vorgehen von Paaß/Wauschkuhn ein Fehlerprozeß spezifiziert (Müller 1991). Um den Problemen einer adäquaten Fehlerschätzung gerecht zu werden, wurden (in Höhe und Struktur) verschiedene Fehlerprozesse spezifiziert, wobei insgesamt fünf unterschiedliche Datenkonstellationen geprüft wurden (vgl. Müller et al. 1991:283).

Bei einer Sicherheitsschwelle von 99 Prozent wurden über alle fünf Situationen insgesamt 29 unterschiedliche Zuordnungen ermittelt. Die Überprüfung der Ergebnisse durch den Treuhänder ist allerdings desillusionierend: von den 29 Zuordnungen bezogen sich nur drei auch in der Realität auf ein und dieselbe Person. Alle drei Zuordnungen waren bereits mit der einfachen Abgleichtechnik ermittelt worden, weil die betreffenden Ausprägungskombinationen kompatibel abgebildet waren. Was sich zunächst als Vorteil dieser Methode darstellt, nämlich die höhere Zahl von Zuordnungen, erweist sich bei näherer Betrachtung insofern als Nachteil, als lediglich der Anteil der Falschzuordnungen steigt. Dieser liegt mit etwa 90 Prozent sogar noch höher als bei einem einfachen Abgleich der Ausprägungskombinationen. Die ausgewiesen hohe Sicherheitsschwelle von 99 Prozent trifft sowohl auf falsche wie auf korrekte Zuordnungen zu. Analog zu einer einfachen Abgleichtechnik ist es daher nicht möglich, zwischen korrekten und falschen Zuordnungen zu unterscheiden. [12)]

Folgerungen für die faktische Anonymität:
Setzt man dieses Ergebnis in Relation zum erbrachten Aufwand, war das Unverhältnismäßigkeitskriterium bei Verwendung der diskriminanzanalytischen Methode bereits durch die Vorbereitungsarbeiten erfüllt. Da das Verfahren nicht standardmäßig zur Verfügung steht und aus plausiblen Gründen auch nur spärlich dokumentiert ist, nahm die Rekonstruktion des Algorithmus, die Programmanpassung an die Datenstruktur und die Implementation auf den vorhandenen Rechner etwa ein Jahr Arbeitszeit in Anspruch. Erst dann war die Generierung der Fehlerverteilung und die Durchführung der Experimente möglich. Hinzu kommen hohe Rechenkosten, da der Algorithmus sehr viel CPU-Zeit und Speicherplatz beansprucht. Die Gesamtkosten beliefen sich auf etwa 261.000 Mark (vgl. Müller et al. 1991:311). Der Einsatz einer solch aufwendigen Methode innerhalb des

Wissenschaftskontextes kann damit ebenso wie die eigene Entwicklung eines äquivalenten Algorithmus mit an Sicherheit grenzender Wahrscheinlichkeit für ein Datenangriffsszenario ausgeschlossen werden. Auf eine weitere Überprüfung dieser Reidentifikationstechnik wurde daher verzichtet.

4.3 Empirische Überprüfung des sozialwissenschaftlichen Szenarios

4.3.1 Datenbasis

Die spezifische Risikokonstellation des Gelehrtenszenarios ergab sich aus dem klaren Bezug auf eine deutlich abgrenzbare Subpopulation und deren nahezu vollständiger Erfassung. Im Gegensatz hierzu weist die sozialwissenschaftliche Erhebung einen relativ kleinen Auswahlsatz auf, der einen repräsentativen Querschnitt der Bevölkerung widerspiegelt. Für die hier erfaßten Personen steht mit 35 Überschneidungsmerkmalen zum Mikrozensus allerdings ein sehr umfangreiches Zusatzwissen zur Verfügung. Von besonderem Interesse sind die zahlreichen Merkmale zum Haushaltskontext der Befragten. In verschiedenen Untersuchungen wird darauf hingewiesen, daß schon einige wenige Haushaltsmerkmale zu einzigartigen Ausprägungskombinationen in einem Datenfile führen können (Brunnstein 1987; Fischer-Hübner 1986; Greenberg 1990). Da davon auszugehen ist, daß gerade im Alltagswissen Informationen über Haushalte von Dritten verfügbar beziehungsweise relativ einfach beschaffbar sind, wird daher ein erhöhtes Reidentifikationsrisiko unterstellt, wenn ein Mikrodatenfile detaillierte Angaben über den Haushaltskontext enthält.

Um das von Haushaltskontextmerkmalen ausgehende Gefährdungspotential empirisch zu präzisieren, wurden die Reidentifikationsexperimente in drei Phasen durchgeführt. In einer ersten Phase wurden nur Merkmale berücksichtigt, die sich auf eine spezifische Person im Haushalt beziehen. In einer zweiten Phase wurde zusätzlich der allgemeine Haushaltskontext einbezogen. In einer dritten Phase schließlich wurden detaillierte Haushaltsinformationen auch über die im Haushalt lebenden Partner berücksichtigt. Eine detaillierte Auflistung der jeweils berücksichtigten Überschneidungsmerkmale gibt Übersicht 2.

4.3.2 Ergebnisse

Analog zum Gelehrtenszenario wird zwischen den Angriffssituationen "keine Teilnahmekenntnis" und "Teilnahmekenntnis" unterschieden. Im folgenden werden zunächst die Ergebnisse der Angriffssituation "keine Teilnahmekenntnis" dargestellt.

Übersicht 2: Sozialwissenschaftliches Szenario: Überschneidungsmerkmale in Zuordnungsphase 1 bis 3

Überschneidungsmerkmale	Anzahl Auspräg.*	Zuordnungsphase: 1	2	3
Personenspezifische Merkmale				
Geschlecht	2	x	x	x
Geburtsjahr	38	x	x	x
Familienstand	4	x	x	x
Schulische Ausbildung	6	x	x	x
Berufsausbildung	8	x	x	x
Stellung im Beruf	11	x	x	x
Erwerbstätigkeit	3	x	x	x
Arbeitslosigkeit	3	x	x	x
wöchentliche Arbeitszeit	7	x	x	x
Arbeitssuche	3	x	x	x
Art des Arbeitsvertrags	4	x	x	x
Ende d. Erwerbstätigkeit	8	x	x	x
persönliches Nettoeinkommen	9	x	x	x
Allgemeine Haushaltsinformationen				
Zahl Kinder im Hhlt unt. 3 Jahren	3		x	
dto von 3 bis unter 6 Jahren	3		x	
dto von 6 bis unter 10 Jahren	3		x	
dto von 10 bis unter 15 Jahren	3		x	
dto von 15 bis unter 18 Jahren	3		x	
dto von 18 bis unter 28 Jahren	5		x	
dto über 28 Jahre	2		x	
Haushaltsnettoeinkommen	9		x	x
Detaillierte Haushaltsinformationen				
Partner: Stellung i. Beruf	11			x
-"- Arbeitszeit (i.Stunden)	8			x
-"- Erwerbstätigkeit	3			x
-"- Arbeitslos	3			x
-"- Arbeitssuche	3			x
Geburtsjahr Kind1	33			x
Geburtsjahr Kind2	27			x
Geburtsjahr Kind3	25			x
Geburtsjahr Kind4	18			x
Geburtsjahr Kind5	10			x
Geschlecht Kind1	2			x
Geschlecht Kind2	2			x
Geschlecht Kind3	2			x
Geschlecht Kind4	2			x
Geschlecht Kind5	2			x
Ausbildung Kind1	4			x
Ausbildung Kind2	4			x
Ausbildung Kind3	4			x
Ausbildung Kind4	4			x
Ausbildung Kind5	2			x
Vorwieg. Unterhalt des Haushalts	7			x
Gesamtzahl Überschneidungsmerkmale		3	21	35

* Diese Angaben beziehen sich auf die sozialwissenschaftliche Erhebung, da nur die dort auftretenden Merkmalsausprägungen für die Zuordnungen mit einer einfachen Reidentifikationsmethode relevant sind.

Tabelle 2: Ergebnisse der experimentellen Überprüfung des sozialwissenschaftlichen Szenarios unter der Annahme, daß *keine Teilnahmekenntnis* vorliegt (Zuordnungskriterium: identische Ausprägungskombinationen)

Phase	I	II	III
Zahl der Überschneidungsmerkmale	13	21	35
Anzahl der Fälle:			
Sowi. Stichprobe	2685	2685	2685
Mikrozensus	94.747	94.747	53.441
Anteil Einzelfälle (gerundet)	20%	79%	84%
Zuordnungen (insgesamt)	1107	298	74
Wahrscheinlichkeit zu einem Fall der sowi. Erhebung den entsprechenden Datensatz im Mikrozensus zu lokalisieren	0	0	0

Tabelle 3: Ergebnisse der experimentellen Überprüfung des sozialwissenschaftlichen Szenarios unter der Annahme, daß *Teilnahmekenntnis* vorliegt (Zuordnungskriterium: identische Ausprägungskombinationen)

Phase	I	II	III
Sowi. Erheb. Fall Nr:	Anzahl der berücksichtigten Überschneidungsmerkmale (in Klammern)	13	21
1)	1 (12)	0	0
2)	55 (11)	2	0
3)	22 (10)	0	0
4)	137 (11)	0	0
5)	53 (12)	12	0
6)	235 (12)	159	1
7)	192 (9)	0	0
8)	25 (9)	0	0
9)	588 (9)	0	0
10)	27 (12)	5	0

Gemessen am Anteil der Datensätze mit einzigartigen Ausprägungskombinationen ist den Haushaltsinformationen ein hohes Risikopotential zuzuschreiben. Stehen nur die personenspezifischen Überschneidungsmerkmale zur Verfügung, weist das Mikrodatenfile eine Einzelfallquote von lediglich 19,6 Prozent auf. Werden alle 35 Merkmale berücksichtigt, sind 84 Prozent der Personen in dem Mikrodatenfile durch eine einzigartige Ausprägungskombination gekennzeichnet.

Wie aus Tabelle 2 hervorgeht, spiegelt sich die mit den haushaltsbezogenen Merkmalen einhergehende Trennschärfe auch in den Zuordnungsquoten wider: Von den 1107 aufgrund von identischen Wertekombinationen ermittelten ein- bzw. mehrdeutigen Zuordnungen in Phase I verbleiben 298, wenn die allgemeinen Haushaltsinformationen berücksichtigt werden. Werden alle 35 Überschneidungsmerkmale als Zusatzwissen eingesetzt, reduziert sich die Zahl der Zuordnungen auf 74. Hiervon sind 35 eindeutig.

Wie Tabelle 2 entnommen werden kann, ist die Einzelfallquote in den Daten auch in diesem Szenario kein Indikator für ein real bestehendes Reidentifikationsrisiko. Alle hier auf der Basis identischer Ausprägungskombinationen ermittelten Zuordnungen waren falsch: für keinen der zehn Fälle, die - nach der Überprüfung des Treuhänders - in beiden Erhebungen erfaßt wurden, war es möglich, die entsprechenden Partnerdatensätze aus Mikrozensus und sozialwissenschaftlicher Erhebung korrekt zuzuordnen.

Noch stärker als im Gelehrtenszenario wirken damit Dateninkompatibilitäten und statistische Doppelgänger als Schutz vor erfolgreichen Reidentifikationsversuchen: Durch einen Vergleich der zehn Partnerdatensätze im Mikrozensus und in der sozialwissenschaftlichen Erhebung konnte gezeigt werden, daß schon in Phase I lediglich fünf von dreizehn Merkmalen für alle zehn Datensätze kompatibel abgebildet waren. Bei Berücksichtigung nur dieser fünf Merkmale wies jedoch keiner der zehn gesuchten Datensätze eine eindeutige Ausprägungskombination auf (Müller et al. 1991:335). Es müßten daher weitere Überschneidungsmerkmale berücksichtigt werden. Alle weiteren Merkmale sind jedoch mit einer gewissen Wahrscheinlichkeit von Inkompatibilitäten betroffen, die eine korrekte Zuordnung verhindern.

Das Dilemma zwischen der notwendigen Eindeutigkeit eines Datensatzes einerseits und der mit jedem zusätzlich berücksichtigten Überschneidungsmerkmal steigenden Wahrscheinlichkeit von Inkompatibilitäten andererseits wird unter der Randbedingung "Teilnahmekenntnis" noch deutlicher. Hier wurde - bezogen auf die personenspezifischen Merkmale - als zusätzliche Annahme unterstellt, daß der Angreifer aufgrund von Plausibilitätsüberlegungen weiß, welche dieser Merkmale kompatibel abgebildet sind. Er könnte

sich bei einem Reidentifikationsversuch in einer ersten Phase daher auf jene Datensätze im Mikrozensus konzentrieren, die für diese Merkmale jeweils kompatibel abgebildet sind. Das Ergebnis dieser Suche ist Tabelle 3/Phase I zu entnehmen.

Wie aus Tabelle 3 hervorgeht, ist es selbst unter dieser äußerst riskanten und höchst unwahrscheinlichen Randbedingung nur für einen Datensatz (Phase I/Fall Nr.1) möglich, eine eindeutige (und korrekte) Zuordnung vorzunehmen. In allen anderen Fällen werden weitere Überschneidungsmerkmale benötigt, um statistische Doppelgänger auszuschließen. In der nächsten Phase (II) sind fünf Fälle aufgrund von Inkompatibilitäten nicht mehr zuordenbar. Bei vier Fällen scheint eine weitere Eingrenzung möglich. Die tatsächlich gesuchten Datensätze sind in diesen Merkmalen jedoch inkompatibel abgebildet und deshalb in der eingekreisten Subpopulation nicht mehr enthalten. Bedingt durch die Inkompatibilitäten würde der Suchprozeß in eine völlig falsche Richtung laufen. Noch deutlicher tritt dieser Sachverhalt in Phase III zutage. Hier ist zwar eine eins-zu-eins Zuordnung möglich, diese ist jedoch falsch. Dieser Sachverhalt würde einem Angreifer verborgen bleiben. Da er aufgrund seiner Teilnahmekenntnis sicher ist, daß die von ihm gesuchte Person im Mikrodatenfile enthalten sein muß, besteht für ihn kein Grund, an der Qualität dieser Zuordnung zu zweifeln: Solange sich für den gesuchten Fall zumindest ein identischer Datensatz in einem Mikrodatenfile ermitteln läßt, enthalten die Daten kein Indiz in bezug auf mögliche Verwechslungen aufgrund von Inkompatibilitäten.

Die im sozialwissenschaftlichen Szenario ermittelten Resultate ergänzen die Befunde aus dem Gelehrtenszenario insofern, als sich zeigt, daß es auch unter der Annahme von Teilnahmekenntnis nahezu unmöglich ist, eine beliebige Person anhand einiger weniger Merkmale zu reidentifizieren. Sofern eine gesuchte Person nicht eine in der Grundgesamtheit äußerst selten vertretene Ausprägungskombination in diesen Merkmalen aufweist, werden sich in der Grundgesamtheit und daher auch im Mikrodatenfile eine Vielzahl von statistischen Doppelgängern finden, die einen Reidentifikationsversuch nachhaltig stören. Das notwendige Zusatzwissen wirkt damit in paradoxer Weise auf den Deanonymisierungsprozeß: Je mehr Überschneidungsmerkmale im Zusatzwissen enthalten sind, desto höher ist der Anteil der einzigartigen Ausprägungskombinationen. Mit jedem zusätzlichen Merkmal erhöht sich jedoch zugleich auch die Wahrscheinlichkeit, daß ein gesuchter Fall in diesem Merkmal inkompatibel abgebildet ist, woraus Falsch- bzw. Nichtzuordnungen resultieren können.

