„Data Sharing“

Mike Kühne, Dirk Meusel

Deutsche Zusammenfassung

Das offene Teilen von Informationen und Forschungsdaten und Kooperation unter Wissenschaftlern entspricht einem wissenschaftlichen Ideal. Der Forschungsalltag wir diesen Prämissen allerdings nicht immer gerecht. Auf der Basis einer Inhaltsanalyse wird die Praxis des Teilens wissenschaftlicher Daten untersucht. Die Art und Weise des Teilens wissenschaftlicher Forschungsdaten unterscheidet sich zwischen Wissenschaftsbereichen. Die Gründe für das Verweigern sind unterschiedlicher Natur und reichen von datenschutzrechtlichen Bedenken bis hin zu urheberechtlichen Einwänden. Eine Weiterentwicklung von Regeln zum Datenaustausch scheint auf unterschiedlichen Ebenen sinnvoll. Zentrale Datenarchive und die Etablierung von Konventionen zum Datenaustausch sind wichtige Möglichkeiten, die Vorteile des Datenaustausches besser zu nutzen.

Englische Zusammenfassung

The open sharing of information and research data as well as the free cooperation between researchers corresponds to the ideal of science. Nevertheless, the everyday scientific routine does not always comply with this ideal. Using a content analysis, the praxis of data sharing among researchers is being examined. The way of sharing research data differs between various disciplines of science. Reasons for withholding research data are of different nature and stretch from reservations related to data protection laws to objections related to authorship rights. A further development and implementation of guidelines for scientific data sharing seems to be appropriate on various levels. The use of centralised data archives and the definition of conventions for scientific data sharing are seen as important options to further make use of the multiple advantages of the scientific ideal of data sharing.

I. Einleitung

In der Januarausgabe 2003 des JAMA stellten Campbell und Kollegen (2002) einen Gegenstand in den Mittelpunkt ihrer Arbeit, welcher bis dahin wenig wissenschaftliche Aufmerksamkeit genoss. In einer Feldstudie befragten sie 3000 Kollegen an 100 Universitäten der USA, inwieweit diese in den letzten drei Jahren in der Lage waren, publizierte wissenschaftliche Ergebnisse der Biowissenschaften anhand der darin beschriebenen Originaldaten zu replizieren. Ihre Grundargumentation dabei war so einfach wie weithin akzeptiert: das freie und offene Teilen von Informationen, Forschungsdaten und sonstiger Materialien hinsichtlich publizierter Forschung ist essentiell für die Replikation publizierter Ergebnisse sowie die Effizienz des wissenschaftlichen Fortschritts. Ihr Forschungsinteresse galt kurz und präzise dem Verständnis von Natur, Ausmaß und Konsequenzen des Zurückhaltens von Forschungsdaten unter den befragten Wissenschaftlern (vgl. Campbell 2002).

In der vorliegenden Studie sollte überprüft werden, ob sich ähnliche Befunde und Ergebnisse in hiesigen Wissenschaftskreisen finden lassen. Dabei sollte überprüft werden, ob sich Anhaltspunkte für Unterschiede zwischen englischsprachigen und deutschsprachigen Forschungspraktiken sowie Unterschiede zwischen verschiedenen Wissenschaftsbereichen erkennen lassen.

1 Hintergrund

Die grundlegende Diskussion um die Bedeutung von Forschungsdaten ist weder neu noch unhinterfragt. Bereits 1973 stellte Robert Merton seinen Ethos der Wissenschaften auf, welcher „ … die funktional notwendige Erfordernis [involviert], dass Theorien oder Verallgemeinerungen in Hinblick auf ihre logische Konsistenz sowie ihre Konsonanz mit Fakten evaluierbar sein sollen“ (Merton 1973). Auch Karl Popper stellt in einem seiner Hauptwerke das Falsifikationsprinzip als Grundlage aller wissenschaftlichen Theoriebildung dar. Kein wissenschaftliches System und keine wissenschaftliche Aussage kann demnach absolute Gültigkeit beanspruchen, es hat als Arbeitshypothese lediglich vorläufigen Modellcharakter (Popper, 1984 (1934)). Michael Polanyi geht einen Schritt weiter, indem er die Wichtigkeit des Datenaustauschs unter Wissenschaftlern betont: „Ohne den freien Austausch publizierter wissenschaftlicher Informationen und ihrer Ressourcen könnten Forscher unwissentlich auf etwas aufbauen, was weit weniger als die totale Akkumulation wissenschaftlicher Erkenntnisse ist oder an Problemen arbeiten, welche bereits gelöst sind“ (Polanyi 1962).

Heute sind große maschinenlesbare Datensätze für die Sozial- und Gesundheitswissenschaften die größte Forschungsressource. Analog zu den reinen Naturwissenschaften können sie in ihrer Bedeutung für die Forschung mit großen Instrumentarien und Laboren für Physik oder Chemie verglichen werden (Fienberg et al. 1985: 19). Aussagen über die Prävalenz einer Krankheit oder den Erfolg bestimmter Interventionsstrategien in der Bevölkerung können nur über statistisch aussagekräftige Daten getroffen

werden. Nur selten kann man verlässliche Erkenntnisse in diesem Gebiet über Tests in Laboren gewinnen. Letztere gehören eher der Grundlagenforschung in den Basisdisziplinen an.

Auch die empirische Sozialforschung bestritt Mitte der siebziger Jahre einen Wechsel im Forschungsparadigma, wobei die Bedeutung von vereinzelten Datenerhebungen singulärer Forschergruppen geschmälert wurde. Stattdessen wurden wissenschaftsöffentliche, repräsentative und regelmäßige Erhebungen für Sekundäranalysen in Zusammenarbeit mit den statistischen Ämtern entwickelt, welche später zu *Datenreports* und *Public Use Files* führten. Beispiele dafür sind die *Allgemeine Bevölkerungsumfrage der Sozialwissenschaften* (ALLBUS), das *International Social Survey Programme* (ISSP) sowie die retrospektiven Kohortenstudien am Max-Plank-Institut für Bildungsforschung. Durch diesen Paradigmenwechsel konnten Forschungsdaten höherer Qualität erzielt werden, welche zudem international vergleichbar und auf Dauer gestellt sind (Diekmann 2002: 45). Ein großer Teil empirischer Studien, vor allem im Bereich weniger globaler Fragestellungen, muss und wird dennoch von kleineren Forschergruppen mittels eigener Datenerhebung durchgeführt. Oft erscheint es unmöglich, Ergebnisse solcher Studien dem oben zitierten Idealbild freier Replizierbarkeit zugänglich zu machen. Knappe Zeit- bzw. Finanzbudgets sowie eine hohe Fluktuationsrate im wissenschaftlichen Personal vermindern die Umsetzungsmöglichkeiten einer idealen Datendokumentation, vor allem in drittmittelgeförderter Forschung.

Allerdings existiert bereits eine Reihe von Datenarchiven, die der Öffentlichkeit zugänglich sind. Sammlungen von Forschungsdaten bestehen auf unterschiedlichen Ebenen. In den Sozialwissenschaften besteht in Form des Zentralarchivs für Empirische Sozialforschung an der Universität zu Köln eine zentrale Institution, die Primärmaterial und Ergebnisse empirischer Untersuchungen aufbereitet und der Öffentlichkeit zugänglich gemacht.