Folgerungen für die faktische Anonymität:
Obwohl in dem hier untersuchten Szenario auf bereits maschinenlesbare Datenfiles zurückgegriffen werden konnte, waren umfangreiche Vorarbeiten notwendig, um die in der sozialwissenschaftlichen Erhebung enthaltenen Informationen als Überschneidungsmerkmale zum Mikrozensus nutzen zu können (vgl. Müller et al. 1991:259ff.). Hierbei entstanden für die Entwicklung des Konzepts, die Anpassung (Rekodierung) der Überschneidungsmerkmale und die Angleichung der Datenstruktur von Zusatzwissen und Mikrodatenfile, Arbeitskosten in Höhe von circa 15.000 Mark. Aufgrund der umfangreichen Vorarbeiten einschließlich der Durchführung der Zuordnungsexperimente entstanden zusätzlich Rechenkosten in Höhe von etwa 13.000 Mark. Der gesamte finanzielle Aufwand betrug circa 28.000 Mark. Angesichts des vorliegenden Ergebnisses, nach welchem selbst unter der Randbedingung "Teilnahmekenntnis" und nur unter äußerst extremen Zusatzannahmen, maximal eine korrekte Zuordnung möglich gewesen wäre, wäre eine Alternativbeschaffung der gewünschten Informationen in diesem Szenario in jedem Fall die kostengünstigere Alternative gewesen.

5. Zusammenfassende Bewertung der empirischen Überprüfung von Reidentifikationsrisiken

Als wichtigstes Ergebnis der empirischen Überprüfung ist festzuhalten, daß das Reidentifikationsrisiko bei der Verwendung von realen Daten weitaus niedriger ist, als dies bislang aufgrund wahrscheinlichkeitstheoretischer Berechnungen und mit ganz oder teilweise synthetisch generierten Daten durchgeführter Experimente zu vermuten war.

Eine wesentliche Erkenntnis war, daß das Vorhandensein einmaliger Ausprägungskombinationen keineswegs schon eine hinreichende Bedingung für eine Reidentifikation bedeutet. Bei den Experimenten wiesen sowohl im Mikrodatenfile wie im Identifikationsfile die überwiegende Zahl der relevanten Fälle einmalige Ausprägungskombinationen auf. Obwohl das Identifikationsfile im Gelehrtenszenario einer Vollerhebung nahekam, war die Zahl der mit Sicherheit richtig vorgenommenen Reidentifikationen äußerst gering. Der wichtigste Grund dafür ist die praktische Unvermeidbarkeit von Inkompatibilitäten zwischen unterschiedlichen Datenbeständen, die in ihrer Wirkung auf Reidentifikationsrisiken bislang unterschätzt wurden.

Gegen das Ergebnis einer "natürlichen" Schutzfunktion von Inkompatibilitäten könnte argumentiert werden, daß es im wesentlichen auf einer empirischen Illustration durch einige wenige Fälle beruht. Dem ist

entgegenzuhalten, daß sich auch in anderen Untersuchungen Hinweise auf Dateninkompatibilitäten finden. So berichten Marsh et al. (1991) von Abweichungen, die im Rahmen einer Nacherhebung zum britischen Zensus 1981 festgestellt wurden. Sie betrugen beispielsweise für das Merkmal Haushaltsgröße 2,4 Prozent, für Haus- und Wohnungseigentum 3,2 Prozent, für Erwerbstätigkeit 7,8 Prozent und für die Zugehörigkeit zu grob definierten Berufsklassen (sechs Kategorien) 13 Prozent. Ähnliche Befunde fanden sich auch bei einer Reanalyse der ALLBUS Test-Retest Studie. In dieser von ZUMA 1984 durchgeführten Studie wurden einer Zufallsauswahl der ALLBUS-Stichprobe im Monatsabstand dreimal dieselben Fragebögen vorgelegt, um die Reliabilität von Umfragedaten zu untersuchen (Porst/Zeifang 1987). Obwohl insbesondere die soziodemographischen Merkmale eine relativ hohe Stabilität aufweisen, unterscheiden sich zwischen Welle 1 und Welle 3 (vgl. Müller et al. 1991:124):

- 45 Prozent der Befragten bei den Angaben zur Einkommenshöhe,
- 18 Prozent hinsichtlich der geleisteten Arbeitswochenstunden,
- 16 Prozent bei den angegebenen beruflichen Ausbildungsabschlüssen,
- 13 Prozent hinsichtlich ihrer überwiegenden Einkünfte.

Auf individueller Ebene zeigen sich von Welle 1 zu Welle 2 sowie von Welle 1 zu Welle 3 ähnliche Abweichungen wie im sozialwissenschaftlichen Szenario. Bei elf berücksichtigten Merkmalen machen von Welle 1 zu Welle 2 lediglich 22 Prozent der Befragten identische Angaben. Knapp 39 Prozent weichen in einem Merkmal und weitere 39 Prozent in mindestens zwei Merkmalen ab (vgl. Übersicht 3).

Übersicht 3: ALLBUS Test-Retest-Studie 1984: Fallspezifische Häufigkeiten von aufgetretenen Inkompatibilitäten bei elf berücksichtigten Merkmalen[13] zwischen Welle 1 und Welle 2 sowie Welle 1 und Welle 3

Von 11 Merkmalen waren inkompatibel	Anteil der betroffenen Fälle von: Welle 1 zu Welle 2 (in %)	Welle 1 zu Welle 3 (in %)
0	22.1	20.8
1	38.7	33.1
2	28.2	31.2
3	8.8	14.3
4	2.2	.6
N	181	154

Quelle: Eigene Berechungen auf Basis der ALLBUS Test-Retest-Studie 1984

Als Maßnahme zum Schutz gegen Deanonymisierung ist deshalb nicht vorrangig auf die Verhinderung einmaliger Ausprägungskombinationen abzustellen. Es ist eher darauf zu achten, daß keine Merkmalsausprägungen ausgewiesen werden, die so selten sind, daß durch sie allein einzelne Personen leicht identifiziert werden könnten (vgl. hierzu auch Brennecke 1980:163).

Während eine massenhafte Deanonymisierung von Datensätzen nahezu ausgeschlossen ist und auch der Versuch, beliebige Personen zu reidentifizieren, in aller Regel scheitern wird, sind der empirischen Analyse jedoch auch Anhaltspunkte für eine Risikokonstellation zu entnehmen, bei welcher eine erfolgreiche Reidentifikation nicht ausgeschlossen werden kann. Dieser - allerdings äußerst unwahrscheinliche - Fall setzt das Zusammentreffen sehr spezifischer Risikofaktoren voraus und kann wie folgt charakterisiert werden: [14)]

1) Eine im Mikrodatenfile gesuchte Person gehört einer sehr kleinen, durch ein spezielles Merkmal identifizierbaren Subpopulation an (sachliche Tiefengliederung);
2) der Mikrodatenfile enthält stark differenzierte Regionalinformationen, so daß in den Regionaleinheiten nur wenige Personen der spezifischen Subpopulation leben (regionale Tiefengliederung);
3) der Datenangreifer weiß, daß die gesuchte Person im Mikrodatenfile enthalten ist (Teilnahmekenntnis);
4) die Merkmale der Person sind genau in der Weise im Mikrodatenfile erfaßt, wie es der Forscher vermutet (Kompatibilität).

Beim Zusammentreffen dieser vier Bedingungen erscheint die Reidentifikation von einzelnen Fällen ohne großen Aufwand als möglich. Bereits wenn eine der Bedingungen nicht gegeben ist, ist die Wahrscheinlichkeit einer Reidentifikation nach den durchgeführten Experimenten als äußerst gering einzustufen. Das gleichzeitige Zusammentreffen aller Bedingungen kann bei Stichprobenerhebungen zwar als außergewöhnlich seltenes Ereignis betrachtet werden. Dennoch sollten bei der Datenübermittlung Vorkehrungen getroffen werden, damit auch eine solche Risikokonstellation ausgeschlossen ist.

Neben obligatorischen vertraglichen Verpflichtungen, die beispielsweise eine Reidentifikation verbieten (vgl. Knoche 1991), sowie technisch-organisatorischen Sicherungsmaßnahmen, die insbesondere der Datenzugriffskontrolle dienen (vgl. Blien 1990), müssen auch datenorientierte Schutzmaßnahmen getroffen werden. In einem weiteren Schritt wurde daher die Wirkung ausgewählter Anonymisierungsmaßnahmen am Beispiel des

Gelehrtenszenarios überprüft (vgl. Müller et al. 1991:386ff.). Die empirische Überprüfung erfolgte im Hinblick auf die oben angeführten Bedingungen eins bis drei).

Zur Verringerung des Reidentifikationsrisikos aufgrund der Zugehörigkeit zu einer kleinen spezifischen Subpopulation oder der Existenz tiefgegliederter Regionalinformationen auf seiten des Angreifers, wurde die von Ausprägungsvergröberungen ausgehende Schutzwirkung geprüft.

Gegen das spezifische Gefährdungspotential der Teilnahmekenntnis wurde die Substichprobenziehung geprüft. Mit fallender Substichprobengröße verringert sich die Wahrscheinlichkeit, daß eine beliebige Person in den übermittelten Mikrodaten enthalten ist. Hierdurch erhöht sich in einem informationstheoretischen Sinn die Unsicherheit eines Angreifers über die Korrektheit möglicher Zuordnungen beträchtlich. Selbst wenn ein Angreifer weiß, daß eine spezifische Person an der Erhebung teilgenommen hat, kann er - bei der Übermittlung von Substichproben - auch bei einer eins-zu-eins Zuordnung nicht mehr sicher sein, ob es sich hierbei um die gesuchte Person oder einen statistischen Doppelgänger handelt.

Im folgenden Abschnitt werden die hieraus für die Wahrung der faktischen Anonymität abgeleiteten datenorientierten Empfehlungen dargestellt (für eine ausführliche Darstellung und Begründung siehe Müller et al. 1991:443ff.). Diese Empfehlungen beziehen sich nur auf den Mikrozensus und die EVS, weil die Untersuchung nur auf diese Datenfiles ausgerichtet war. Inwieweit diese Empfehlungen auf andere Datenbestände übertragbar sind, müßte gesondert untersucht werden.

6. Datenorientierte Empfehlungen für die Übermittlung faktisch anonymer Daten

Datenorientierte Schutzvorkehrungen beruhen letztendlich immer auf einer Reduktion der in den Daten enthaltenen Informationen, womit auch eine Verringerung des Analysepotentials einhergeht. Wenn ein gewisser Informationsverlust aus Datenschutzgründen auch unvermeidbar ist, so sollten die Anonymisierungsmaßnahmen dennoch so gestaltet sein, daß das Analysepotential möglichst geringfügig beeinträchtigt wird.

Das wissenschaftliche Potential von amtlichen Mikrodaten liegt in der Präzision von Aussagen über sachlich oder regional tiefgegliederte Bevölkerungsgruppen. Nur mit Mikrodaten können auch zahlenmäßig kleine

Bevölkerungsgruppen in ihrer Größe präzise bestimmt und in Veränderungen genau analysiert werden. Dies gilt auch für regionalspezifische Analysen. Nur aufgrund der umfangreichen Stichproben vom amtlichen Erhebungen, wie z.B. des Mikrozensus, können aussagekräftige Analysen auch regional disaggregiert durchgeführt werden. Es ist davon auszugehen, daß die Wissenschaft amtliche Mikrodaten genau zu diesen Zwecken benötigt.

Die gleichzeitige regionale und sachliche Tiefengliederung ist jedoch ein wesentlicher Faktor der oben dargestellten Risikokonstellation. Bei bestimmten Analyseverfahren (tabellarischen Aufgliederungen) ist die gleichzeitige sachliche und regionale Tiefengliederung nicht sinnvoll, da Zellenbesetzungen sehr klein werden und wegen großer Zufallsschwankungen wenig aussagefähig sind. Hier sind entweder in der regionalen oder sachlichen Analysedimension aus statistischen Gründen Vergröberungen vorzunehmen. Bei multivariaten Analyseverfahren sind solche Aggregierungen nicht erforderlich und beeinträchtigen das Analysepotential. Dennoch erschien es als die sinnvollste Lösung, aus Datenschutzgründen bei der Datenweitergabe für den Regelfall für den Mikrozensus zwei unterschiedliche Datenfiles vorzusehen: ein sogenanntes Grundfile und ein sogenanntes Regionalfile. Mit Hilfe des Regionalfiles soll es möglich sein, Mikrodaten auf einer Ebene von Regionaleinheiten zu analysieren, für welche die Daten entsprechend dem Stichprobenplan noch als repräsentativ gelten. Bei der EVS sind dies die Bundesländer. Eine weitergehende Regionalisierung als im Grundfile des Mikrozensus vorgesehen, erschien daher für die EVS nicht sinnvoll.

Im Grundfile sind die Regionalinformationen nur in undifferenzierter Form enthalten. Sie schließen die Angabe über das Bundesland (außer für die Bundesländer Bremen und Saarland) und eine Klassifikation des siedlungsstrukturellen Typs ein. Empfohlen wurden hierbei Typisierungsmerkmale, wie zum Beispiel die von der Bundesforschungsanstalt für Landeskunde und Raumordnung entwickelte Gemeindetypologie, beziehungsweise eine vergröberte Klassifikation der Gemeindegrößenklasse als Alternative. Alle übrigen Merkmale sollen in möglichst großer Differenzierung enthalten sein, wobei auch der Haushaltszusammenhang der Befragten erhalten bleiben soll.

Das Regionalfile enthält stärker differenzierte Regionalinformationen, grenzt dafür aber die Differenzierungstiefe bei den übrigen Variablen ein. Damit wurden zwei Elemente entkoppelt, die in der Verbindung (und nur in der Verbindung) zu der oben dargestellten begrenzten Risikokonstellation führen. Werden die Daten ohne kleinräumigen Regionalbezug übermittelt,

dann sind sie - das haben die Experimente gezeigt - bereits durch die Entfernung der personenbezogenen Angaben faktisch anonym. Als zusätzliche Sicherung wird für das Grundfile empfohlen:

Mikrozensus:

Festlegung eines Minimums in den univariaten Randverteilungen, so daß jede ausgewiesene Merkmalsausprägung für die Grundgesamtheit der Bundesrepublik mindestens 5000 Fälle umfaßt. Dies entspricht circa 50 Fällen im Datensatz des Mikrozensus.

- Es darf keine einzelne Gemeinde eingrenzbar sein, die weniger als 500.000 Einwohner umfaßt.
- Ein Gemeindetyp (z.B. Gemeindegrößenklasse), dem mehrere Gemeinden zugehören, darf in keinem Bundesland weniger als 400.000 Einwohner umfassen.
- Angaben über Staatsangehörigkeit werden nur so weitergegeben, daß eine Nationalität oder eine identifizierbare Gruppe von Nationalitäten in der Bundesrepublik insgesamt wenigstens 50.000 Einwohner umfaßt. Dies entspricht circa 500 Fällen im Mikrozensus.

Einkommens- und Verbrauchsstichprobe:

"Sichtbare" oder über die Zeit vergleichsweise stabile Merkmale - wie Geburtsjahr, Stellung im Beruf oder Besitz auffälliger Konsumgüter - sollen so aggregiert werden, daß nur Merkmalsausprägungen ausgewiesen werden, die für die Grundgesamtheit der Bundesrepublik mindestens 5000 Fälle umfassen. Dies entspricht circa zehn Fällen in der EVS.

Bei öffentlich wenig bekannten oder über die Zeit wenig stabilen, jedoch differenziert erfaßten Merkmalen - im wesentlichen die nichtgruppierten Einkommens-, Vermögens- und Ausgabenbeträge - sollen die jeweils fünf niedrigsten und fünf höchsten Ausprägungen eines Merkmals nur als Mittelwert dieser Ausprägungen ausgewiesen werden. Die übrigen Ausprägungen im untersten und obersten Dezil der Verteilung eines solchen Merkmals sollen mit einem Zufallsfehler von bis zu plus oder minus ein Prozent des jeweiligen Merkmalswertes überlagert werden.

Im Regionalfile ist die faktische Anonymität durch weitere Ausprägungsvergröberungen bei den sehr differenziert erfaßten Merkmalen Beruf, Wirtschaftszweig, Geburtsjahr und Nationalität zu sichern. Im Detail wurden folgende Einzelmaßnahmen für das Regionalfile des Mikrozensus vorgeschlagen:

Regionalfile:

- Durch die Kombination von Regionalklassifikationen soll keine Regionaleinheit ermittelbar sein, die eine Einwohnerzahl von weniger als 100.000 Personen aufweist.
- Die Überschneidungsmerkmale Beruf, Wirtschaftszweig, Nationalität und Alter sollen so vergröbert werden, daß keine Ausprägungen ausgewiesen werden,
  - die in der Grundgesamtheit nicht wenigstens 50.000 Einwohner umfassen;
  - die pro übermittelter Regionaleinheit (ohne Substichprobenziehung) nicht mindestens drei Fälle im Mikrodatenfile enthalten. Merkmalsausprägungen, die im Mikrodatenfile nur einen oder zwei Fälle enthalten, werden nur in einer stärker aggregierten Weise ausgewiesen.
- Alle übrigen Variablen sollen - falls erforderlich - so aggregiert werden, daß jede ausgewiesene Merkmalsausprägung für die Grundgesamtheit der Bundesrepublik mindestens 5000 Fälle umfaßt.

Weiterhin wurde die Ziehung von Substichproben empfohlen. Die Substichprobenziehung verhindert, daß ein Datenangreifer mit Sicherheit weiß, ob eine bestimmte Person im übermittelten Mikrodatenfile enthalten ist. Hierdurch wird der Unsicherheitsfaktor bei einem Reidentifikationsversuch erhöht. Zugleich verringert sie prinzipiell das Reidentifikationspotential. Die Ziehung von Substichproben ist in jedem Fall mit Einschränkungen in der Präzision von Analyseergebnissen verbunden, aber sie wirkt sich im Regionalfile schwerwiegender aus als im Grundfile.

Nach dem neuen Stichprobenplan bilden die unterste Ebene, für die der Mikrozensus regional repräsentative Aussagen zuläßt, Regionaleinheiten in der Größe von circa 200.000 Einwohnern. Für solche Einheiten enthält der Mikrozensus circa 2000 Fälle. Bei nur wenigen weiteren Aufgliederungen sind die Fehlerbereiche bei Stichproben dieses Umfangs schon sehr groß. Jede Substichprobenziehung bedeutet deshalb eine empfindliche Einschränkung des Analysepotentials. Es wurde deshalb empfohlen, die Substichprobe für das Regionalfile keinesfalls niedriger als bei 85 Prozent anzusetzen.

Beim Grundfile muß eine Substichprobenziehung ebenfalls bei vielen Analysen als empfindlicher Informationsverlust gewertet werden. Insbesondere in Analysen mit multivariaten tabellarischen Aufgliederungen werden auch bei sehr umfangreichen Stichproben wie dem Mikrozensus sehr schnell die Grenzen sichtbar, unterhalb derer eine Substichprobenziehung an der

Substanz der Analysemöglichkeiten des Mikrozensus rührt. Es wurde deshalb empfohlen, beim Grundfile als unterste Grenze eine Substichprobe von 70 Prozent anzusetzen.

Bei der EVS soll das Datenfile in Abhängigkeit der benötigten Erhebungsteile zukünftig mit folgenden Auswahlsätzen weitergegeben werden:

- 98 Prozent:
  Haushalts- und Personenmerkmale aus dem Grundinterview + 1 Erhebungsteil, [15)]
- 90 Prozent:
  Haushalts- und Personenmerkmale aus dem Grundinterview + 2 Erhebungsteile, [15)]
- 80 Prozent:
  Haushalts- und Personenmerkmale aus dem Grundinterview + 3 Erhebungsteile. [15)]

Ergänzend wurde die Weitergabe einer Ein-Prozent-Substichprobe aus dem Mikrozensus empfohlen. Diese sollte, mit Ausnahme der Regionalangaben, sämtliche Merkmale des Mikrozensus ohne weitergehende Anonymisierungsmaßnahmen enthalten. Durch den Wegfall von Regionalinformationen ist die starke Substichprobenziehung eine hinreichende Schutzmaßnahme. Dieses Subfile ist für Analysen gedacht, in welchen eine Massenbasis nicht erforderlich ist oder für Wissenschaftler, die eine Massendatenanalyse nicht selbst vornehmen, jedoch an einem kleineren Datenfile geplante Analysen testen wollen.

7. Ausblick

Die hier dargestellten Empfehlungen sind eine erste Konkretisierung für faktisch anonymisierte Mikrodatenfiles der EVS und des Mikrozensus, die in einem Standardfall angewandt werden können. Nach einiger Zeit der Praxis und Erfahrungssammlung sollten sie nochmals überprüft und gegebenenfalls revidiert werden. Da insbesondere noch keine genauen Erfahrungen dazu vorliegen, auf welchem Niveau der neue Stichprobenplan des Mikrozensus regional verwertbare Analysen zuläßt, sind vor allem die für das Regionalfile gemachten Vorschläge als Orientierungsgrößen zu betrachten, die möglicherweise schon bald in Abstimmung mit Regionalforschern an neue Erkenntnisse, Erfahrungen und einen neuen Bedarf anzupassen sind. Es ist zwar angestrebt, durch diese Epfehlungen eine gewisse Routinisierung und damit auch Aufwandseinsparung, sowie - falls möglich - eine

Kostensenkung bei der Datenweitergabe zu erreichen. Es wird jedoch weiterhin möglich sein, für spezifische Forschungszwecke durch eine unterschiedliche Ausgestaltung verschiedener Anonymisierungs- und Sicherungsmaßnahmen (z.B. Merkmalsvergröberung, Stichprobenziehung, technisch-organisatorische Maßnahmen) Lösungen vorzusehen, die bei einem vergleichbaren Schutz vor Deanonymisierung auf spezifische Forschungsvorhaben abgestimmt sind (Knoche 1991).

Anmerkungen

1) Mikrodaten beziehen sich - im Gegensatz zu Makrodaten oder Aggregatdaten - auf Informationen über einzelne Elementareinheiten, daher werden sie gelegentlich auch als Einzelangaben bezeichnet. Als Elementareinheiten kommen hierbei sowohl Personen bzw. Individuen, Betriebe, Institutionen oder sonstige Objekte in Frage, sofern sie Gegenstand einer Datensammlung sind. Beziehen sich die Mikrodaten auf Individuen, werden sie auch als Individualdaten bezeichnet (vgl. Müller et al. 1991:1).
2) Vom 3. - 5. März 1986 fand im Statistischen Bundesamt das wissenschaftliche Kolloquium "Nutzung von anonymisierten Einzelangaben aus Daten der amtlichen Statistik - Bedingungen und Möglichkeiten" (vgl. Statistisches Bundesamt 1987) statt.
3) Die Finanzierung der Hauptkosten des Anonymisierungsprojektes (1988-1990) erfolgte durch das Bundesministerium für Forschung und Technologie. Eine ausführliche Darstellung der Projektergebnisse findet sich in Müller/Blien/Knoche/Wirth (1991): Die faktische Anonymität von Mikrodaten.
4) Im allgemeinen spricht man hierbei von Personenbeziehbarkeit. Die Abgrenzungsprobleme zwischen Personenbezug, Personenbeziehbarkeit und Anonymität werden unter anderem bei Brennecke (1980) diskutiert. Für eine kritische Diskussion von Personenbeziehbarkeit als Prinzip siehe Scheuch (1987).
5) Die wichtigsten allgemeinen theoretischen Ansätze zu Fehlerquellen bei der Datenerhebung basieren einerseits auf Grundlagen der Kognitions-, Emotions- und Motivationspsychologie (Schwarz et al. 1988, 1989; Hippler et al. 1987; Schwarz 1990; Sudman/Bradburn 1974; Bradburn/Sudman 1979), andererseits auf einer allgemeinen Theorie des sozialen Handelns, bei welcher das Verhalten in Befragungssituationen als Spezialfall sozialen Handelns angesehen wird (Esser 1984a, b, c, d, 1986).
6) Für eine ausführliche Darstellung siehe Müller et al. (1991:49-85); Bender (1990); Müller, M. (1991).
7) Die angesprochenen Dimensionen können an einem Beispiel verdeutlicht werden. Hypothetisch wird unterstellt, daß ein Angreifer 100 Personen im Mikrozensus deanonymisieren will: Unter der Annahme, mit einer Reidentifikationstechnik könnte jede zehnte Person, deren Daten sowohl im Zusatzwissen als auch im Mikrozensus enthalten sind, reidentifiziert werden, wäre bedingt durch den Auswahlsatz des Mikrozensus von einem Prozent nur jeder tausendste Fall des Identifikationsfiles auch im Mikrozensus auffindbar. Für eine Reidentifikation von 100 Fällen, müßte das Identifikationsfile demnach 100.000 Datensätze enthalten.
8) In dieser Untersuchung wurden über 10.000 Bundesbürger befragt. Bezogen auf Nordrhein-Westfalen enthielt sie 2685 Fälle. Aus Datenschutzgründen wird, in Abweichung von der üblichen Praxis, der Name der Erhebung und der beteiligten Institutionen nicht genannt.
9) Da hier die spezifische Subpopulation der Gelehrten betrachtet wird, die in aller Regel eindeutig durch Beruf und Branche gekennzeichnet ist, bedeutet dies für die Anzahl

der Ausprägungen von Wirtschaftszweig 1 und Beruf 1, daß knapp 99 Prozent der erfaßten Fälle hier jeweils in eine Kategorie fallen. Die verbleibenden Kategorien werden von "Ausreißern" eingenommen und sind nur schwach besetzt. Demgegenüber erklärt sich die erhöhte Anzahl von Ausprägungen bei Wirtschaftszweig 2 und Beruf 2 damit, daß hier jeweils breiter gestreute Alternativen zu der Tätigkeit der erfaßten Personen aufgenommen wurden.

10) Im Falle einer mehrdeutigen Zuordnung liegt keine eins-zu-eins, sondern eine 1:n, n:1 beziehungsweise n:m Zuordnung von Datensätzen vor.
11) Für den ALLBUS 1990 wurde dem Umfrageinstitut pro Interview etwa 130 Mark bezahlt (inklusive Datenfilerstellung und Plausibilitätskontrollen).
12) Für eine detaillierte Erklärung dieser Ergebnisse siehe Müller et al. 1991:302ff.
13) Berücksichtigte Variablen (in Klammern: Anzahl der Merkmalsausprägungen): Geschlecht (2); Alter (58); Familienstand (5); Allgemeiner Schulabschluß (5); Beruf. Ausbildungsabschluß (5); Beruf. Erwerbstätigkeit derz.(22); Arbeitslos (2); Berufliche Stellung derz. (22); Berufliche Stellung früher (20); Arbeitswochenstunden (74); Monatliches Netto-Einkommen (56).
14) Vgl. hierzu auch die Ergebnisse der argumentativen Analyse einzelner Szenarien (Müller et al. 1991:351ff.).
15) Erhebungsteile in diesem Sinn sind das Schlußinterview, der Erhebungsteil über die Nahrungs- und Genußmittel sowie der Erhebungsteil über die Jahresrechnung.

Literatur

Beckmann, P., 1988: Die Bedeutung des Zusatzwissens vor dem Hintergrund einer potentiellen Deanonymisierung von Mikrozensus und Einkommens- und Verbrauchsstichprobe, Arbeitsbericht aus dem Anonymisierungsprojekt Nr.5.

Bender, S., 1990: De-Anonymisierung von Individualdaten bei statistischen Erhebungen. Eine Diskussion des diskriminanzanalytischen Verfahrens von Paaß/Wauschkuhn (Diplomarbeit, Universität Mannheim).

Bender, S./Blien, U./Müller, M., 1990a: Grundidee der diskriminanzanalytischen Methode von PAASS und WAUSCHKUHN zur Zuordnung anonymisierter Datensätze, Arbeitsbericht aus dem Anonymisierungsprojekt Nr.11.

Bender, S./Blien, U./Müller,M., 1990b: Implementation und erste Tests der diskriminanzanalytischen Methode von PAASS und WAUSCHKUHN zur Zuordnung anonymisierter Datensätze.Erfahrungsbericht und Abschätzung des notwendigen Arbeitsaufwandes, Arbeitsbericht aus dem Anonymisierungsprojekt Nr.16.

Bethlehem, J.G./Keller, W.J./Pannekoek, J., 1990: Disclosure Control of Microdata. Journal of the American Statistical Association, Volume 85:38-45.

Blien, U., 1989: Technisch-organisatorische Sicherungsmaßnahmen gegen unbefugte Datenzugriffe bei faktisch anonymen Daten, Arbeitsbericht aus dem Anonymisierungsprojekt Nr.8.

Blien, U./Müller, M., 1991: Empirische Überprüfung der Anonymität des Mikrozensus mit der diskriminanzanalytischen Methode von Paaß und "Kürschners Gelehrtenkalender" als Zusatzwissen, Arbeitsbericht aus dem Anonymisierungsprojekt Nr.18.

Block, H./Olsson, L., 1976: Bakvägsidentifiering, in: Statistiskal Tidskrift 14:133-144. (Engl. Version: Backwardsidentification (unveröffentlicht)).

Bradburn, N./Sudman, S. and Associates, 1979: Improving Interview Method and Questionnaire Design. San Francisco: Jossey Bass.

Brennecke, R., 1980: Kriterien zur Operationalisierung der faktischen Anonymisierung, in: Kaase et al.:158-175.

Brennecke, R./Schneider H., 1977: Zur Problematik des Bundesdatenschutzgesetzes für die Forschung. SPES-Arbeitspapier Nr.63, Sozialpolitische Forschungsgruppe Frankfurt/ Mannheim.