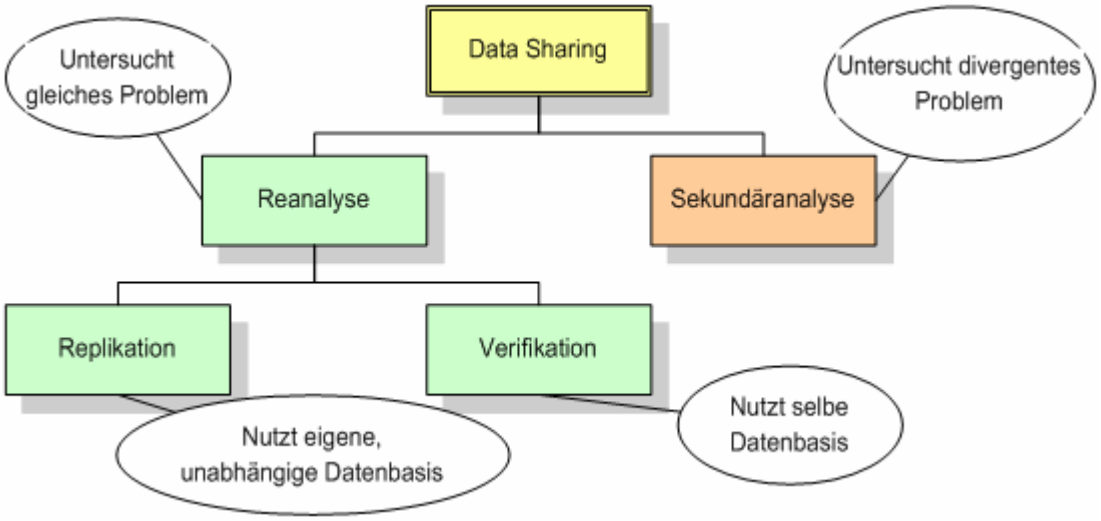
Die Zugänglichkeit von Forschungsergebnissen ist in vielfältiger Weise festgeschrieben. In den einzelnen Hochschulgesetzen der Bundesländer existieren Paragraphen, welche die Veröffentlichung von Forschungsergebnissen fordern. In fachspezifischen Gremien werden Kodizes, wie zum Beispiel der Ethik-Kodex der Deutschen Gesellschaft für Soziologie und des Berufsverbandes Deutscher Soziologen entwickelt, die Wissensproduktion, -verwertung und vor allem -weitergabe regeln sollen. Trotzdem scheinen zwischen Anspruch und Wirklichkeit größere Differenzen zu herrschen (Schnell 2002). Die bereits angesprochenen Studien zu diesem Thema haben gezeigt, dass die Routine im Wissenschaftsbetrieb den Idealen des Datenaustauschs nicht gerecht wird.

Die Literatur zum Thema der Weitergabe wissenschaftlicher Daten enthält aufgrund der geschilderten Diskrepanz Hinweise auf eine gesteigerte soziale Erwünschtheit in den durchgeführten Befragungsstudien. Aufgrund dieser Problemlage wurde in dieser Studie ein alternativer Weg der Datenerhebung genutzt.

2 *Begriffe und Abgrenzung*

Die Weitergabe wissenschaftlicher Forschungsdaten kann nicht nur institutionell organisiert stattfinden (wie oben beschrieben), sondern kann auch direkt von Forscher zu Forscher vollzogen werden. Dafür wurde in der englischsprachigen Literatur der Begriff des *Data Sharing* (das gemeinsame Nutzen von Forschungsdaten) eingeführt. Um Vorteile wie auch Nachteile der Aufbereitung wissenschaftsöffentlicher Datensätze sowie des D*ata Sharing* darlegen zu können, ist eine definitorische Abgrenzung von vier Begriffen notwendig, welche den Zweck der Verwendung solcher Daten beschreiben: (1) die *Reanalyse*; (2) die *Replikation*; (3) die *Verifikation*; und (4) die *Sekundäranalyse*.

Abbildung 1: Systematik des Data Sharing



Fienberg et al. (1985: 9) bieten für diese Begriffe eine zweckmäßige Definition. (1) Eine *Reanalyse* untersucht das gleiche Problem oder die gleiche Fragestellung, welche durch die ursprüngliche Forschergruppe schon untersucht worden ist. Dabei kann die gleiche Datenbasis verwendet werden, sowie auch eine andere. Ziel der Reanalyse ist die Überprüfung der Reliabilität eines Ergebnisses oder einer Aussage der originalen Forschungsarbeit. (2) Wenn von der Originalstudie abweichende, unabhängig erhobene Daten verwendet werden, um ein und dasselbe Problem zu untersuchen, dann wird die Reanalyse *Replikation* genannt. (3) Wenn die gleiche Datenbasis der Originalstudie verwendet wird, um ein und dasselbe Problem zu untersuchen, dann wird die Reanalyse *Verifikation* genannt. (4) In einer *Sekundäranalyse* hingegen werden Daten, welche für ein bestimmtes Set von Fragestellungen gesammelt worden sind, für die Beantwortung eines anderen Sets von Fragestellungen verwendet. Meistens, aber nicht zwingend, werden für die Sekundäranalyse Public Use Files allgemeiner Bevölkerungsumfragen verwendet, welche als Mehrzweckumfragen von vornherein

geplant sind. Abbildung 1 illustriert die Beziehungen der Begriffe untereinander. In der folgenden Diskussion von Vorteilen und Nachteilen wird auf diese in ihrer hier ausgeführten Bedeutung zurückgegriffen.

3 Vorteile des Data Sharing

Die Vorteile eines avancierten *Data Sharing* sind vielfältig und reichen von einer gesteigerten Transparenz und Replizierbarkeit publizierter Forschungsergebnisse, über die Möglichkeit von Sekundäranalysen bis hin zur Verwendung der Forschungsdaten als realistisches Studienmaterial zur Einführung in die statistische Analyse. Da von den Autoren dieses Beitrages die Erstellung von Public Use Files kleinerer Datensätze als ein Mittel zur Verbesserung des *Data Sharing* gesehen wird (Meusel et a 2005), wird im Folgenden die Begriffe *Data Sharing* und *Aufbereitung von Public Use Files* in gleicher Absicht verwendet.

(1) *Replizierbarkeit* und *Transparenz* sind die Grundpfeiler einer guten Wissenschaftspraxis (Hall 2000). Um wissenschaftliche Ergebnisse nachvollziehen zu können, leitet sich die Notwendigkeit der sorgfältigen Beschreibung der Forschungsprozedur ab. King (1995) führt dazu erweiternd aus, dass eine empirische Analyse nur dann voll verstanden und evaluiert werden kann, wenn eine vollständige Information über den Prozess der Datengenerierung und Datenanalyse vorliegt. Auch wenn dieser Umstand in allen empirisch arbeitenden Wissenschaftsdisziplinen weithin akzeptiert ist, fehlen oftmals Strategien zur praktischen Umsetzung. Schnell (2002) berichtet unter anderem von einer schlechten Replizierbarkeit deutschsprachiger methodologischer Arbeiten auf dem Gebiet der Sozialwissenschaften, weil kaum eine Publikation explizit den Datensatz oder gar das zur Analyse verwendete Setup zur Verfügung stellt.

Die Aufbereitung von Forschungsdatensätzen für die wissenschaftliche Öffentlichkeit in Form von Public Use Files kann diesem Problem entgegenwirken. Dies fordert neben anderen auch Diekmann (2002: 48): „Wir müssen ... die Fehlerkontrolle und Replikationsmöglichkeiten verbessern, z. B. indem Fachzeitschriften fordern, dass Datensätze in wohldokumentierter Form für Reanalysen verfügbar gemacht oder am besten gleich ins Internet gestellt werden“. Durch *Data Sharing* wird Forschung nachvollziehbar, indem die Argumentation des originalen Forscherteams nachvollzogen werden kann.