Brunnstein, K., 1987: Über die Möglichkeit der Re-Identifikation von Personen aus Volkszählungsdaten, in: Appel, R. (Hrsg.): Vorsicht Volkszählung! Köln: Volksblattverlag, (2.Auflage).

Burkert, H., 1979: Die Eingrenzung des Zusatzwissens als Rettung der Anonymisierung? Datenverarbeitung im Recht, Bd. 8, Heft 1:63-75.

Burkert, H., 1980: Das Problem des Zusatzwissens, in: Kaase et al.:143-147.

Dalenius, T., 1977: Towards a Methodology for Statistical Disclosure Control. Statistical Review 5/1977:429ff.

Dalenius, T., 1986: Finding a Needle in a Haystack or Identifying Anonymous Census Records. Journal of Official Statistics 2:329-336.

Dalenius, T., 1988: Controlling Invasion of Privacy in Surveys, Continuing Education Program, Statistics Sweden.

Dittrich, K./Schlörer, J., 1985: Anonymisierung von Forschungsdaten. Bericht im Auftrag des Ministeriums für Wissenschaft und Kunst Baden-Württemberg, Mai 1985.

Dorer, P./Mainusch, H./Tubies, H., 1988: Bundesstatistikgesetz. Gesetz über die Statistik für Bundeszwecke mit den Leitsätzen des Volkszählungsurteils, Mikrozensusgesetz und Volkszählungsgesetz. Kommentar, München: C. H. Beck.

Esser, H., 1984a: Fehler bei der Datenerhebung, Kurseinheit 1: Methodologische Probleme bei der empirischen Kritik von Theorien, Fernuniversität Hagen.

Esser, H., 1984b: Fehler bei der Datenerhebung, Kurseinheit 2: Meßfehler bei der Datenerhebung und die Techniken der empirischen Sozialforschung, Fernuniversität Hagen.

Esser, H., 1984c: Fehler bei der Datenerhebung, Kurseinheit 3: Datenerhebung als sozialer Prozeß, Fernuniversität Hagen.

Esser, H., 1984d: Fehler bei der Datenerhebung, Kurseinheit 4: Meßfehler in Kausalmodellen, Fernuniversität Hagen.

Esser, H., 1986: Können Befragte lügen? Zum Konzept des "wahren Wertes" im Rahmen der handlungstheoretischen Erklärung von Situationseinflüssen bei der Befragung. Kölner Zeitschrift für Soziologie und Sozialpsychologie 37:314-336.

Fischer-Hübner, S., 1986: Zur Anonymität und Reidentifizierbarkeit statistischer Daten, Mitteilungen Nr. 143 des Fachbereichs Informatik der Universität Hamburg.

Greenberg, B., 1990: Disclosure Avoidance Research at the Census Bureau, paper presented at the 1990 Annual Research Conference, Bureau of the Census, Arlington, Virginia.

Hamacher, B., 1980: Resümee zu Datenschutzmaßnahmen, in Kaase et al.:219-224.

Helmcke, T., 1989: Allgemeine Kosten-Nutzen-Überlegungen zu Deanonymisierungsversuchen, Arbeitsbericht aus dem Anonymisierungsprojekt Nr.6.

Hippler, H.-J./Schwarz, N./Sudman, S. (Hrsg.), 1987: Social Information Processing and Survey Methodology, New York/Heidelberg: Springer.

Kaase, M./Krupp, H.-J./Pflanz, M./Scheuch, E.K./Simitis, S. (Hrsg.), 1980: Datenzugang und Datenschutz. Königstein/Ts.:Athenäum.

Knoche, P., 1989: Außerberufliche Motive, Arbeitsbericht aus dem Anonymisierungsprojekt Nr.12.

Knoche, P., 1991: Der neue Leitfaden des Statistischen Bundesamtes für die Weitergabe von Einzeldaten des Mikrozensus und der Einkommens- und Verbrauchsstichprobe. Vortrag auf der Tagung "Faktisch anonyme Einzeldaten der amtlichen Statistik" in Mannheim, Dezember 1991.

Koch, A., 1986: Wie zuverlässig lassen sich Berufs- und Bildungsvariablen messen? Ergebnisse einer Test-Retest-Studie zur Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften 1984, Diplomarbeit, Universität Mannheim.

Krupp, H.-J./Preißl, B., 1989: Die Neufassung des BDSG und die wissenschaftliche Forschung. Computer und Recht 5/2:121ff.

Marsh, C./Skinner, C./Arber, S./Penhale, B./Openshaw, S./Hobcroft, J./Lievesley, D./Walford, N., 1991: The Case for Samples of Anonymized Records from the 1991 Census. Journal of the Royal Statistical Society, Vol.154.

Mohler, P., Ph./Kaase, M., 1980: Formen der Erhebung in der empirischen Sozialforschung, in: Kaase et al.:107-110.

Müller, M., 1991: Reidentifikation von Individualdaten: Experimentelle Überprüfung im Rahmen eines sozialwissenschaftlichen Szenarios mit der Methode von Paaß, Wauschkuhn (Diplomarbeit, Universität Mannheim).

Müller, W., 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, W./Blien, U./Knoche, P./Wirth, H., 1991: Die faktische Anonymität von Mikrodaten (Band 19 der Schriftenreihe Forum der Bundesstatistik). Metzler-Poeschel, Stuttgart.

Müller, W./Hauser, R., 1987: Der Bedarf der Wissenschaft an anonymisierten Einzelangaben, in: Statistisches Bundesamt (1987).

Neue juristische Wochenschrift, 1984: Urteil des Bundesverfassungsgerichts zum Volkszählungsgesetz (8):419-428.

Paaß, G., 1987: Re-Identifikationsrisiko von Einzelangaben, in: Statistisches Bundesamt (Hrsg.): Nutzung von anonymisierten Einzelangaben aus Daten der amtlichen Statistik. Bedingungen und Möglichkeiten, Stuttgart, Mainz: Kohlhammer.

Paaß, G./Wauschkuhn, U., 1985: Datenzugang, Datenschutz und Anonymisierung. Analysepotential und Identifizierbarkeit von anonymisierten Individualdaten, München, Wien: R. Oldenbourg.

Porst, R./Zeifang, K., 1987: Wie stabil sind Umfragedaten? Beschreibung und erste Ergebnisse der Test-Retest-Studie zum ALLBUS 1984. ZUMA Nachrichten 20:8-31.

Scheuch, E.K., 1980: Die Weiterentwicklung des Datenschutzes als Problem der Sozialforschung, in: Kaase et al.:252-275.

Scheuch, E.K., 1987: Risikointerpretation beim Datenschutz, in: Statistisches Bundesamt 1987: 121-145.

Schnell, R./Hill, P.B./Esser, E., 1988: Methoden der empirischen Sozialforschung, Oldenbourg, München.

Schlörer, J., 1980: Anonymisierung von Mikrodaten: Technische Aspekte, in: Kaase et al.:118-142.

Schwarz, N., 1990: Assessing Frequency Reports of Mundane Behaviors: Contributions of Cognitive Psychology to Questionnaire Construction, ZUMA-Arbeitsbericht 1988/10, abgedruckt in: Hendricks, C., Clark, M. (Hrsg.): Research Methodology (Review of Personality and Social Psychology, Vol. 11), Beverly Hills: Sage.

Schwarz, N./Hippler, H.-J./Noelle-Neumann, E., 1989: Response Order Effects in Long Lists: Primacy, Recency, and Asymmetric Contrasts Effects, Mannheim, ZUMA-Arbeitsbericht Nr. 89/18.

Schwarz, N./Hippler, H.-J./Noelle-Neumann, E., 1989: Einflüsse der Reihenfolge von Antwortvorgaben bei geschlossenen Fragen. ZUMA-Nachrichten 25:24-38.

Schwarz, N./Hippler, H.-J./Strack, F., 1988: Kognition und Umfrageforschung: Themen, Ergebnisse und Perspektiven. ZUMA-Nachrichten 22:15-28.

Skinner, C.J./Marsh, C./Openshaw, S./Wymer, C., 1990: Disclosure Avoidance for Census Microdata in Great Britain, in: Annual Research Conference, Proceedings, 1990, U.S. Department of Commerce, Bureau of the Census.

Spruill, N., 1983: Testing Confidentiality of Masked Business Microdata, Working Paper, PRI 83-07.09, The Public Research Institute, Alexandria, Virginia .

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

Statistisches Bundesamt (Hrsg.), 1987: Nutzung von anonymisierten Einzelangaben aus Daten der amtlichen Statistik. Bedingungen und Möglichkeiten, Stuttgart, Mainz: Kohlhammer.

Südfeld, E., 1987: Anonymisierungsstandards und generelle Abwicklungsregelungen für Anforderungen nach anonymisierten Einzelangaben im Statistischen Bundesamt, in: Statistisches Bundesamt (1987).

Sudman, S./Bradburn, N., 1974: Response Effects in Surveys. Chicago: Aldine Publishing Company.

Zapf, W., 1985: Der Zugang der Wissenschaft zur statistischen Information - Forderung und Realität. In: Statistisches Bundesamt (1985).

Anhang

Die Ergebnisse des Anonymisierungsprojektes wurden auf einer Tagung am 11.12.1991 an der Universität Mannheim vorgestellt. Nachfolgend geben wir die auf der Tagung stattgefundene Podiumsdiskussion in leicht gekürzter Form wieder. Thema der Podiumsdiskussion war vor allem die Übertragbarkeit der Ergebnisse des Anonymisierungsprojektes auf andere Forschungsbereiche.

Podiumsdiskussion: Die neuen Erkenntnisse zur faktischen Anonymität und ihre Übertragbarkeit auf andere Daten- und Forschungsbereiche in der Wissenschaft und der amtlichen Statistik

Prof. Dr. Allerbeck, Universität Frankfurt:
Der Titel unserer Diskussion umfaßt vier Zeilen aus denen ich als operative Worte eigentlich zwei entnehme - "Übertragbarkeit" und "andere". Der Kreis, der hier am Podium sitzt ist so illuster, daß es zuviel Zeit kosten würde ihn vorzustellen. Ich will dies unterlassen und einfach die Teilnehmer in der Reihenfolge der Liste bitten, ihren ersten Beitrag abzugeben.

Dr. Schmidt, Bundesbeauftragter für den Datenschutz, Bonn:
Zunächst möchte ich mich für die Gelegenheit bedanken, daß ich hier so interessante und in jeder Richtung fundierte Beiträge zu einem schwierigen Thema hören konnte, das nicht zu unrecht als Dilemma bezeichnet wurde. Der Versuch ist hiermit gelungen, die in Paragraph 16, Absatz 6 Bundesstatistikgesetz als Voraussetzung einer Datenübermittlung an die Forschung genannte faktische Anonymisierung zu beschreiben und sie nicht nur zu beschreiben, sondern auch erreichbar zu machen. Über diese abstrakte Bemerkung hinaus ist auch noch plausibel dargelegt worden, nicht nur daß, sondern auch mit welchen einfachen Mitteln und aus meiner Sicht für die Forschung auch erträglichen Mitteln, die Daten faktisch hinreichend gut anonymisiert werden können. Wenn nun trotzdem auch für diese faktisch anonymisierten Daten Sicherungsmaßnahmen und weitere Auflagen und Einschränkungen geboten sind, und das bedeutet im Ergebnis, daß der gedankliche Abstand zum sogenannten public file noch ganz erheblich ist, dann sollen diese flankierenden Maßnahmen den aufgezeigten Weg nicht versperren, sondern gewährleisten, daß er auch wirklich mit Erfolg beschritten werden kann. Ich halte es deshalb für angebracht, daß die Überlassung faktisch anonymisierter Daten maßgeschneidert durchgeführt wird und das fallweise geprüft wird. Beides wird durch die hier diskutierten Erkenntnisse aus meiner Sicht erheblich erleichtert und praktikabel. Damit

scheint mir eine Entwicklung in Richtung auf eine wirklich zu Buche schlagende Aufwandsminderung durch wiederholbare Vorgehensweise sehr leicht möglich. Dies dürfte die Durchführbarkeit von Forschungsvorhaben schon in einer sehr frühen Planungsphase weit besser kalkulierbar machen.

Prof. Dr. Dr. Häfner, Zentralinstitut für seelische Gesundheit, Mannheim:
Ich muß zunächst darauf verweisen, daß ich Mediziner bin - mein Arbeitsgebiet ist die psychiatrische Epidemiologie. Vorweg möchte ich das Fazit ziehen, daß dieses Projekt meiner Wissenschaft direkt wenig Nutzen bringt, aber dennoch in einem weiteren Zusammenhang erheblichen Nutzen bringen könnte, so hoffe ich jedenfalls. Relativ wenig Nutzen haben wir insofern, als der größte Teil der medizinischen Forschung, und dazu zählt auch die analytische Epidemiologie, auf Identifikatoren oder mindestens sprechende Codes angewiesen ist, damit Informationen aus unterschiedlichen Quellen und über Zeit einem Individuum zugeordnet werden können. Der zweite Grund liegt darin, daß die Daten der amtlichen Statistik der Bundesrepublik über die Gesundheit der Bevölkerung relativ wenig tief gegliederte, valide Informationen enthalten. Es gibt im Grunde außer der Todesursachenstatistik, deren Daten wenigstens in einigen Bereichen hinreichend valide sind, um als zuverlässige Gesundheitsindikatoren auf der Makroebene zu dienen, eigentlich nur die Daten des Mikrozensus. Sie stellen ein Gemenge aus subjektiven und objektiven Gesundheitsvariablen dar. Ihre geringe Validität läßt sich bereits an einem schlichten Vergleich der Erhebungsergebnisse über verschiedene Querschnitte hinweg feststellen. Sie weisen in einigen Kategorien so erhebliche Unterschiede auf, daß eine Erklärung durch Morbiditätstrends nicht mehr plausibel ist.

Daß die deskriptive Epidemiologie und die Gesundheitssystemforschung unter diesen ungünstigen Bedingungen leiden, läßt sich auch an ihren Entwicklungsdefiziten ablesen. Besonders deutlich wird dies beim Defizit der Public-Health-Forschung in der Bundesrepublik, zu deren Grundlagenwissenschaften das epidemiologische Studium der Beziehung zwischen wirtschaftlichen, sozialen und ökologischen Indikatoren einerseits und Gesundheitsindikatoren andererseits zählt. Der Bundesforschungsminister hat dieser Tage ein hochdotiertes Programm für die Förderung und Institutionalisierung von Public-Health-Forschung aufgelegt, das die Situation im Lande in eindrucksvoller Weise widerspiegelte: einen erschreckenden Mangel an anspruchsvoller epidemiologischer Forschung, aber auch an hinreichend zuverlässigen Gesundheitsdaten sowohl auf Bundesebene als auch auf der Ebene der Länder, Gebietskörperschaften und Kommunen, die als Grundlagen für gute epidemiologische Public-Health-Forschung geeignet wären. Man kann Public-Health-Forschung nur sinnvoll betreiben, wenn ein hinreichender Pool zuverlässiger Daten auf diesem

Gebiet zur Verfügung steht. Hier ist nicht nur eine bessere Kooperation zwischen Datenschutz, Wissenschaft und Gesetzgeber gefordert; hier muß auch mehr Problembewußtsein, vor allem gegenüber dem von unkritischen Datenschützern oft gebrauchten Wort der "Datensammelwut", zum Tragen kommen.