Diesen Punkt einen Schritt weiter zu denken bedeutet, dass die Dissemination wissenschaftsöffentlicher Datensätze die *Verifikation*, *Widerlegung* und *Verfeinerung* ursprünglicher Forschungsergebnisse fördert. Auf der Grundlage einmal erstellter Public Use Files könnten publizierte Ergebnisse direkt nachgerechnet und auf ihre Richtigkeit hin verifiziert werden. Zusätzliche oder alternative Analysen dieser Datensätze könnten die Robustheit der originalen

Schlussfolgerungen überprüfen bzw. diese unter verschiedenen Annahmen weiterführend testen. Diese Strategie kann die ursprünglichen Ergebnisse einer Studie stärken und einer breiteren Akzeptanz zuführen. Andererseits könnten aber Fehler der ursprünglichen Studie oder Inkonsistenzen in der Datenbasis entdeckt werden, welche die Validität der Ergebnisse in Zweifel ziehen. Letztendlich könnte aber auch die Verfeinerung der ursprünglichen Ergebnisse einer Studie das Resultat der Reanalyse bilden. Reanalysen der ursprünglichen Datenbasis unter geänderten Grundannahmen könnten beispielsweise eine verbesserte Effektstärke zum Vorschein bringen und somit die ursprünglichen Ergebnisse unterstreichen (Fienberg et al. 1985: 10).

(2) Die *Stimulierung neuer Forschungsfragen* kann durch existierende Forschungsdaten angeregt werden. Reanalysen können neben den unter (1) angesprochenen Punkten auch Tests der Allgemeingültigkeit von Forschungsergebnissen beinhalten. Zur Verallgemeinerung von partikulären Forschungsergebnissen benötigt man den Vergleich verschiedener Datensätze aus unterschiedlichen Quellen, welche Aussagen über Zeit- und Ortsgrenzen hinaus zulassen. Public Use Files tragen demzufolge dazu bei, Datensätze verfügbar zu halten und diese für Allgemeingültigkeitstests zugänglich zu machen.

Des Weiteren können mehrere vorhandene Datensätze untereinander „verlinkt“ werden, um somit eine *neue vergrößerte Datenbasis* zu schaffen. Diese könnte wiederum die Beantwortung neuer Forschungsfragen oder das Testen neuer Hypothesen ermöglichen. Anwendungsgebiete sind vor allem Längsschnittstudien über mehrere Jahre, wobei Daten fehlender Jahrgänge unter Umständen durch bestehende Public Use Files ersetzt werden könnten. Bei diesem Vorgehen sind umfangreiche Voraussetzungen der Datenbasis zu gewährleisten, wobei vor allem soziodemographische Informationen vergleichbar sein müssen. Die flächendeckende Implementierung der in den Sozialwissenschaften akzeptierten Standarddemographie (Statistisches Bundesamt 2004) sollte hierfür forciert werden.

(3) Die Evaluierung von Forschungsdesign und der daraus entstandenen Datenbasis durch Wissenschaftler, welche von den ursprünglichen Forschern verschieden sind, kann zur *Verbesserung von Mess- und Datenerhebungsmethoden* führen (Altman 1994). Wenn die Datenbasis gut dokumentierter Public Use Files nachgerechnet werden kann, steigt damit auch die Wahrscheinlichkeit, dass Verbesserungsvorschläge von dritten Wissenschaftlern an die ursprünglichen Forscher weitergereicht werden können. Mit Verbesserungsvorschlägen sind hier Datenerhebungsmethoden und statistische Analyseverfahren gemeint. Wichtig sind Rückmeldungen dieser Art vor allem für andauernde Surveys, welche als Follow-up geplant sind. Hierbei kann nachträgliche, die Datenerhebungsmethodik anzweifelnde Kritik die Arbeit von vielen Jahren vernichten bzw. in Frage stellen. Andererseits können Verbesserungsvorschläge schwerlich an ein

Forschungsteam weitergeleitet werden, wenn die genauen Umstände der Datenerhebung sowie die resultierende Datenbasis einer direkten Evaluation und Verifikation nicht offen stehen.

(4) Die Aufbereitung wissenschaftsöffentlicher Datensätze kann Wissenschaftler dazu ermutigen, *vielfältigere Perspektiven* für die Lösung einer Forschungsfrage einzubeziehen. Wie Fienberg et al. (1985: 13) ausführen, steigt unter der Voraussetzung weithin zugänglicher Datensätze zu einer breiten Palette wissenschaftlicher Themengebiete die Wahrscheinlichkeit, dass Forscher auf Informationen anderer Wissenschaftsgebiete treffen, die für ihre eigene Untersuchung wichtig sind. Die Verwendung von Daten verschiedener Disziplinen für eigene Forschungsfragen erweist sich als guter Test für die eigenen theoretischen Schlussfolgerungen. Außerdem fördert dies den Aufbau persönlicher Kontakte zu Wissenschaftlern anderer Disziplinen, wodurch wiederum die *Multidisziplinarität* der Forschung verstärkt wird.

(5) Die volle *Weiterverwendbarkeit* einmal erhobener Datensätze für Sekundäranalysen wird zum einen durch die Struktur der Wissenschaftsförderung untergraben, zum anderen aber auch durch unzureichende Datenmanagementpraktiken im Wissenschaftsalltag (Meier 2003). Vor allem in den Gesundheitswissenschaften Public Health werden Kooperationen von Forschern verschiedener Wissenschaftsrichtungen gebildet, welche an der Analyse einer spezifischen Fragestellung zusammenarbeiten. Mit dem Ende eines solchen Projektes trennen sich häufig die Arbeitsgruppen mit der Folge, dass das Know-how über die erhobenen Datensätze verloren geht.

Daraus resultiert der Umstand, dass selbst die Experten der statistischen Analyse die Beziehungen zwischen den Variablen des Forschungsdatensatzes nicht mehr mit den korrespondierenden Items der Feldinstrumente verknüpfen können. Noch schwerer gestaltet sich die Reidentifizierung berechneter Konstrukte des Datensatzes, welche bereits auf Analysen der verwendeten Erhebungsinstrumente zurückgehen (Meusel et al. 2001). An dieser Stelle kann die wissenschaftsöffentliche Aufbereitung als eine unabhängige Instanz von dem Forscher gesehen werden, welcher die Daten ursprünglich erhob und jenen, welche die Daten später für Sekundäranalysen verwenden.

Die standardisierte Aufbereitung von Public Use Files erzielt dabei mehrere Vorteile für beide Seiten: So fallen beispielsweise die Kosten der Dokumentation von Datensätzen nicht allein dem Forscher zu Lasten, welcher schon mit der Erhebung der Daten betraut war, sondern können über eine Nutzungsgebühr der späteren Public Use Files umverteilt werden. Weiterhin stellen Inkompatibilitäten in Datensätzen, komplexe Dateistrukturen oder ungenügend dokumentierte Variablen häufig die größten Hindernisse für eine Sekundäranalyse dar. Durch die Entwicklung und Anwendung von Standards und Konventionen für die Speicherung, Aufbereitung und Dokumentation von Forschungsdaten innerhalb wissenschaftsöffentlicher Datensätze bzw. Public Use Files können diese Hürden umgangen werden .Der Vorteil der Aufbereitung wissenschaftsöffentlicher Datensätze kann

damit als *Schutz vor Verlust von Forschungsdaten und Metadaten über diese Datenbasis* beschrieben werden.

(6) Die *Einführung in wissenschaftlich-statistische Analysen für Studenten* empirisch orientierter Wissenschaften kann durch die Verwendung wissenschaftsöffentlicher Forschungsdatensätze stark aufgewertet werden. King (1995) sieht in der Reproduktion und Erweiterung existierender hochqualitativer Forschungsergebnisse ein wichtiges pädagogisches Hilfsmittel. Auch Fienberg et at. (1985: 13) bezeichnen die Verfügbarkeit einer Vielzahl von sorgfältig dokumentierten Datensätzen als einen großen Gewinn für statistisch-wissenschaftliche Analysen. Den Autoren folgend bieten Daten realer Problemstellungen für Studenten zwei Vorteile. Erstens kann der Prozess der Datenerhebung im Hinblick auf Genauigkeit, Relevanz der Fragestellung sowie Effizienz des Studiendesigns erlernt und bewertet werden. Zweitens können diese Daten verwendet werden, um verschiedene Analysetechniken der Originalstudie nachzustellen, um verschiedene Schlussfolgerungen abzuleiten sowie um eigenständige Analyseansätze zu testen.