Wenn Public-Health-Forschung mehr an den Veränderungen des Gesundheitszustands der Bevölkerung im Zusammenhang mit Risikofaktoren und im Kontext von ökologischen und Bevölkerungsvariablen interessiert ist, so sind die analytische Epidemiologie und der größte Teil der medizinischen Forschung überhaupt, wie schon eingangs angesprochen, auf Individualdaten angewiesen. Die Quellen, aus denen institutsübergreifende medizinische Forschung, und das gilt für epidemiologische Projekte nahezu in allen Fällen, Informationen über krankheitsrelevante Sachverhalte gewinnen muß, sind in der Regel Ärzte oder ärztliche geleitete Institutionen. Hier kommt mit der ärztlichen Schweigepflicht das Thema der faktischen Anonymisierung ins Spiel. Der Paragraph 203 StGB wird durch Standesrecht ausgefüllt. Die entsprechenden Bestimmungen in der Bundesärzteordnung und in den Landesärzteordnungen erlauben eine Weitergabe ärztlicher Daten zu Zwecken der Forschung nur mit Einwilligung des Patienten oder anonymisiert. Wo die Einwilligung des Kranken nicht erlangt werden kann, beispielsweise weil er nicht einwilligungsfähig oder bereits verstorben ist oder die Einholung einer Einwilligung nur unter unbilligem Aufwand an Zeit und Kosten zu erreichen wäre, konkretisiert sich die Problematik der Anonymisierung. In den vergangenen Jahren ist, nicht zuletzt im Kontext der von einzelnen Datenschutzbeauftragten willfährig geschürten öffentlichen Datenschutzpsychose, der Anspruch an Anonymisierung so hoch geschraubt worden, daß eine Zuordnung der übermittelten Informationen zu einem einzelnen Fall mit hoher Zuverlässigkeit verhindert wurde.

Die Folge davon ist, daß die epidemiologische Forschung in der Bundesrepublik empfindlich beeinträchtigt wurde und eine Reihe von gravierenden medizinischen Problemen, etwa Altersdemenz bzw. Alzheimersche Erkrankung, dort ununtersuchbar wurden, wo Einwilligung und Einwilligungsfähigkeit unabdingbare Voraussetzungen für die Gewinnung der notwendigen Daten sind. Nicht weniger schwerwiegend sind die Risiko- und Therapieforschung betroffen. Wenn über Behandlungserfolge und -risiken ausgesagt werden soll, dann müssen institutionsüberschreitend Informationen über den weiteren Verlauf und muß beispielsweise die Information über einen Todesfall zuverlässig und mit der Möglichkeit der Zuordnung zu dem betreffenden Kranken gewonnen werden. Das gleiche gilt für die analytische Risikoforschung, die die kausale Beziehung zwischen Risikofaktoren und später eintretenden Gesundheitsschäden erfassen will,

ein Thema, das in der modernen Ökologie und Umweltpathologie zentrale Bedeutung einnimmt, aber einer soliden wissenschaftlichen Untersuchung in weiten Bereichen entzogen worden ist.

Für die Untersuchung kleiner Risikopopulationen, etwa solcher, die besonderer Exposition ausgesetzt oder mit spezifischen Vulnerabilitäten belastet sind, ist die Registerforschung ein wichtiges Instrument. Auch hier ist die Bundesrepublik in einer extrem restriktiven Position. Derzeit ist die Forschung mit Krankheitsregistern nur auf der Grundlage einer spezialgesetzlichen Regelung zulässig, wobei der Gesetzgeber alle zu registrierenden Items im Gesetz festgelegt. Diese Lösung ist in zweierlei Hinsicht unsinnig. Einmal führt sie dazu, daß nur eine minimale Anzahl von Registern und diese ausschließlich für solche Krankheiten, die im öffentlichen Bewußtsein stark präsent sind - sprich Krebserkrankungen - legalisiert werden. Zum anderen ist die Festschreibung der zu registrierenden Informationen im Gesetz eine unglückliche Regelung, denn ein Register muß jederzeit für den wissenschaftlichen Fortschritt offen sein, und wissenschaftlicher Fortschritt ist schneller als der Fortschritt der Gesetzgebung. Schließlich soll nicht übersehen werden, daß Risiken, die nur sehr kleine Populationen betreffen, und das gilt auch für die Exposition gegenüber Umweltgiften oder Gefahrensituationen am Arbeitsplatz, in der Tat nur mit nicht anonymisierten Individualdaten untersucht werden können. Hier muß die Abwägung zwischen konfligierenden Grundrechtsgütern doch noch anders gesehen werden als in der politischen und sozialwissenschaftlichen Forschung, denn hier geht es um die Bedrohung der Gesundheit und mitunter um die Gefährdung des Lebens von Menschen.

Damit möchte ich zum Schluß kommen: Wir haben in den letzten Jahren wegen der unserer Meinung nach irrationalen rechtlichen Beurteilung von Fallregistern einen wichtigen Teil unserer psychiatrisch-epidemiologischen Forschung mit den Daten des nationalen Dänischen Fallregisters und mit Daten der Weltgesundheitsorganisation aus zehn verschiedenen Ländern durchführen müssen. Das ist eine Situation, die eigentlich eines zivilisierten, vernünftigen Staatswesens, das die gegenwärtigen und künftigen Interessen seiner Bürger an der Erhaltung und Wiederherstellung der Gesundheit ernstnimmt, nicht würdig. Meine ganze Hoffnung baut darauf, daß allmählich die Vernunft oder eine unvoreingenommene, besonnene Abwägung zwischen bescheidenen Risiken des Mißbrauchs ärztlicher Daten in der Forschung und den Folgen der massiven Einengung medizinischer Forschung für die betroffenen Kranken und für die medizinische Wissenschaft die Oberhand gewinnen. Ich sehe in den Ergebnissen des Projekts zur faktischen Anonymität insofern einen echten Schritt vorwärts, weil es aus

dem Konsens von Sozialwissenschaft, Statistik und Bevölkerungswissenschaft und Politik - wenn auch nicht der Mehrheit aller Politiker - geboren, einen bedeutsamen Schritt auf die Notwendigkeit empirisch-sozialwissenschaftlicher Forschung zum Nutzen von Daten der amtlichen Statistiken getan hat. Ich sehe vor allem in dem Konsens, der sich in der Beurteilung der Ergebnisse dieses Projekts herausgebildet hat, einen hoffnungsvollen Schritt zu einer vernünftigeren Behandlung des Problems Geheimnisschutz und Forschung und hoffe, daß sich dieser Schritt zu einem Trend entwickeln möge, der auch der medizinischen Forschung wieder diejenigen Möglichkeiten zurückgibt, die ohne substantielles Risiko der Verletzung von Geheimnisschutz den Wiedereinstieg in die Bearbeitung einiger großer, ungelöster Forschungsfragen erlaubt.

Dr. Nowak, Statistisches Bundesamt, Wiesbaden:
In den Mittelpunkt dieser Diskussion wurde die Frage gestellt, inwieweit die Ergebnisse des hier vorgestellten Projekts auf andere Bereiche übertragen werden können. Für die Bundesstatistik gilt es dabei, sowohl die rechtlichen als auch die inhaltlichen Aspekte dieser Frage zu sehen. Den rechtlichen Rahmen bildet Paragraph 16 des Bundesstatistikgesetzes. Es hat wenig Zweck, die statistischen Ämter des Bundes und der Länder wegen dieser rechtlichen Grenzen zu schelten. Wir müssen uns einfach daran halten. Innerhalb dieser gesetzlichen Rahmenbedingungen gilt es, die inhaltlichen Kriterien zu konkretisieren. Ich glaube, hier hat uns das Projekt geholfen, eine ganze Reihe weiterer Kriterien zu erkennen. Kriterien die ansetzen an dem Begriff der Unverhältnismäßigkeit und an den Fragen: was kostet es den Angreifer, wie leicht fällt ihm eine Zuordnung und was nutzt sie ihm. Damit ist auch deutlich geworden, daß man hier die Frage stellen kann, ob es alternative Wege gibt, wie er zu diesen Informationen kommen kann. Wir haben anhand der Zahlen des Projekts gesehen, wie wenig rational es ist, den Versuch zu machen, über eine Reidentifikation vier Ergebnisse zu bekommen und dafür 100.000 Mark zu zahlen, wenn das gleiche Ergebnis über einen noch so schlechten Privatdetektiv wahrscheinlich für einen Bruchteil dieser Summe erhältlich ist. Ob man die Ergebnisse des Projekts zu Fragen der Inkompatibilität der Datensätze kurzerhand auf andere Bereiche übertragen kann, wird zu prüfen sein. Ich nehme an, daß es weiter bei der Einzelfallprüfung bleiben wird, da man auch zukünftig die Entwicklungen im Bereich der Inkompatibilität untersuchen und im Auge behalten muß.

Es ist hier mehrfach gesagt worden, daß die Ergebnisse des Projektes sich nur auf personenbezogene Informationen beziehen, nicht jedoch auf den Bereich der Wirtschaftsstatistik. Für einen großen Teil der Wirtschaftsstatistiken dürfte nach meinen Überlegungen das hier gezeigte Verfahren der

faktischen Anonymisierung ausscheiden. Man muß andere Ansätze prüfen und wird dann sehen, wie man weiterkommt.

Prof. Dr. Heinz, Arbeitsgruppe strafrechtliche Rechtstatsachenforschung und empirische Kriminologie, Institut für Rechtstatsachenforschung, Universität Konstanz:
Die Ergebnisse des hier vorgestellten Projektes sind zweifelsohne beeindruckend. In vielen Teilbereichen der sozialwissenschaftlichen Forschung wird damit das Problem der faktischen Anonymisierung lösbar sein. Allerdings gibt es auch, wie soeben schon Herr Häfner kritisch angemerkt hat, wissenschaftliche Bereiche, für die die Ergebnisses dieses Projektes deshalb geringe oder überhaupt keine Relevanz haben, weil diese Bereiche auch weiterhin auf die Erhebung von personenbezogenen Einzeldaten angewiesen sind, die erst im Prozeß der statistischen Analyse aggregiert werden können. Zu diesen Teilbereichen zählt auch die Kriminologie, die entweder die amtliche Statistik selbst als Forschungsgebiet hat, indem sie z.B. die Reliabilität der Strafverfolgungsstatistik durch Abgleich mit Daten aus anderen Quellen zu bestimmen versucht, oder aber mittels der Daten der amtlichen Statistik Untersuchungen durchführt. In beiden Fällen ist kriminologische Forschung regelmäßig auf personenbezogene Einzelangaben angewiesen. Illustrieren will ich dies an zwei konkreten Beispielen aus der empirischen kriminologischen Forschung, wobei im ersten Fall die Daten der amtlichen Statistik Gegenstand der Auswertung sind, im zweiten dagegen Grundlage für die Ziehung einer repräsentativen Stichprobe.

Wie schon aus den amtlichen Rechtspflegestatistiken hervorgeht, ist Kriminalität, jedenfalls in ihren schwereren Erscheinungsformen, ein relativ seltenes Ereignis. Die Strafverfolgungsstatistik der Bundesrepublik Deutschland weist dementsprechend schon bei der Gesamtzahl der Verurteilten gelegentlich nur einen einzigen Verurteilten aus. In Vergangenheit und Gegenwart hat man immer wieder versucht, die Strafzumessungspraxis der Gerichte im zeitlichen Längsschnitt und im regionalen Querschnitt darzustellen, insbesondere die Frage von Gleichmäßigkeit oder Ungleichmäßigkeit zu klären. Früheren Versuchen, die sich auf die veröffentlichten Daten stützten, wurde zu Recht vorgehalten, keine vergleichbaren Gruppen gebildet zu haben. Voraussetzung hierfür ist die Kontrolle der strafzumessungsrelevanten Faktoren, insbesondere der personenbezogenen Informationen der Strafverfolgungsstatistik (der Verurteilung zugrundeliegende Tat, Alter, Geschlecht und Zahl der Vorstrafen der Verurteilten). Die elektronische Datenverarbeitung ermöglicht es der sozialwissenschaftlichen Forschung, die Rohdaten z.B. der Strafverfolgungsstatistik zur Bildung derartiger homogener Gruppen zu nutzen. Bei entsprechender Gruppenbildung ist es aber, da die Strafzumessungspraxis in kleinen regionalen

Einheiten, z.B. Landgerichtsbezirken, miteinander verglichen werden soll, von vornherein nicht auszuschließen, daß man auf Einzelfälle stößt. Dies ist selbst bei der Untersuchung von insgesamt häufiger vorkommenden Delikten erwartbar. Erst bei Bildung derartiger merkmalshomogener Gruppen können Art und Höhe der Strafe in den verschiedenen Regionen auf Unterschiede hin überprüft werden. Und erst nachdem diese homogenen Gruppen gebildet sind, können etwa auftretende Einzelfälle von der weiteren Auswertung ausgeschlossen werden. Würde stattdessen von vornherein auf die Übermittlung und Auswertung von Einzelangaben verzichtet, wäre die Bildung homogener Gruppen prinzipiell verhindert bzw. derart erschwert, daß nicht mehr entschieden werden könnte, ob festgestellte Unterschiede auf Eigenschaften der abhängigen (regionale Einheit) oder der unabhängigen Variablen (Strafzumessungsfaktoren) beruhen.

Das andere Beispiel der Verwendung der Daten der amtlichen Statistik als Grundlage für die Ziehung einer repräsentativen Stichprobe bildet eine Untersuchung, die meine Mitarbeiter und ich in den letzten Jahren durchgeführt haben. Hier ging es unter anderem darum, die Effekte unterschiedlicher Erledigungs- und Sanktionierungsstrategien im Jugendstrafrecht auf die Wiederauftretenswahrscheinlichkeit der betroffenen jugendlichen Straftäter zu ermitteln. Als Grundgesamtheit wurde von uns die Gesamtzahl aller im Jahr 1979 in Baden-Württemberg durch die Staatsanwaltschaften durch Verfahrenseinstellung oder durch Anklage erledigten Jugendstrafverfahren gewählt. Die Daten zu diesen Verfahren wurden durch eine Totalerhebung aus dem Rohdatensatz der Staatsanwaltschaftsstatistik für das Land Baden-Württemberg gewonnen, den das Statistische Landesamt Baden-Württemberg auf Magnetband übermittelt hatte. Nach Ziehung einer quotierten Stichprobe wurden anhand der im Rohdatensatz enthaltenen Aktenzeichen bei den Staatsanwaltschaften die Personalien der Beschuldigten ermittelt, gegen die sich die Verfahren gerichtet hatten. Für diese Personen wurden Auskünfte aus dem Zentral- bzw. Erziehungsregister eingeholt. Erst danach war es möglich, die Daten durch Aggregation zu anonymisieren.