Die Implementierung dieser Strategie in der deutschen sozialwissenschaftlichen Methodenausbildung wird jedoch als defizitär und im Allgemeinen als nicht praxis- und berufsfeldbezogen beschrieben (Pötschke und Simonson 2003: 73). Die Autoren verweisen darauf, dass neben der Vermittlung von Kompetenzen zur selbstständigen Datenerhebung auch die Befähigung zur Sekundäranalyse bereits vorhandener Datenbestände stärker in den Mittelpunkt der akademischen Ausbildung gerückt werden soll. Absolventen sollten somit in der Lage sein, fremde wissenschaftliche Forschungsbeiträge und Diskussionen verstehen und interpretieren zu können. Daher fordern sie im Weiteren, dass eine bedeutsame Aufgabe universitärer Ausbildung darin zu sehen ist, Studierenden den Zugang zu den neuesten Forschungsergebnissen im jeweiligen Fachgebiet zu ermöglichen (Pötschke und Simonson 2003).

Für die Gesundheitswissenschaften Public Health kann erweiternd festgehalten werden, dass Public Use Files eine realistischere, am späteren Berufsbild orientierte Einführung in die statistische Datenanalyse ermöglichen. Einerseits findet die Public Health Ausbildung in Deutschland zum großen Teil innerhalb von Aufbaustudiengängen berufsbegleitend mit begrenzten zeitlichen Rahmen statt. Andererseits erfordern gesundheitswissenschaftliche Fragestellungen zu einem hohen Anteil epidemiologische Studiendesigns. Dies heißt, um Aussagen über Krankheitsprävalenzen bzw. Interventionswirksamkeiten in einer Zielpopulation treffen zu können, müssen statistisch umfangreiche Erhebungen durchgeführt werden, welche den zeitlich verfügbaren Rahmen der statistischen Ausbildung deutlich überschreiten. An dieser Stelle sollten Public Use Files bzw. andere wissenschaftsöffentliche Datensätze helfen, in enger Verbindung zur jüngsten gesundheitswissenschaftlichen Forschung die statistisch-epidemiologische Ausbildung an realen Fragestellungen sowie an einer de facto zu dieser Fragestellung erhobenen Datenbasis durchzuführen.

4 Nachteile des Data Sharing

Den Vorteilen der Aufbereitung wissenschaftsöffentlicher Datensätze steht eine Reihe von Nachteilen gegenüber.

(1) *Datenschutzrechtliche Ansprüche* an die Forschung sind der wichtigste Nachteil der Aufbereitung wissenschaftsöffentlicher Datensätze. Vor allem gesundheitswissenschaftliche Fragestellungen zielen auf Aussagen ab, welche von den beteiligten Studienteilnehmern zu ihrem intimsten Privatleben gerechnet werden. Der Bereich dieser Aussagen reicht von standardisierten Antworten zu eigenen Krankheitsbildern über weit reichende Beschreibungen eigener täglicher Lebensgewohnheiten bis hin zu enthüllenden Geständnissen über Süchte und Abhängigkeiten, etwa zu Alkoholkonsum, Drogen und Nikotingebrauch. Für den Forscher, welcher solche privaten Bekundungen zum Untersuchungsgegenstand hat, ist es eine unabdingbare Voraussetzung, ein sensibles Vertrauensverhältnis zwischen ihm und den Untersuchungsteilnehmern zu schaffen. Je diffiziler solche Aussagen sind, desto aufwendiger und kompromissreicher gestaltet sich dieses Vertrauensverhältnis für beide Seiten.

Für den Forscher sind in diesem wechselseitigen Verhältnis zwei zentrale Komponenten zu beachten: *Vertraulichkeit* und *Privatsphäre*. Fienberg et al. (1985: 19) definieren diese beiden Begriffe wie folgt: „Confidentiality refers to not disclosing responses to questions that could be identified as belonging to an individual organization or person. Privacy refers to the right of an individual not to make personal information available to another“. Während die Einhaltung von Vertraulichkeit gegenüber dem Studienteilnehmer unumstößliche Voraussetzung im Forschungsprozess ist, wird auch die Einhaltung der Privatsphäre für den Forschungsprozess immer bedeutender. Durch die steigende Anzahl von Befragungen und Interviews unterschiedlich seriöser Medien, Marktforschungsinstitute und wissenschaftlicher Forschungseinrichtungen wächst zum einen die Skepsis gegenüber solcher Datensammlung, zum anderen aber auch das Bewusstsein der Bevölkerung, Informationen über die eigene Person sparsamer nach außen zu tragen.

Um der Bevölkerung ausreichenden Schutz vor der Missachtung ihrer Privatsphäre zu bieten, wurden in Deutschland, wie auch in vielen anderen modernen Industriestaaten, umfangreiche Datenschutzgesetze verfasst (Krappweis et.al. 1997; Peto et.al. 2004). Dazu zählen Gesetzgebungen und Richtlinien in unterschiedlichen administrativen Ebenen, welche zum einen die Möglichkeiten der Datenerhebung eingrenzen, vor allem aber Restriktionen beim Gebrauch einmal erhobener Daten vorschreiben. Für die Aufbereitung und Dissemination wissenschaftsöffentlicher Datensätze hat dies weitreichende Konsequenzen, welche sich in der Pflicht zur teilweisen bis vollständigen Anonymisierung der Daten bzw. in der gänzlichen Unterbindung der Weitergabe von Forschungsdaten zeigen. Für den Forscher, welcher Public Use Files für eigene Analysen verwendet, bedeutet die

Anonymisierung jedoch einen ernsthaften Einschnitt in die Brauchbarkeit der Daten. Werden zum Beispiel personenbezogene Daten vollständig gelöscht, können auch andere Informationen des Datensatzes nicht mehr auf Individuenebene miteinander verbunden werden.

Datenschutzrechtliche Ansprüche haben deshalb tief greifenden Einfluss sowohl auf die direkte Weitergabe von Forschungsdaten (*Data Sharing*) sowie auf die Erstellung von Public Use Files auf Grundlage von Studien kleineren Umfangs. Andererseits kann das Argument des Datenschutzes auch als Scheinargument für die Zurückhaltung von Forschungsdaten (Data Withholding) genutzt werden. Eine standardisierte Verfahrensweise, welche mit aktuell gültigen Datenschutzgesetzen koordiniert ist, könnte dem datenerhebenden Forscher mehr Sicherheit und Rückhalt für das Weitergeben seiner Forschungsdaten geben.

(2) Ein weiterer Nachteil wissenschaftsöffentlicher Datenaufbereitung kann im *hohen Aufwand* der Datensammlung, der Datenaufbereitung und der Archivierung gesehen werden. Die Datensammlung setzt umfangreiche Kontakte mit den Forschergruppen voraus, welche die ursprüngliche Datenerhebung ausführten. Für jedes zu erstellende Public Use File müssen das jeweilige Studiendesign genau analysiert und damit assoziierte Materialien, Dokumente und Erhebungsinstrumente identifiziert werden. Die Ausführlichkeit, mit welcher dieser Schritt ausgeführt wird, entscheidet über die letztendliche Verwertbarkeit des resultierenden Public Use File.

Die nachfolgende Datenaufbereitung beinhaltet die Dokumentation der zuvor gesammelten Materialien. Wie Fienberg et al. (1985: 16) ausführen, sind Forschungsdaten oft schlecht dokumentiert. Wissenschaftler glauben die Details der Datenerhebung, der Variablenkonstruktion und besondere Eigenheiten der Datenbasis eher im Gedächtnis behalten zu können und diese nicht niederschreiben zu müssen. Zudem werden der Einfachheit halber oft routinemäßige Verfahren der Datenvorbereitung und Datendokumentation verwendet, mit welchen die Forscher vertraut sind. Diese genügen zwar meist dem ursprünglichen Forschungsziel, nicht jedoch den akzeptierten Standards der Weitergabe dieser Datenbasis an andere Forscher. Weiterhin ermöglichen die finanziellen und personellen Ressourcen eines Forschungsprojektes vielfach nicht eine optimale Datendokumentation innerhalb der Projektarbeit. In der Konsequenz bedarf die Aufbereitung der ursprünglichen Daten in wissenschaftsöffentliche Datensätze meist einer aufwändigen nachträglichen Bearbeitung.

Für eine sinnvolle Datenarchivierung mehrerer Public Use Files ist zudem die Konzeption und Implementierung einer relationalen Datenbank notwendig. Die Brauchbarkeit des in ihr abgebildeten Datenbestandes hängt in hohem Masse von dessen Aktualität ab. Aus diesem Grund reicht es nicht, eine systematische Archivierung eines einmalig erstellten Datenbestandes zu erstellen. Vielmehr müssen Möglichkeiten gefunden werden, wie dieser Datenbestand so aktuell wie möglich gehalten werden kann. So ist es beispielsweise sinnvoll, auf einer ursprünglichen Studie aufbauende

Publikationen und Folgepublikationen im dazugehörigen Public Use File abzulegen, damit Wiederholungen umgangen und bereits erstellte Analysen rekapituliert werden können. Dies jedoch erfordert eine Instanz, welche sich konstant um die Erweiterung, Aktualisierung und Pflege des Datenbestandes kümmert.

(3) Die Möglichkeit *falscher Auswertungen von Konstrukten* ist ein weiterer Nachteil der Aufbereitung wissenschaftsöffentlicher Datensätze. Dieser bezieht sich direkt auf die spezifischen Charakteristika der gesundheitswissenschaftlichen Forschung. Ein zentraler Punkt des wissenschaftlichen Interesses der Public Health Forschung ist die Erhebung und Auswertung medizinischer und/oder psychologischer Diagnosen in Bezug zu ihrem Auftreten innerhalb der untersuchten Population. Die korrekte Interpretation vieler dieser Diagnosen setzt ein umfassendes Basiswissen der Konstrukte voraus, welche durch die erhobenen Daten widergegeben werden sollen. Dieses Basiswissen ist wiederum häufig an den fundamentalen Wissensbestand einer Disziplin geknüpft. Beispielsweise ist für die sensible und fachkundige Interpretation klinisch-psychologischer Diagnosen zu verschiedenen Angstzuständen die Kenntnis der theoretischen Fundamente dieser Konstrukte notwendig. Hierbei können keine direkten Parameter gemessen werden. Vielmehr handelt es sich immer um eine Sammlung indirekter Parameter, welche in ihrer Gesamtheit interpretiert und auf Fehler überprüft werden müssen.

Die ungehinderte Veröffentlichung solcher Forschungsdaten kann somit zu Fehlinterpretationen oder Fehleinschätzungen und damit zu falschen Empfehlungen auf Grundlage einer richtigen Datenbasis führen. Diesem Nachteil muss durch geeignete Vorsichtsmaßnahmen auf jeden Fall in solchen Fällen entsprochen werden, in denen die Einschätzung der ursprünglichen Forscher dies nahelegt. Ein geeignetes Mittel wäre unter Umständen eine Einweisung in einen Datensatz durch das ursprüngliche Forscherteam, sobald der Gebrauch eines Public Use Files einer solchen Studie angefordert wird.

(4) Weitere Nachteile ergeben sich aufgrund der Notwendigkeit der *Wahrung der Interessen der ursprünglichen Forscher* an ihrer eigenen Arbeit. Trotz der vielen Vorteile, welche das Weitergeben von Forschungsdaten (*Data Sharing*) für den Forschungsprozess haben kann, liegt es meist nicht im Interesse der Forscher, ihre selbst erhoben Daten weiterzugeben. Dies wiederum hat verschiedene Gründe. Fienberg et al. (1985: 17) zählen dazu Folgende: Die ursprünglichen Wissenschaftler könnten befürchten, (a) mit Fehlern in ihrer originalen Arbeit konfrontiert werden; (b) dass nachfolgende Analysten Resultate eher publizieren als sie selbst (diese Punkt ist vor allem brisant bei Panelstudien, welche sich in mehreren Erhebungswellen über viele Jahre erstrecken); (c) dass Reanalysen ihrer eigenen Daten nur publiziert werden, wenn deren Ergebnisse denen der Originalstudie widersprechen (ein Anreiz, welcher eine Verzerrung in Richtung negativen Feedbacks produziert); (d) dass Reanalysen durch schlecht qualifizierte Forscher den Effekt (c) noch verstärken und damit ihre

wissenschaftliche Reputation Schaden erleidet; und (e) dass sie die Kontrolle über ihre Daten, über den Nutzungszweck sowie über die angewendeten Analysemethoden verlieren.

Auch im Kontext der deutschen Public Health Forschung der letzten Jahre sind diese Kriterien durchaus anwendbar. Die Einwerbung von Drittmittelprojekten ist sehr zeitaufwendig und erfordert ein hohes Maß an Zeit, Recherchen und Verhandlungen mit möglichen Projektpartnern. Dennoch erhält der Forscher, welcher diesen Aufwand betreibt, seine Gratifikation nur über die Publikation von Ergebnissen in wissenschaftlichen Fachzeitschriften. Forscher, welche sich möglicherweise auf Reanalysen wissenschaftsöffentlicher Datensätze spezialisieren würden, erhielten ein unfair hohes Maß an wissenschaftlicher Gratifikation, ohne selbst jemals den Aufwand der Vorbereitung von Drittmittelprojekten realisiert zu haben. Diese Möglichkeit hätte wiederum negative Konsequenzen für die Forschung im Allgemeinen, weil dem Aufwand qualitativ hochwertiger Datenerhebungen innerhalb des Wissenschaftssystems verminderte Anreizstrukturen geboten würden. Die Veröffentlichung von Public Use Files muss also Strategien zur Lösung dieser Interessenkonflikte bieten.

II. Anlage der Erhebung „Data Sharing“

1 Fragestellung

Zusammenfassend: in dieser Studie soll als erstes geklärt werden, ob sich die Erfahrungen von Campbell (2002) auch für den deutschsprachigen Raum bestätigen lassen und inwieweit sich Unterschiede ergeben. Des Weiteren soll der Frage nachgegangen werden, ob die Ergebnisse aus den Biowissenschaften auch für die Sozialwissenschaften gelten.

Weiterhin sollte diese Studie einen Einblick geben, ob institutionalisierte *Data Sharing*-Instrumente auch Auswirkungen auf die sozialwissenschaftliche Publikationspraxis haben. Wenn man davon ausgehen kann, dass Daten erst nach dem Ende eines Projektes und den meist damit verbundenen Publikationen weitergeben werden können, ist mit längeren Zeiträumen zu rechnen. So müssten in älteren Publikationen mehr Verweise auf bestehende wissenschaftsöffentliche Datensätze und Instrumente gefunden werden als in Veröffentlichungen jüngeren Datums. Grund zu dieser Annahme für die Sozialwissenschaften gibt vor allem das Zentralarchiv in Köln.

2 Methode

Wie bereits eingangs ausgeführt, kann in der von Campbell (2002) verwendeten Methodik ein Problem gesehen werden. Der Inhalt des verwendeten Fragebogens lässt eine Antwortverzerrung der befragten Forscher/innen in Richtung des sozial erwünschten Ortes einer Antwort vermuten. Daraus

könnte sich eine stärker positiv bewertete Einschätzung des Teilens von wissenschaftlichen Daten in der Forschungsgemeinschaft entstehen.