Im Unterschied also zu dem hier vorgestellten Anonymisierungsprojekt, das die faktische Anonymisierung bereits bei Datenerhebung gewährleisten will, ist die Situation in der Kriminologie grundlegend anders. In der Kriminologie ist es regelmäßig so, daß es notwendig ist, entweder Einzeldaten in die Auswertung einzubeziehen oder Einzeldaten aus verschiedenen Datenquellen miteinander zu verknüpfen, etwa bei der Sanktions- und Wirkungsforschung. Dementsprechend stellt sich auch das Datenschutzproblem anders. Für kriminologische Forschung ist in einer ersten Untersuchungsphase der Zugang zu personenbezogenen Einzeldaten der amtlichen Statistik unverzichtbar. Dies gilt zum einen uneingeschränkt für Untersuchungen, bei denen diese Daten als Basis für Sekundärdatenanalysen benötigt werden.

Dies gilt aber auch, jedenfalls solange die Geschäftsstellenautomation noch nicht flächendeckend eingeführt ist, auch für die Stichprobenziehung für Zwecke von Aktenanalysen. Daraus resultiert die Forderung der Kriminologie nach entsprechenden gesetzlichen Forschungsklauseln und nach gesetzlichen Regelungen über die Übermittlung von Daten zu Zwecken der wissenschaftlichen Forschung bei strenger Zweckbindung, die gewährleistet, daß Daten nur im Rahmen des Forschungsvorhabens verwertet und zum frühestmöglichen Zeitpunkt anonymisiert werden.
Bei der automatisierten Datenverarbeitung selbst muß dann freilich der hohen Sensibilität dieser Daten durch entsprechende technische und organisatorischen Maßnahmen (vgl. Paragraph 9 BDSG) Rechnung getragen werden. Von besonderer Bedeutung sind vor allem die Anforderungen bezüglich der Speicher-, Zugriffs-, Übermittlungs- sowie der Eingabekontrolle. In Konstanz haben wir, um auch hier ein konkretes Beispiel zu erwähnen, dieses Problem dadurch gelöst, daß wir zum einen den Zugang nur über virtuelle Maschinen erlauben, auf die nur über dedizierte Terminals zugegriffen werden kann, und wir zum anderen ein maschinenseitiges Protokollierungsprogramm einsetzen, das jeden Datenzugriff mitprotokolliert. Um freilich den Stand der Technik in vollem Umfang zu erreichen, bedarf es professioneller Lösungen, die sich derzeit typischerweise an den Sicherheitsanforderungen des vom US-amerikanischen National Computer Security Center erarbeiteten Kriterienkatalogs zur Bewertung von Informationstechnik-Systemen ("Trusted Computer System Evaluation Criteria"), dem sogenannten Orange Book, orientieren. Dem wird z.B. bei der Neuanschaffung eines Rechners im Hochschulrechenzentrum der Universität Konstanz Rechnung zu tragen versucht. Die Anforderungen des Datenschutzrechts, die die Forschung zu beachten hat, können nur bei entsprechender Ausstattung der Hochschulrechenzentren erfüllt werden. Diese müssen entsprechend ausgestattet werden. Hierauf zu bestehten, ist Aufgabe auch der Forschung.

Dr. Vorschulte, Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen, Düsseldorf:
Das statistische Bundesamt und die statistischen Ämter der Länder geben ja bereits heute faktisch anonymisierte Einzeldatensätze an die Wissenschaft weiter. Sie tun das in wenigen Fällen und sie tun das nicht abgestimmt und mit recht unterschiedlichen Verfahren. Diejenigen von Ihnen, die bereits solche Unterlagen bekommen haben wissen, wie lange Gespräche geführt worden sind, bis schließlich ein vereinbarter Datensatz vorlag. Wir geben diese Angaben weiter an Hochschulen und an sonstige Einrichtungen mit der Aufgabe unabhängiger wissenschaftlicher Forschung. Das hier nun vorgestellte Forschungsprojekt zeigt die Anonymisierung von Einzeldatensätzen im Mikrozensus und in der Einkommens- und Verbrauchsstichprobe. Es

hat dazu als Fazit der Untersuchung Empfehlungen gegeben und es gibt, an die Empfehlungen angelehnt, einen Leitfaden, der im Statistischen Bundesamt entwickelt und hier vorgetragen wurde. Diese Empfehlungen sind nach meiner Auffassung und Einschätzung auch auf andere Statistiken in modifizierter Form übertragbar. Und ich meine, nachdem diese Arbeit vorliegt und die Summe der Empfehlungen Ihnen bekannt ist, ist es jetzt an der Wissenschaft, mit Datenwünschen auf uns zuzukommen und wir müssen dann gemeinsam überlegen, wie diese Modifizierungen letztlich aussehen müssen, damit sie den Datenbedarf befriedigt bekommen, der für die Durchführung Ihrer Aufgaben erforderlich ist. Das wollen wir Ihnen liefern. Wir sind schließlich und endlich wissenschaftsfreundlich und nicht wissenschaftsfeindlich und wollen Ihnen die Informationen, die mit viel Mühe und Aufwand erarbeitet worden sind, zugänglich machen, aber wir müssen letztendlich auch darauf achten, daß die Grundforderung der Statistik Bestand behält, daß Einzelangaben nicht einer bestimmten Person zugeordnet werden können.

Prof. Dr. Brennecke, Freie Universität Berlin:
Als Sozialmediziner gehöre ich einem Fachgebiet an, welches sowohl personenbezogene Daten als auch anonymisierte Daten braucht. Ich denke, daß dieses Forschungsprojekt auch für personenbezogene Daten und für Untersuchungen damit ein sehr deutliches Ergebnis liefert. Wer sich an die alte Diskussion erinnert, wird wissen, daß es von Wissenschaftsrichtungen und z. T. auch von politisch oder ideologisch ausgerichteten Gruppen die Argumentation gab, daß Mediziner eigentlich keine personenbezogenen Daten bräuchten, sondern, da anonymisierte Daten auch zur Zuordnung und zur Verknüpfung von Datenbeständen benutzt werden könnten, man auch auf diesem Hintergrund medizinische Forschung betreiben könne. Dieses ist mit diesem Projekt ganz sicher widerlegt. Eine solche Zusammenführung gibt es nicht, sie ist nicht sicher. Registerforschung läßt sich nicht durchführen, indem man auf dieser Ebene arbeitet. Ich halte das für eine Unterstreichung, daß es für bestimmte abgegrenzte Forschungsvorhaben nach wie vor personenbezogene Daten geben muß. Ich denke, daß auch in bezug auf die anonymisierten Daten, die genauso benötigt werden, wesentliche Fortschritte erreicht worden sind. In bezug auf die Übertragbarkeit der Ergebnisse schweben mir zwei Bereiche vor, Einen, den ich sehr global sehe: die amtliche Statistik und alles was damit zusammenhängt. Da gibt es von wissenschaftlicher Seite natürlich sehr viele Anforderungen, die nicht genau so wie in diesem Forschungsprojekt, aber doch ähnlich sind. Ein Beispiel hierfür ist die Mikrozensus-Gesundheitserhebung. Obwohl die genaue Einordnung der Diagnosen sicherlich zweifelhaft ist, kann man doch, wenn man nach Krankheitsgruppen aggregiert, einen Einblick in die Morbidität bekommen, den sonst keine andere Statistik bietet. Auch hier

wäre es notwendig, anonymisierte Individualangaben zu bekommen. Und ich hoffe, daß bei einer Prüfung tatsächlich eine Übertragbarkeit dieser Ergebnisse beispielsweise auf die Mikrozensus-Gesundheitsstatistik möglich ist. Ein zweiter Bereich, und da haben mich die Ausführungen von Herrn Nowak bedenklich gestimmt, ist die Kostenstrukturstatistik, insbesondere die Kostenstrukturstatistik der Ärzte, die, wie ich eben gehört habe, wohl eher "Unternehmensstatistiken" zugerechnet werden. Dennoch hoffe ich, daß - da ja viele Praxisinhaber auch Personen sind - sich zumindestens Teile der Forschungsergebnisse zur faktischen Anonymisierung übertragen lassen. Ein Bereich, der mir ganz wichtig erscheint sind jedoch die prozeßproduzierten Daten. Ich habe den Eindruck, daß dieses Projekt erste Anstöße gegeben hat, in welche Richtung man überlegen muß, um die Anonymisierungsproblematik bei prozeßproduzierten Daten etwas besser in den Griff zu bekommen. Sicher ist heute noch die juristische Problematik ein Hindernis sowie der Umstand, daß mit dem Sozialdatengeheimnis ein grundsätzlich anderer juristischer Tatbestand geschaffen ist als derjenige im Bundesstatistikgesetz. Das Problem, wird sich aber lösen lassen. Wie geht man jedoch bei prozeßproduzierten Daten für die faktische Anonymisierung vor, bei denen ja eine Genauigkeit der Merkmale Voraussetzung für die Verarbeitung ist und damit, wenn ich das mal so bezeichnen darf, das "Geräusch" in den Merkmalsausprägungen, die Unsicherheiten, vermutlich geringer sein werden. Wie geht man dort mit der Anonymisierung vor und worauf hat man zu achten? Ich habe den Eindruck, daß wir solche Daten brauchen werden, auch im Rahmen der Gesundheitsberichterstattung als Querschnitt, vielleicht aber auch als Längsschnitt. Man kann hierfür aus dem Projekt viel Nutzen ziehen und weiterarbeiten. Dafür bin ich Ihnen sehr dankbar. Ich denke, als eine Folgerung ergibt sich, wie häufig bei Forschungsprojekten, ein noch viel größeres Arbeitsgebiet und ich bin zuversichtlich, daß mit dem entsprechenden Elan auch daran gegangen wird.

Prof. Dr. Zimmermann, Universität München:
(...) Nun darf ich vielleicht etwas sagen zu den Interessen meines Fachbereichs, der ja hier nicht unmittelbar beteiligt war. Es ist so, in der Bundesrepublik stehen die Wirtschaftswissenschaften oder auch die Statistik an den Universitäten ganz weit hinter den Sozialwissenschaften, was die empirische Forschung anbetrifft, das auch im Gegensatz dazu wie Ökonomen international forschen und lehren. International ist es üblich, Wissenschaft in einer Verbindung aus Theorie, empirischer Analyse und statistischen Tests zu betreiben. Hier haben die Deutschen - auch in meinem Fachbereich - einen Wettbewerbsnachteil, der begründet ist in der mangelnden Verfügbarkeit von Datenmaterial, das die amtliche Statistik zur Verfügung stellt, und auch in der Tradition der Wirtschaftswissenschaften,

die keine eigenen Daten erhebt. Man mag das bedauern, aber es ist so und insoweit sind wir wirklich in einer Notlage. Und insoweit wird das, was jetzt begonnen worden ist, uns weiterhelfen. Denn zumindest die, die in der Bundesrepublik Haushaltsökonomie betreiben, werden mit Freude hören, daß die Einkommensstichprobe und der Mikrozensus, die zentralen Instrumente, hier nun mittelfristig zur Verfügung stehen werden. Das wird die Forschung mit Sicherheit deutlich befruchten, insbesondere auch vor dem Hintergrund, daß die Mikroökonometrie, die Statistik der Mikrodaten in den Wirtschaftswissenschaften, eine phantastische Blüte in den letzten Jahren erfahren hat, daß wir die methodischen Verfahren haben, um diese Daten auch adäquat zu untersuchen. Die andere Seite ist die Unternehmensstatistik. Der Ökonom möchte natürlich beide Marktseiten, die Haushalte und die Unternehmen, zur Verfügung haben. Es wäre sicherlich aus der Sicht meiner Disziplin wünschenswert, den Weg weiter zu gehen, der begonnen worden ist und ich sehe eigentlich keine prinzipiellen Probleme, warum man hier nicht weitermachen sollte. Lassen sie mich zum Schluß noch einige kurze Kommentare zu den Empfehlungen geben. Überwiegend verstehe ich, daß es diese Auflagen geben muß. Die Stichprobenlösung ist etwa für den Bereich der Statistik und Ökonometrie ein Problem, das uns nicht besonders bedrückt. Die Frage, die schwieriger ist, ist etwa die der Nutzungsbegrenzung, die zu enge Einordnung der Daten, die a priori zur Verfügung gestellt werden. Ich will auch noch einmal darauf hinweisen, daß es Sinn macht, eine Standardisierung dieses Materials vorzunehmen. Aus mehrfacher Hinsicht ist es für beide Seiten nicht nur kostengünstiger und praktikabler, sondern es erleichtert auch den Vergleich der Ergebnisse und die Forschungsergebnisse müssen vergleichbar bleiben. Insofern würde ich auch aus inhaltlicher Sicht dafür plädieren, ein Paket zu machen. Einen Datensatz zu produzieren und den bei hinreichend guten Argumenten zur Verfügung zu stellen.

Prof. Dr. Allerbeck, Universität Frankfurt:
Ich meine, daß natürlich leicht vernachlässigt wird, daß es in den Jahren, die seit den Tagungen verstrichen sind, auf die Herr Kaase eingangs Bezug nahm, natürlich hier und da Fortschritte gegeben hat. Ich denke an eine sicherlich weithin unbekannte Regelung, wie die Ausführungsbestimmungen zum Wissenschaftsparagraphen des hessischen Datenschutzgesetzes, die festlegen, daß die Anonymisierung durch den Wissenschaftler oder die wissenschaftliche Institution vorgenommen werden kann, wenn die Stelle, die die Daten abgibt, selbst dazu nicht in der Lage ist. Also es gibt schon Dinge, die außerhalb unseres Kontextes hier geschehen sind, die Wege weisen, wie man das zur allgemeinen Befriedigung machen kann. Jedenfalls habe ich nicht gehört, daß es in Hessen über diese Bestimmung zu irgendwelchen großartigen Konflikten gekommen ist. Insofern gibt es auch

für diese Identifikationsprobleme hier und da Lösungen, die auch Rechtsnormen darstellen. Auch wenn der Bereich im großen und ganzen unbefriedigend - um es sehr zurückhaltend zu sagen - geregelt ist. Für unsere Diskussion sollten wir vielleicht so vorgehen, daß wir noch eine Runde am Podium haben und dann Fragen oder Stellungnahmen aus dem Rest des Saals willkommen sind.

Prof. Dr. Müller, Universität Mannheim:
Darf ich eine kurze Bemerkung machen. Wenn ich mir die in der Diskussion vertretenen Bereiche ansehe, fällt mir auf, daß wir einen Bereich vernachlässigt haben, nämlich die empirische Sozialforschung. Und ich bedaure, daß wir dazu niemanden eingeladen haben. Wir dachten, daß dieser Bereich ohnehin bekannt ist. Aber wir haben Herrn Mochmann hier, den Geschäftsführer des Zentralarchivs für empirische Sozialforschung, der vielleicht diese Lücke füllen könnte. Entschuldigen Sie Herr Mochmann, daß ich sie nicht vorher explizit auf das Podium gebeten habe.