Um das Problem sozialer Erwünschtheit bei der Antwortgabe zu vermeiden, sollte eine einfache Idee umgesetzt werden: Wie antworten Forscher/innen im wissenschaftlichen Alltag auf eine reale Anfrage zur Übersendung Ihrer Forschungsdaten? Dazu wurde eine Auswahl von publizierten wissenschaftlichen Aufsätzen getroffen, und die Autor/innen dieser Artikel um weiterführende Daten, zugrunde liegende Instrumenten und Auswertungsroutinen per E-Mail gebeten. Um die Anfrage realistisch zu gestalten, wurde von den Autoren die Recherche zu einem fiktiven Dissertationsthema vorgegeben.

Die Art und Weise der Antworten sollten Aufschluss über das Maß der Bereitschaft geben, eigene Forschungsdaten mit anderen Wissenschaftlerinnen und Wissenschaftlern zu teilen.

a) Auswahl der Artikel

Um die Fragestellungen der Arbeit in der Auswahl der Artikel abbilden zu können, wurden für zwei Wissenschaftsbereiche, Soziologie und Gesundheitswissenschaften, Zeitschriften ausgewählt, die vor allem Aufsätze mit empirisch-quantitativem Schwerpunkt veröffentlichen.

Zudem wurden für den Wissenschaftsbereich Soziologie zwei Jahrgänge ausgewählt, welche ein ausreichend großes Zeitfenster abbildeten, um eine mögliche Archivierung der Forschungsdaten im Zentralarchiv zu erfassen. Weiterhin wurden für beide Wissenschaftsbereiche sowohl deutschsprachige wie auch englischsprachige Zeitschriften ausgewählt.

Aufgrund des explorativen Charakters der Studie stand nicht die umfassende Berücksichtigung aller möglichen Zeitschriften im Vordergrund, vielmehr wurden einige Zeitschriften auf der Grundlage ihrer Reputation in der Forschungsgemeinschaft ausgewählt. Tabelle 1 fasst die ausgewählten Zeitschriften zusammen. Zur Wahrung der Anonymität der kontaktierten Wissenschaftler werden die ausgewählten Jahrgänge nicht benannt.

Tabelle 1: Ausgewählte Zeitschriften

Soziologie (jeweils zwei Jahrgänge)	Gesundheitswissenschaften (jeweils ein Jahrgang)
• Kölner Zeitschrift für Soziologie und Sozialpsychologie • ZUMA Nachrichten • ZA Information • The American journal of sociology	• Public Health Forum • Das Gesundheitswesen • Zeitschrift für Gesundheitswissen-schaften • European Journal of Public Health

Vor der Sichtung der Artikel wurde eine Datenbank angelegt, in der die relevanten Informationen gesammelt wurden. Für jedes Heft im ausgewählten Jahrgang wurden folgende Aspekte in der Datenbank erfasst:

- Zeitschrift und Ausgabe
- Titel
- Autorennamen und Kontaktadressen sowie E-Mail
- Abstract, soweit elektronisch verfügbar

Die in der Datenbank erfassten Informationen wurden anhand folgender Kategorien gesichtet:

- Hat der Artikel einen quantitativ empirischen Schwerpunkt?
- Wurden die Ergebnisse mit Hilfe einer Sekundäranalyse produziert?
- Wurde ein eigenständiges Instrument verwendet? (nur bei Primärerhebungen)
- Ist ein Datensatz verfügbar?
- Wurden ausreichend nachvollziehbare Auswertungsroutinen angegeben?
- Wurde ein Verweis auf den Zugang zu den Forschungsdaten geäußert?

Mit der ersten Überprüfung wurden alle Artikel aus der Erhebung ausgeschlossen, die Buchbesprechungen zum Gegenstand hatten. Ebenfalls nicht berücksichtigt wurden redaktionelle Mitteilungen sowie Forschungsarbeiten, die auf qualitativ erhobenen Daten beruhte. Für qualitative Daten stehen in der empirischen Sozialforschung ebenfalls Archive, wie beispielsweise das Archiv für Lebenslaufforschung in Bremen, zur Verfügung. Allerdings erscheint ein Vergleich des Teilens so unterschiedlicher Daten aufgrund der zum Teil erheblichen Differenzen in den Forschungsdesigns als nicht angebracht.

Für Sekundärdatenanalysen wurde recherchiert, ob ein Verweis auf die Forschungsdaten besteht. Bei den Primärerhebungen wurde der Abstract zusätzlich nach Hinweisen auf ein Instrument durchsucht.

Letztlich wurde entschieden, ob eine Auswertungsroutine zum Nachvollziehen des Lösungsweges vom Datensatz bis hin zu den Ergebnissen nötig sein könnte, beziehungsweise inwieweit das Setup der Datenanalyse bereits ausreichend beschrieben wurde. Von den 361 gesichteten Artikeln gingen letztlich 203 in die Untersuchung ein.

b) Erstkontakt

Es wurden Erst- bzw. Zweitautor/in aller ausgewählten Artikel per E-Mail kontaktiert und auf den betreffenden Artikel angesprochen. Unter Vorgabe einer Recherche für eine Dissertation wurden die den Ergebnissen zugrunde liegenden Informationen erfragt.

Die Literaturrecherche für eine Dissertation sollte der Bedeutung einer Anfrage Gewicht verleihen und berechtigtes Interesse repräsentieren. Die betreffende Person wurde persönlich angesprochen und um die Unterstützung des Promotionsvorhabens gebeten. Der relevante Artikel wurde mit Titel sowie den Erscheinungsdaten genannt. Damit sollten regelmäßig publizierenden Autor/innen und Verfasser/innen älterer Artikel die Zuordnung der Anfrage erleichtert werden.

Der Text setzte sich aus vorgefertigten Passagen zusammen. Somit konnten in Abhängigkeit der zuvor beschriebenen Sichtung detailliert pro Aufsatz um die Übersendung von Forschungsdaten, Instrumenten sowie Auswertungsroutinen gebeten werden. Die Textpassagen befinden sich im Anhang. Der Kontakt erfolgte für deutschsprachige Artikel in Deutsch und für englischsprachige Artikel in Englisch.

Bei allen Primärerhebungen wurde nach dem Instrument gefragt. War im Artikel nicht ersichtlich, ob es sich bei dem Datensatz um einen bereits der Öffentlichkeit zur Verfügung stehenden handelt, wurde ebenfalls nach dem Datensatz bzw. dem Zugang zum Datensatz gefragt. Bis auf zwei Personen, welche die Auswertungsroutine im Anhang des Artikels bereits ausführlich veröffentlichten sowie weiteren drei Personen bei denen ein Setup nicht notwendig war, wurden alle nach den Auswertungsroutinen gefragt.

c) Zweiter Kontakt

Nach der ersten E-Mail mit der Anfrage wurde fünf Tage später eine zweite E-Mail versendet. Diesmal informierten wir alle Teilnehmer/innen der Studie über das Forschungsinteresse und die Anlage der Studie. Für weiterführendes Interesse präsentierten wir einen Link zur Forschungsskizze im Internet sowie die Kontaktmöglichkeiten zu den Autoren, jeweils in Deutsch und Englisch. Des Weiteren enthielt die zweite E-Mail, einen Link zu einer kurzen Online-Befragung zum Thema *Data Sharing* (Ergebnisse werden hier nicht berichtet).

d) Auswertung

Die Rückmeldungen per E-Mail wurden inhaltsanalytisch ausgewertet. Dabei wurden über ein vorher erstelltes Kategoriensystem die Häufigkeiten unterschiedlicher Informationen in den E-Mails erfasst und in die Datenbank übertragen. Ziel war es, systematisch und intersubjektiv nachvollziehbar die inhaltlichen und formalen Merkmale der einzelnen E-Mails zu beschreiben (Früh 1991). Die uns zugesandten Informationen wurden im Anschluss an die Erfassung gelöscht. Zur Kategorisierung der Art und Weise des Rücklaufs wurden beispielsweise die folgenden Aspekte erfasst:

- Wurde der Datensatz gesendet?
- Wurde das Instrument gesendet?
- Auswertungsroutinen geschickt?

Alle drei Kategorien wurden anhand der Optionen *geschickt, explizit verweigert, vorerst verweigert, kein Bezug in der E-Mail* (das komplette Kategoriesystem befindet sich im Anhang).

e) Teilnahme

Von den 361 Artikeln in den ausgewählten Zeitschriften erfüllten 203 die relevanten Kriterien. Die Aufsätze verteilten sich dabei wie folgt auf die Kategorien Sprache und Wissenschaftsdisziplin.

Tabelle 2: Auswahlgesamtheit und Teilnahmeverhalten

	Artikel gesamt		davon kontaktiert		davon teilgenommen	
Gesamt	(n=361)	(%)	(n=203)	(%)	(n=101)	(%)
Englisch	87	24	63	31	25	25
Deutsch	274	76	140	69	76	75
Gesundheitswissenschaften	237	66	136	67	74	73
Sozialwissenschaften	124	34	67	33	27	27

Wir erhielten von 101 Personen eine Antwort auf eine der beiden E-Mails (vgl. Tabelle 2). Von den 203 angeschriebenen Personen konnte zu insgesamt 36 kein Kontakt hergestellt werden. In den meisten Fällen waren dafür technische Ursachen verantwortlich (z.B. E-Mail veraltet, inkorrekte Schreibweise). Vereinzelt wurden auch automatische Antworten generiert (z.B. Abwesenheitsnotiz aufgrund von Urlaub). Schließt man diese 36 Fälle bei der Berechnung einer Teilnahmequote aus, haben etwa 60 % aller kontaktierten Personen auf die Anfrage geantwortet.

Ausgehend von den Verteilungen der kontaktierten Personen nach Sprache und Wissenschaftsgebiet ergeben sich Unterschiede im Rücklauf. So finden sich etwas mehr deutschsprachig publizierende Personen in dem Datensatz als Verfasser/innen englischsprachiger Artikel. Bezüglich der Teilgebiete sind die sozialwissenschaftlichen Autoren etwas unterrepräsentiert.

III. Ergebnisse

1 Art und Weise der Antwort

Die Menge der Publikationen, für welche die Forschungsdaten zugänglich sind, setzen sich zusammen aus den Publikationen, welche bereits im Text oder Anhang einen Verweis auf den wissenschaftsöffentlichen Charakter ihrer Forschungsdaten enthielten sowie jenen, bei denen der Autor auf die E-Mail Nachfrage mit dem Zusenden der gewünschten Informationen antwortete. Insgesamt wurden dabei angefragt:

- 150 Instrumente
- 175 Datensätze
- 198 Auswertungsroutinen

Von den 101 antwortenden Autor/innen wurden 91 um die Übersendung von Forschungsdaten gebeten, 79 um die Weitergabe von Instrumenten sowie 99 um die Übermittlung von Auswertungsroutinen. Bezüglich des verwendeten Instruments stellte sich die Praxis des Teilens wissenschaftlicher Materialien wie folgt dar. 3 Personen haben das Instrument geschickt (vgl. Tabelle 3). Jeweils 5 Personen haben das Instrument verweigert beziehungsweise die Weitergabe von weiterem Informationsaustausch abhängig gemacht. Bei 3 Instrumenten wurde ein öffentlicher Zugang genannt. Allerdings haben etwas mehr als drei Viertel aller Teilnehmer/innen keinen Bezug auf das Instrument genommen.

Tabelle 3: Reaktionen auf die Anfrage zum Instrument

	Instrument	
	(n=79)	(%)
Keinen Bezug	63	80
Gesandt	3	4
Explizit verweigert	5	6
Vorerst verweigert	5	6
Öffentlich zugänglich	3	4

Von den teilnehmenden Personen haben 7 Autoren die Weitergabe des Datensatz nach weiterem Informationsaustausch in Aussicht gestellt (vgl. Tabelle 4). Bei 12 Anfragen wurde ein Verweis auf öffentlich zugängliche Datensatz gegeben (z.B. Zentralarchiv für Empirische Sozialforschung an der Universität zu Köln). 13 Personen haben die Weitergabe aus unterschiedlichen Gründen explizit verweigert. Etwas mehr als die Hälfte aller Personen hat in der Antwort E-Mail keinen Bezug auf den Datensatz genommen.

Tabelle 4: Reaktionen auf die Anfrage zum Forschungsdatensatz

	Forschungsdatensatz	
	(n=91)	(%)
Keinen Bezug	56	62
Gesandt	3	3
Explizit verweigert	13	14
Vorerst verweigert	7	8
Öffentlich zugänglich	12	13

Die Auswertungsroutinen haben 3 Personen sofort zur Verfügung gestellt. 6 angefragte Personen haben die Weitergabe ausgeschlossen und 8 waren zur Weitergabe unter Maßgabe einer ausführlicheren Schilderung des Forschungsanliegens bereit. Der überwiegende Teil erwähnte die Routinen in der Rückantwort überhaupt nicht.

Tabelle 5: Reaktionen auf die Anfrage zur Auswertungsroutine

	Auswertungsroutinen	
	(n=99)	(%)
Keinen Bezug	82	83
Gesandt	3	3
Explizit verweigert	6	6
Vorerst verweigert	8	8
Öffentlich zugänglich	0	0

Vergleicht man die drei angefragten Aspekte so fällt auf, das die meisten Reaktionen in der E-Mail auf den Datensatz erfolgen. Im Vergleich zu den Instrumenten sind die Datensätze auch deutlich häufiger zugänglich. Auf die Auswertungsroutinen wird in den Antworten sehr selten Bezug genommen.

Ein großer Teil der angefragten Personen haben sich für das Interesse am Thema bedankt (n=31). Von fast zwei Drittel der Wissenschaftler, die zu keiner direkten Weitergabe an Informationen bereit waren, wurden unterschiedliche Gründe genannt. Die am häufigsten angeführten Gründe waren dabei, zum Teil in Kombination:

- Persönliche Rücksprache gefordert (n=8);
- Das Interesse Dritter verhindert die Weitergabe (n=7);
- Weitere Publikationen sind geplant (n=4);
- Ein Verweis auf den Artikel unter dem Hinweis, dass dieser vollständig sei und alle angefragten Angaben enthalte (n=22);
- Verweis an einen dritten Besitzer der Daten (n=13);
- Verweis auf andere weiterführende Literatur (n=22);
- Fehlende Zeit zum Aufbereiten der Materialien (n=5);
- Expliziter Einwand aufgrund von Datenschutzbedenken (n=7);

Die Reaktionen ähneln den Antworten der Studie von Campbell (2002). Die Motivationen für das Zurückhalten von Forschungsdaten scheinen über kulturspezifische Unterschiede und Wissenschaftsgebiete hinweg übereinzustimmen. Dabei sind in dieser Studie keine größeren Unterschiede in den Beweggründen zwischen Sozialwissenschaften und Gesundheitswissenschaften feststellbar. Die Unterschiede in den Häufigkeiten der einzelnen Antwortmuster zwischen den beiden Wissenschaftsdisziplinen sind aufgrund der geringen Fallzahlen kaum interpretierbar.