Ekkehard Mochmann, Zentralarchiv für empirische Sozialforschung, Köln:
Wenn hier eine besondere Legitimation, die Sozialforschung zu vertreten, gefragt ist, so darf ich mich Ihnen auch als Sprecher des Vorstandes der Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen, kurz GESIS, vorstellen. Ich möchte aufgreifen, was Herr Zimmermann schon gesagt hat, nämlich die standardisierte Bereitstellung der Daten. Aus der praktischen Erfahrung des Zentralarchivs spricht im Sinne der intersubjektiven Überprüfbarkeit sehr viel dafür, einen einmal aufbereiteten Datensatz standardisiert und nicht mit individueller Variation verfügbar zu machen. Ganz aus der Praxis gesprochen, sind jetzt Gewichtungsprobleme bei einer in Amerika aufbereiteten Version der Eurobarometer aufgetreten. Wir haben die Absicht, die Nationengewichtung in den Datensatz einzufügen und nicht nur das vorhandene Europagewicht bereitzustellen. Das führt dazu, daß der Datensatz mit anderen Recodierungen bei einzelnen Merkmalen aufbereitet wird und, wenn man nicht aufpaßt, daß in der Literatur in Zukunft zwei Datenbasen für die gleiche Behauptung oder gegenläufige Behauptungen genommen werden. Niemand weiß dann, ist es ein Analyseartefakt oder ist es ein Artefakt, das aus den Daten selbst resultiert? Ich trete hier nachdrücklich dafür ein, daß man für Datensätze einen Standard für Datensatzkennzeichnungen entwickelt, z.B. eine Internationale Datensatz Nummer (ISDN), daß man also wie bei Publikationen weiß, auf welche Ausgabe man sich bezieht. Das zweite Argument in dieser Richtung ist genauso gefallen und ich kann das nochmals aus der praktischen Erfahrung unterstreichen. Wir haben uns im Zentralarchiv Zurückhaltung auferlegt, Spezialsubsets von bestimmten Daten zu erstellen. Für bestimmte Zwecke, z.B. die Lehre, kann es sinnvoll sein, so etwas zu machen, aber in der Regel

stellen die Forscher dann in der Analysephase fest, daß sie doch noch die eine oder andere Variable brauchen, und dann hat man zwei oder drei verschiedene Datenversionen. Zur Frage der Sozialforschung: Die positive Einschätzung des Projektes, die hier wiederholt gegeben worden ist, teile ich auch. Darüber hinaus haben wir zunehmend einen Bedarf an der Verknüpfung nicht nur von Mikrodaten der amtlichen Statistik, sondern insbesondere auch von Meinungs- und Einstellungsfragen mit den regionalen Kontexten. Das ist eine europaweite Entwicklung, die ich als ganz wichtig ansehe. Ich weiß noch nicht, welche Effekte dies nun hat, weil wir ja gerade die Regionalisierungsmerkmale, ganz abgesehen von den Gemeindekennziffern, beschneiden müssen, um die Identifizierbarkeit zu reduzieren. Andererseits erscheint mir aber eine größere Genauigkeit für diesen Bereich und die Verknüpfbarkeit mit Meinungs- und Einstellungsdaten bezogen auf unterschiedliche Regionen zunehmend wichtiger.

Prof. Dr. Müller, Universität Mannheim:
Eine Frage an Herrn Nowak im Hinblick auf den Punkt, den auch Herr Brennecke schon angesprochen hat. Zwar sind nicht alle Punkte, die das Anonymisierungsprojekt untersucht hat, auf andere Datenkonstellationen übertragbar, aber kann man nicht dennoch etwas gewinnen für andere Typen von Daten, die von der amtlichen Statistik aufbereitet werden, insbesondere wenn es sich um sehr ähnliche Datenkonstellationen handelt? Ich denke zum Beispiel an die Zeitbudgetstudie, die das Statistische Bundesamt mittlerweile durchgeführt hat oder an die Beschäftigtenstatistik der Bundesanstalt für Arbeit. Für Sozialdaten gibt es allerdings Spezialregelungen, die zu berücksichtigen sind. Es gibt Testerhebungen oder Sondererhebungen, die das Statistische Bundesamt weiterhin auf der Basis von Paragraph 16 Bundesstatistikgesetz durchführen kann. Auch Daten solcher Erhebungen sind für Sozialwissenschaftler von Interesse. In der Bildungsstatistik gibt es im Grunde eine ganz ähnliche Datenkonstellation: Es gibt Individualdaten einer Reihe von Merkmalen von bestimmten Personen, die man in einer ähnlichen Weise anonymisieren könnte.

Dr. Nowak, Statistisches Bundesamt, Wiesbaden:
Da Sie mich direkt ansprechen, will ich auch direkt antworten. Meine Aussage bezog sich auf den Bereich der Wirtschaftsstatistiken. Herr Vorschulte hat zu recht gesagt, daß man auch hier prüfen müssen wird, ob sich die für die personenbezogenen Statistiken vorliegenden Untersuchungen - Mikrozensus und EVS - im Grundansatz übertragen lassen. Für die Bereiche, die Sie genannt haben und die ich im weitesten Sinne als personenbezogene Statistiken ansehen würde, glaube ich, daß man diese Grundprüfung ähnlich durchsetzen könnte, daß heißt, man wird auch hier prüfen müssen, welche Überschneidungsmerkmale vorliegen, welches Risiko

einer Deanonymisierung sich daraus entwickelt, und ob der Gedanke der Dateninkompatibilität genauso zu sehen ist, wie in den Fällen, die untersucht worden sind, ohne daß man die sehr aufwendigen Szenariotechniken nochmals nachstellen muß. Damit stellt sich auch die Frage, die Herr Brennecke schon angesprochen hat: Sind prozeßproduzierte Daten im Hinblick auf Dateninkompatibilität genauso zu betrachten wie andere, die auf unterschiedlichen Erhebungswegen eingeholt worden sind oder habe ich eine geringere Inkompatibilität, wenn ich aus der gleichen Quelle schöpfe? Ich glaube, das sind Fragen, die man noch untersuchen muß. Wir müssen dabei allerdings den Rahmen, den der Gesetzgeber mit dem Paragraph 16 des Bundesstatistikgesetztes gezogen hat, mit allen Bedingungen sehen, da können wir als Statistiker nicht einfach darüber hinwegspringen.

Bertram Raum, Bundesbeauftragter für den Datenschutz, Bonn:
Ergänzend will ich etwas zu der Möglichkeit sagen, vielleicht einmal ein anonymisiertes File zu erstellen. Ich stehe diesem Wunsch sehr skeptisch gegenüber. Grundsätzlich ist es so, daß Daten an die Forschung wie auch an andere Stellen nur weitergegeben werden können, soweit sie für diesen speziellen Zweck erforderlich sind. Bei einem public use file würde sie jedoch auch Daten bekommen, die für das spezielle Forschungsvorhaben nicht erforderlich sind. Und deshalb müssen diese Stellen sagen, diese Daten brauche ich für dieses Forschungsvorhaben. Dann besteht die Möglichkeit, daß sie diese Daten auch bekommen. Und dann zu dem Problem, das von Herrn Häfner, Herrn Brennecke und auch Herrn Allerbeck angesprochen wurde. Dieses Forschungsprojekt hier bezog sich ja nur auf statistische Daten, die in Paragraph 16 Bundesstatistikgesetz, dem Statistikgeheimnis geregelt sind. Auch Herr Heinz hat Probleme in seiner Forschung, die er auf personenbezogene Daten zurückführt. Das sind aber Sonderprobleme. Sie brauchen da keine statistischen Daten, keine Daten die dem Statistikgeheimnis unterliegen. Für die Daten, die Sie brauchen wird ja schon lange an einer Wissenschaftsklausel in der Strafprozeßordnung gearbeitet. Jüngsten Gerüchten zufolge hat man die Arbeit darüber wieder aufgegriffen. Hier gelten zunächst einmal auch nicht die Wissenschaftsklauseln in den Datenschutzgesetzen, sowohl weder das Bundesdatenschutzgesetz (innerhalb ihrer Forschungsstelle: Paragraph 40 BDSG; hinsichtlich der öffentlichen Stellen, die Daten übermitteln Paragraph 14 BDGS), noch die Datenschutzgesetze der Länder. Beispielhaft ist ja schon das sehr fortschrittliche Datenschutzgesetz von Hessen genannt worden.

Prof. Dr. Heinz, Arbeitsgruppe strafrechtliche Rechtstatsachenforschung und empirische Kriminologie, Institut für Rechtstatsachenforschung, Universität Konstanz:

Auf die von Herrn Raum angesprochene Frage der bereichsspezifischen Regelung im Referentenentwurf eines Gesetzes zur Änderung und Ergänzung des Strafverfahrensrechts (StVÄG) vom 3. Nov. 1988, durch das die Einsicht in Strafverfahrensakten für wissenschaftliche Zwecke die notwendige gesetzliche Grundlage erhalten soll, will ich nicht näher eingehen. Nur zur Erläuterung des von Herrn Raum angesprochenen Problems sei darauf hingewiesen, daß neben Befragung und Beobachtung die Dokumentenanalyse die dritte anerkannte Erhebungsmethode der empirischen Sozialforschung ist. Ein Großteil der kriminologischen Forschung beruht auf Aktenanalysen. Derzeit fehlt immer noch die gesetzliche Grundlage für Aktenauskunft und Akteneinsicht. Kriminologische Aktenanalysen sind, da die Akteneinsicht für wissenschaftliche Vorhaben acht Jahre nach dem Volkszählungsurteil des BVerfG vom 15.12.1983 schwerlich noch auf Nr. 185a RiStBV in Verbindung mit einem "Übergangsbonus" gestützt werden kann, derzeit nicht mehr möglich. Deshalb ist die kriminologische Forschung dringend auf die von Herrn Raum erwähnte Regelung im StVÄG angewiesen. Eingehen will ich vielmehr auf einen hier noch nicht angesprochenen Effekt des hier vorgestellten Anonymisierungsprojekts: Ich befürchte, daß dieses Projekt einen unerwünschten Nebeneffekt haben könnte. Diesen unerwünschten Nebeneffekt sehe ich darin, daß nunmehr angenommen wird, Paragraph 16 Absatz 6 BStatG sei eine Regelung, mit der die Forschungsbedürfnisse vieler Wissenschaftszweige hinreichend befriedigt seien, weil die Probleme der Verhältnismäßigkeit usw. jetzt geklärt seien, so daß die sozialwissenschaftliche Forschung mit faktisch anonymisierten Daten arbeiten könne. Dies ist jedoch nicht für alle sozialwissenschaftliche Disziplinen der Fall. Für die Kriminologie jedenfalls gilt, daß sie auf die personenbezogenen Daten der Strafverfolgungsstatistik angewiesen ist, wie ich bereits am Beispiel der Stichprobenziehung und der Forschungen zur Strafzumessungspraxis zu zeigen versucht habe. Alternativen zu den Daten der Strafverfolgungsstatistik gibt es derzeit - und auf absehbare Sicht - nicht. Die Eintragungen im Bundeszentralregister enthalten z.B. keine Freisprüche, auch die Mehrzahl der Einstellungen werden nicht eingetragen. Ob das geplante "länderübergreifende staatsanwaltschaftliche Informationssystem" die erforderlichen Angaben enthalten wird, steht noch nicht fest. Würde z.B. der Arbeitsentwurf "Strafverfolgungsstatistikgesetz" vom 30.5.1989 in Kraft treten, dann würde Paragraph 16 Abs. 6 BStatG auch für die Kriminologie gelten und damit das Ende für einen nicht unerheblichen Teil kriminologischer Forschung bedeuten, insbesondere für den rechtspolitisch besonders wichtigen Bereich der Sanktions- und Wirkungsforschung. Möglich wären dann vielleicht noch Untersuchungen zur Sanktionspraxis von Massendelikten, wie z.B. "Diebstahl geringwertiger Sachen" in ausgewählten, sehr großen Landgerichtsbezirken. Nicht mehr möglich wären dagegen Untersuchungen in den rechtspolitisch besonders

wichtigen Fallgruppen, etwa der Gewaltkriminalität, weil in diesen relativ seltenen Ereignissen die Einzelangaben, die strafzumessungsrelevant sind, nicht übermittelt werden dürfen. Paragraph 16 Abs. 6 BStatG bedarf deshalb der Ergänzung durch den Gesetzgeber, entweder durch eine weitere Wissenschaftsklausel, in der die Belange der von mir angesprochenen Wissenschaftszweige im Sinne einer "praktischen Konkordanz" von Wissenschaftsfreiheit und dem Recht auf informationelle Selbstbestimmung geregelt werden, oder durch eine bereichsspezifische Regelung in den speziellen Statistikgesetzen, also z.B. im Strafverfolgungsstatistikgesetz. Insoweit verweise ich z.B. auf Paragraph 40 Abs. 2 Bundeszentralregistergesetz. Nur durch eine derartige Regelung wird gewährleistet, daß auch künftig kriminologische Forschung die Daten der amtlichen Statistik für die Durchführung von Forschung, sei es als unmittelbarer Gegenstand, sei es für die Planung und Durchführung von Aktenanalysen (Stichprobenziehung) wird nutzen können. Für viele sozialwissenschaftliche Disziplinen wird der hier vorgestellte Weg der faktischen Anonymisierung gangbar und ausreichend sein. Ich plädiere aber nochmals dringend dafür, daß wissenschaftlichen Forschungsvorhaben, die auf personenbezogene Daten angewiesen sind, auch künftig diese Daten zur Verfügung gestellt werden. Aufgabe und selbstverständliche Verpflichtung der in diesen Bereichen tätigen Forscher ist es, einerseits durch die erforderlichen technisch-organisatorischen Maßnahmen und andererseits durch Anonymisierung der Daten zum frühestmöglichen Zeitpunkt sicherzustellen, daß der einzelne durch den Umgang mit seinen Daten in seinem Persönlichkeitsrecht nicht beeinträchtigt wird.

Prof. Dr. Brennecke, Freie Universität Berlin:
Vielen Dank, daß Sie das nochmal gesagt haben. Das ist auch in meinem Sinn. Ich möchte in bezug auf das, was Herr Mochmann und Sie über das public use file gesagt haben, noch etwas hinzufügen. Ich denke, der Begriff public use file ist sehr unglücklich. Ganz sicher wird es nicht so aussehen - und so stelle ich mir das auch nicht vor -, daß das Statistische Bundesamt ein wie auch immer geartetes File an X gibt und von X an Y und Z usw. Aber die Erfahrung der Auswertung von Mikrodaten hat doch eigentlich gezeigt, und das ist implizit heute auch angeklungen, daß man im Prinzip alle Möglichkeiten eines solchen Datensatzes nutzen möchte. Im Gegenteil, meistens ist die Situation so, daß man sich an vielen Stellen fragt, ob es nicht gut wäre, wenn dieses oder jenes Merkmal noch im Datensatz wäre. Ich vermute deshalb, daß die Anforderungen, die von verschiedenen Seiten bezüglich des Mikrozensus kommen werden, in bezug auf die Merkmale und Variablen sehr ähnlich sein werden. Und ich vermute deshalb, daß es sich von seiten des Statistischen Bundesamtes lohnen wird, diesen Gedanken vorzugreifen und für alle jene Fälle, die mit ähnlicher und begründeter

Einzelanforderung kommen, aber in der Struktur gleich sind, ein einheitliches File zu konstruieren. Das würde unter Umständen den Arbeitsaufwand minimieren und vielleicht auch den Forderungen oder dem Wunsch der Forscher relativ nahe kommen.

Prof. Dr. Dr. Häfner, Zentralinstitut für seelische Gesundheit, Mannheim:
Wenn wir Themen angeschnitten haben, die weit über den Paragraph 16 des Statistikgesetzes hinausgehen und, was mich angeht, nicht nur das Datenschutzgesetz, sondern auch den Paragraph 203 des Standesrechts zum Inhalt hatten, dann hat das damit zu tun, daß man am Ende einer solchen Tagung ein wenig das Recht fühlt, sich zu fragen, ob es über den begrenzten direkten Effekt für mein Fach auch eventuell einen indirekten Effekt gibt. Auch da ist natürlich das Problem - das versuchte ich deutlich zu machen -, daß die Bedürfnisse der Forschung auf meinem Gebiet so sind, daß ich nicht nur de lege lata sondern auch de lege ferenda denken oder diskutieren muß. Und ein zweiter Punkt, der noch ein bißchen weiter weg geht vom heutigen Thema, den ich aber auch mit angeschnitten habe, ist die Fragmentierung der Gesundheitsstatistiken in der Bundesrepublik. Wir haben mit diesem Problem solche negative Folgen für die Forschung, daß wir im Grunde genommen Public-Health Forschung und Epidemiologie nur in Grenzbereichen führen können. Es ist nun leider nicht so, daß uns die Regelungen, wie etwa Anonymisierung, die den Zugang zu Daten einzelner Statistiken erleichtern würden, viel helfen würden. Denn die unterschiedlichen Institutionen die Daten sammeln und zur Verfügung stellen können, haben außerdem noch unterschiedliche Variablendefinitionen, unterschiedliche Erhebungsmethoden, aber auch unterschiedliche Grundgesamtheiten. Wenn ich beispielsweise die Arbeitslosenstatistik vergleichen will mit der Mortalitätsstatistik des Statistischen Bundesamtes, um zu prüfen, ob Arbeitsplatzverlust oder längerfristige Arbeitslosigkeit Suizidraten erhöht, entsteht das Problem, daß ich unterschiedliche Grundgesamtheiten habe, die ich im Grunde nicht vergleichen kann. Das Problem bleibt also, daß man die Ausgangssitutation für Gesundheitsdaten in diesem Land verbessern muß.

Dr. Schmidt, Bundesbeauftragter für den Datenschutz, Bonn:
Ich möchte etwas sagen zu der Problematik, die aus meiner Sicht mit diesem Projekt kaum etwas zu tun hat. Bei den Verlaufsstatistiken, die wir natürlich vor allen Dingen da brauchen, wo es auf die Folgen von Maßnahmen ankommt, wie zum Beispiel bei Verurteilungen, ist sicher eine Anonymisierung, wie heute in diesem Projekt vorgestellt, überhaupt nicht denkbar. Die Richtung, die zur Zeit wohl sehr konsequent gedacht ist, ist die, daß man Verlaufsregister mit einem Anonymisierungstreuhänder schafft, das scheint mir ein aussichtsreicher Weg. Und ich glaube, das Hindernis liegt hier nicht

im Bereich der Datenschützer oder der Datenschutzbeauftragten. Ganz konkret: es gab zum Beispiel in der DDR Register, die unter diesen Umständen hier nie geführt worden wären, die aber gleichwohl für die Forschung auch jetzt noch von erheblicher Bedeutung sein können. Hier geht die Bereitschaft der Datenschutzbeauftragten, über die Wahrung der Interessen der Betroffenen entsprechende Forschung möglich zu machen, sicher sehr weit und ich würde auch mal sagen, weiter als die Bereitschaft mancher, die jetzt das Dach über dem Kopf und den Pförtner bezahlen müßten, damit die Daten auch aufbewahrt werden können. Also da sieht sich zumindest die Datenschutzseite von dem Vorwurf frei, hier ein Hindernis zu bieten. Es sind andere, die da bremsen. Zum public use file möchte ich sagen, dies ist eine Geschichte, die sollte man nach Möglichkeit nicht mit diesem Projekt verknüpfen. Man könnte die Ergebnisse dieses Projekts überstrapazieren, wenn man damit allgemein benutzbare - und sei es nur in der Forschung allgemein benutzbare - Standarddatensätze machen würde. Dann hätte man nämlich vermutlich ein ganz neues Angriffsszenario. Die Interessen, auch die publizistischen Interessen, eine definierte Datenbasis als verletzbar hinzustellen, könnten zu Zusammenarbeiten führen, die ein ganz anderes Szenario aufzeigt. Deshalb meine ich, daß man zumindest mit den jetzt aufgestellten Ergebnissen gut daran tut, weiterhin Forschungsvorhaben bezogenene Weitergaben zu erlauben, die vernünftigerweise aus einer konkreten, dafür geeigneten Basis mit einfachen Mitteln gezogen werden können. Zu denen man dann vielleicht auch leichter als das bisher möglich war, auch mal noch ein Datenfeld nachliefern kann. Aber zu sagen, jeder vernünftig scheinende Antrag bekommt die selben Daten, dafür scheinen mit die Ergebnisse dieses Projektes nicht ausreichend tragfähig. Wir tun deshalb gut daran, zunächst die Erfahrungen mit der individuellen Prüfung abzuwarten und zu sehen, wie das ausgeht.

Dr. Nowak, Statistisches Bundesamt, Wiesbaden:
Ich möchte mich der Meinung anschließen, daß das Wort "public use" für unsere Diskussion irreführend ist. Faktisch anonymisiertes Einzelmaterial ist nach den Bestimmungen des Gesetzgebers nur für einen sozusagen privilegierten Empfängerkreis und nicht für die Öffentlichkeit zugänglich. Aber es ist hier auch - und ich erinnere an die Aussage von Frau Marsh heute Mittag - gesagt worden, Kernpunkt der Akzeptanz ist das Vertrauen. Dem Auskunftgebenden für die Statistik reicht sozusagen das subjektive Mißtrauen, um der Statistik die Mitarbeit aufzukündigen. Das ist etwas sehr Sensibles, woran wir denken müssen. Was Herr Schmidt hier aufführt ist ein Angriffsszenario, das man unter diesem Blickwinkel sehr ernst nehmen muß. Erlauben Sie mir noch eine Anmerkung zu dem, was Herr Heinz gesagt hat. Wir sollten nicht nur gebannt auf den Paragraph 16 Absatz 6 des Bundesstatistikgesetzes sehen. Der Paragraph 16 sagt ja ausdrücklich, daß

der Gesetzgeber im begründeten Fall auch andere Formen der Weiterleitung von Einzelangaben im jeweiligen Gesetz zulassen kann. Er muß es dort aber regeln. Und daher müssen diejenigen, die solche Angaben benötigen an den Gesetzgeber appellieren und nicht nachträglich den Statistikern sagen, nun gebt mir endlich die Daten, auch wenn es im Gesetz so nicht erlaubt ist. Dasselbe gilt auch für die Frage, darf denn ein Wissenschaftler Daten an einen anderen weitergeben. Nachdem wie ich den Paragraph 16 des Bundesstatistikgesetztes lese, steht im Abschnitt 8 genau das Gegenteil. Da steht, daß er es nicht darf. Und letztlich sollten wir auch deshalb hier nicht so gebannt nur auf den Paragraph 16 Absatz 6 des Bundesstatistikgesetztes sehen, weil die Leistung der amtlichen Statistik nicht nur die Übermittlung von Einzelangaben ist - anonymisiert oder wie auch immer - sondern die Übermittlung von statistischen Ergebnissen in Aggregatform. Wir sollten nicht vergessen, daß es dafür auch ganz gute Zwecke gibt und man damit auch ganz gut arbeiten kann. Ein allerletztes Wort zu dem, was Herr Häfner sagte: Eine der Grundlagen für eine fachliche Konzentration der Arbeit der amtlichen Statistik ist die Koordinierungsfunktion, damit Sie eben nicht immer wieder vor unterschiedlichen Statistiken stehen, deren Ergebnisse vielleicht als Insellösung isoliert ganz brauchbar sind, die aber nicht mit anderen kombiniert verwendet werden können.

Prof. Dr. Allerbeck, Universität Frankfurt:
Da das Stichwort public use drei- oder viermal gefallen ist, können sie vielleicht verstehen, daß ich mich selbst auf die Rednerliste setze. Zunächst ist einmal klar, daß unter dem Absatz des Paragraphen, mit dem wir uns die ganze Zeit beschäftigt haben, ein public use file nicht möglich ist. Das ist vollständig unstrittig. Der Aspekt, den Herr Mochmann angesprochen hat ist, daß die Leute, die Daten haben oder mit Daten arbeiten, diese ja fortlaufend immer ein klein wenig modifizieren, weil sie etwas merken, was nicht in Ordnung ist, was besser sein könnte. Die Befürchtung, die da angesprochen worden ist, und die ich teile ist, daß das Statistische Bundesamt, so lange die Daten im Hause sind an den Daten schneidert und sicherlich die eine oder andere Korrektur vornehmen wird. Dieses maßgeschneiderte Verfahren der Variablenauswahl wäre für mich als Wissenschaftler nur dann akzeptabel - und jetzt ist das Stichwort nicht Publizistik sondern einfach Wissenschaft - wenn lediglich Variablen eliminiert würden und sonst nichts mehr mit dem Datensatz geschieht sobald er für fertig erklärt ist. Es darf dann nichts mehr geändert werden. Sonst ist dieses Verfahren der Maßschneiderung unbrauchbar und für die Wissenschaft einfach nicht akzeptabel. Was nun die Szenarien angeht, so wäre manche Diskussion vor der letzten Volkszählung bei uns sehr viel einfacher gewesen, wenn nicht die Informatikdiplomanden sondern unsere Studenten Zugang zu einem public use file gehabt hätten oder einem subset

eines public use file. Weil sie dann nämlich gewußt hätten, was in dieser Volkszählung typischerweise vorkommt. Und das hätte der Statistik nicht geschadet, sondern definitiv geholfen. Wenn man in Ihrem Amt ist, Herr Schmidt, dann bekommt man natürlich immer Alarmmeldungen aus aller Welt. Und hat von diesen einen reichen Vorrat und stellt sich dann vor, was könnte noch dazu kommen. Man könnte sich natürlich auch Strategien der amtlichen Statistik vorstellen, ihr Veröffentlichungsprogramm über das Erscheinen des statistischen Jahrbuchs zu verbreitern. Und eine Form dieser Verbreiterung wäre die Publikation von Daten in einer zeitgenössischen Form.

Dr. Vorschulte, Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen, Düsseldorf:
Dazu kann ich gleich sagen, daß wir das jedes Jahr versuchen. Wir haben schon im letzten Jahr eine CD-Rom mit den Ergebnissen der Volkszählung auf kleinster räumlicher Ebene erstellt. Natürlich nur für Nordrhein-Westfalen. Und wir werden in diesem Jahr eine Diskette vorlegen mit Ergebnissen nach dem Bundesraumordnungsprogramm. Für die Herrichtung wird sicherlich vieles getan und da bemühen wir uns sehr, und im übrigen bleibt ja noch festzustellen, daß jeder Forscher die Möglichkeit hat, sich an das jeweilige Statistische Landesamt oder das Statistische Bundesamt zu wenden. Dort sind alle Datensätze vorhanden und bestimmte Probleme kann man dort abarbeiten lassen. Das hat ja Herr Mochmann heute Morgen als Normalfall für England, Dänemark oder Schweden vorgestellt, daß zentral gerechnet wird. Dort werden die Dinge nicht aus dem Haus gegeben, sondern die Probleme werden besprochen und dann wird zentral gerechnet. Das ist natürlich ein Angebot, das die amtliche Statistik hier in der Bundesrepublik auch immer gemacht hat. Wir jedenfalls in Nordrhein-Westfalen wären dazu jederzeit bereit.

Prof. Dr. Allerbeck, Universität Frankfurt:
Das Angebot ist natürlich in der Praxis die Aufforderung, für das Statistische Bundesamt einen Blankoscheck auszustellen. Der heißt natürlich nicht so, sondern der heißt Kostenübernahmeerklärung. Wir sind weit fortgeschritten in der Zeit, vielleicht sollte Herr Müller noch die Gelegenheit bekommen, ein abschließendes Wort in dieser Diskussion anzufügen.

Prof. Dr. Müller, Universität Mannheim:
Ja, gerne. Als erstes möchte ich sagen, daß ich diese Diskussion nicht angeregt habe, damit wir soviel Lob bekommen - worüber ich mich natürlich gefreut habe -, es hätte ja auch anders ausgehen können. Sondern es ging mir darum, Vertreter aus den verschiedenen Bereichen zu Wort kommen zu lassen, genau deshalb, weil die Wissenschaft ein sehr differenziertes

Unternehmen ist und weil klar ist, daß nicht zu allen Zwecken die gleichen Mittel taugen. Wenn man faktische Anonymität diskutiert und sieht, daß es damit für bestimmte Bereiche nun einen Schritt weitergeht, muß man gleichzeitig sagen, daß dieses nicht alle anderen Probleme löst, die wir haben. Deshalb waren die Ergänzungen, caveats und die Hinweise darauf, wie zu den verschiedenen Bereichen Problemlösungen aussehen müßten, ganz in meinem Sinn. Das heißt auch, man möge dieses Projekt und seine Ergebnisse nicht überfrachten und sie nicht auch für Bereiche nutzen wollen, für die sie nicht primär gedacht waren. Deshalb ist wichtig, was zum public use file gesagt worden ist. Man muß da sehr vorsichtig sein. Man soll die Ergebnisse des Anonymisierungsprojektes zunächst nur dazu nehmen, wozu das Projekt gedacht worden ist. Zum Abschluß möchte ich noch sagen, daß ich den heutigen Tag so wahrgenommen habe, wie das ganze Projekt. Ich fand die Zusammenarbeit zwischen den verschiedenen Stellen - zwischen Statistik, Wissenschaft und Datenschutz - überaus fruchtbar. Und wenn wir heute eine vergleichsweise unverkrampfte Diskussion gehabt haben und eine Diskussion, aus der man auch etwas lernen kann, war das zu Beginn des Projektes nicht so. Ich weiß noch, wie wir uns zunächst sehr vorsichtig begegnet sind. Hinter jedem Argument haben wir gewissermaßen eine verdeckte Attacke vermutet. Das hat sich im Laufe der Zeit jedoch in großem Maße verändert. Es liegt mir sehr daran, wenn dieses Projekt jetzt gelobt worden ist, dies wirklich an alle Beteiligten weiterzugeben: Neben den beteiligten Wissenschaftlern an den Datenschutz und die amtliche Statistik, die in vielfältiger Weise mitgewirkt haben. Sie haben nicht nur die Daten zur Verfügung gestellt, sondern auch unkonventiolle Lösungen ermöglicht. Ich habe zum Beispiel noch die lange Diskussion in Erinnerung, bei der wir überlegt haben, wie denn die faktische Anonymität mit Daten geprüft werden kann, die eigentlich nur als faktisch anonym weitergegeben werden dürfen. Daß hier Lösungen gefunden worden sind, ist das Verdienst aller, die mitgewirkt haben. Ich kann heute Abend nur einen großen Dank aussprechen und dabei auch alle einschließen, die sich heute als Referenten und Teilnehmer am Podiumsgespräch zur Verfügung gestellt haben und die dazu beigetragen haben, daß wir einen sehr interessanten Tag verbringen konnten.