In den Antworten wurden zum Teil Wege aufgezeigt und Verfahrensweisen angeboten, die vorerst geäußerte Zurückhaltung zu überwinden. So wurde mehrfach der Datenzugang in der jeweiligen Einrichtung angeboten, wenn die Weitergabe beispielsweise durch Auftraggeber oder andere Dritte nicht möglich war. In einigen Fällen wurde bei einer zukünftig etablierten und gleichberechtigten Kooperation Datenweitergabe in Aussicht gestellt. Dabei sollten Datenaustauschregeln auf einem sehr einfachen Niveau institutionalisiert werden. Ein Autorenteam war nur zur Datenweitergabe bereit, wenn daraus resultierende Publikationen mindestens eine der Personen als Koautor berücksichtigen würde. Vereinzelt wurden massive Hindernisse für eine Datenweitergabe genannt. Beispielsweise konnten Daten nicht weitergegeben werden, da Datenverlust durch Diebstahl der Technik sowie formatierte Festplatten entstand.

2 Unterschiede in den beiden Disziplinen

Die eingangs angesprochenen Unterschiede in der Etablierung von Institutionen der Datenarchivierung gaben Grund zu der Annahme, dass innerhalb der Soziologie häufiger auf bereits öffentlich zugängliche Datensätze verwiesen würde. Schon im Vorfeld wurde bei der Zusammenstellung der nachzufragenden Artikel deutlich, dass in der Soziologie deutlich häufiger auf die Zugänglichkeit der Datensätze hingewiesen wird.

Auch bei den Autoren, die nicht explizit in Ihrem Artikel auf die freie Zugänglichkeit der Daten eingegangen sind, haben in unserer Untersuchung mehr Verfasser in den soziologischen Zeitschriften auf frei zugängliche Datensätze hingewiesen (vgl. Tabelle 6).

Tabelle 6: Reaktion auf die Anfrage nach Wissenschaftsgebiet

	Soziologie		Public Health	
	(n=20)	(%)	(n=71)	(%)
kein Bezug auf den Datensatz	8	40	48	68
Datensatz geschickt	2	10	1	1
Datensatz explizit verweigert	2	10	11	15
Datensatz vorerst verweigert	1	5	6	8
Öffentlich zugänglich	7	35	5	7

Die höhere Rate an expliziten Verweigerungen der Datensätze auf Seiten der Gesundheitswissenschaften deckt sich zumindest vorerst mit der Vermutung, dass Vertraulichkeit und Anonymität in den Studien der Gesundheitswissenschaften häufiger eine Rolle spielt als in der Soziologie. Unterschiede in den Reaktionen auf die Anfrage, die sich zwischen den Sprachen der Publikationen ergaben, konnten fast vollständig auf den Einfluss der Wissenschaftsdisziplin zurückgeführt werden, da in den ausgewählten Zeitschriften für den Bereich Gesundheitswissenschaften deutlich häufiger englischsprachig publiziert wurde.

3 Differenzen im Erscheinungsjahr

Die Forschungsfrage, ob Publikationen älteren Datums zumindest in der Soziologie häufiger in einem öffentlich zugänglichen Archiv zu finden sind, als Publikationen jüngeren Datums kann anhand der Daten nicht abschließend geklärt werden. Die Zellenbesetzung der einzelnen Gruppen ist dafür zu gering.

IV. Diskussion und Schlussfolgerungen

Die mit dem Teilen wissenschaftlicher Forschungsdaten verbundenen Problemlagen finden sich auch in den Sozialwissenschaften und Gesundheitswissenschaften. Auf Grund der genannten Motive ist es wahrscheinlich, dass die Verweigerung der Weitergabe von Forschungsdaten im wissenschaftlichen Alltag unabhängig von der wissenschaftlichen Disziplin zu finden ist. Disziplinspezifische Unterschiede können sich beispielsweise auf Grund der unterschiedlichen Forschungsgegenstände ergeben. So sind datenschutzrechtliche Bedenken bei epidemiologischen Studien mittels Anamnesebögen eher anzutreffen, als beispielsweise bei selbst rekrutierenden Onlinebefragungen. Allerdings muss darauf hingewiesen werden, dass der Grund „Datenschutz“ in der vorliegenden Untersuchung relativ selten als Verweigerungsgrund genannt wurde.

Die Etablierung möglichst umfassend kompatibler Datenarchive erscheint aus der Sicht der Autoren ein wichtiger Bestandteil künftiger Forschungspragmatik. Dabei sollten vor allem finanziell, personell und zeitlich sehr eng gefassten Projekten trotzdem Möglichkeiten zur Verfügung stehen, Forschungsdaten öffentlich zugänglich zu machen, vorausgesetzt die Rahmenbedingungen (z.B. Datenschutzauflagen) stimmen.

Für die Einordnung der Teilnahmequote ist es wichtig zu beachten, dass die Differenz zwischen den 203 gesendeten E-Mails und den 101 Antworten zu einem großen Teil auf offensichtliche technische Probleme, wie z.B. veraltete bzw. falsche E-Mail Adressen zurückzuführen ist. Zusätzlich kann ein Teil der Nichtteilnehmer durch Gründe wie Dienstreisen, Auslandsaufenthalte und andere Verpflichtungen erklärt werden, die eine Antwort auf die E-Mail verzögerten. Vor diesem Hintergrund ist die Antwortrate auf die Anfrage per E-Mail von etwa 60 % plausibel.

Die Rückmeldungen der einzelnen Autoren waren bei dem Umfang nicht umfassend zu überprüfen. Beispielsweise wurde darauf verwiesen, dass die im Artikel beschriebenen Verfahren ausreichen würden, um die Ergebnisse nachvollziehen zu können. Aufgrund des eingangs beschrieben Forschungsinteresses wurde auf eine erneute Überprüfung der Artikel verzichtet.

Die Art und Weise der Datenerhebung als verdeckte Erhebung wurde von einigen Autoren kritisiert. Sicherlich ist die gewählte Erhebungsmethode ungewöhnlich. Aber an dieser Stelle soll noch einmal betont werden, dass dem Forscherteam die Persönlichkeitsrechte aller Beteiligten wichtig sind und waren. Es wurden alle gesendeten Anhänge vernichtet. Die Auswertung ist im Sinne empirischer Forschung nicht an der einzelnen Person interessiert. Die angeschrieben Forscher waren nur als Merkmalsträger relevant.

Eine Reihe positiver Rückmeldungen hat allerdings gezeigt, dass durchaus ein Bedarf an Diskussion und Informationsaustausch zu diesem Thema auch bei anderen Kollegen herrscht und diese Studie, wenn auch mit unkonventionellen Mitteln geführt, mit Interesse verfolgt wurde.

Die geringe Bezugnahme auf die Auswertungsroutinen ist aus verschiedenen Gründen fraglich. Welche Informationen über die angewendeten ergebnisproduzierenden Strategien müssen offen gelegt werden? Reicht die Angabe des verwendeten Verfahrens aus, um die Ergebnisse zu prüfen? Ob die Setups der einzelnen Statistikpakete notwendig sind, wird vor allem bei der Umsetzung komplexerer Modellierungen deutlich. In wie weit Datentransformationen und Selektionsregeln aus dem Text und dem idealer weise vorhandenen Datensatz nachvollzogen werden können, ist nicht sicher. Mit steigendem Komplexitätsgrad einer Auswertung dürfte die Wahrscheinlichkeit sinken, ausgehend von einem Datensatz die präsentierten Ergebnisse in einer eigenen Auswertung zu erhalten. Zugleich besteht aber möglicherweise auch ein Interesse am Schutz innovativer Ideen und aufwendiger Setups.

V. Literaturverzeichnis

Altman, Douglas, 1994: The scandal of poor medical research. BMJ: 283-284.

Campbell, Eric G., Brian R. Clarridge, Manjusha Gokhale, Lauren Birenbaum, Stephen Hilgartner, Neil A. Holtzman, David Blumenthal, 2002: Data Withholding in Academic Genetics – Evidence From a National Survey. JAMA 287: 473-480.

Diekmann, Andreas, 2002: Soziologie und Empirische Sozialforschung – Von den siebziger Jahren bis Heute. In: *Jan van Deth* (Hg.): Von Generation zu Generation: ZUMA Nachrichten Spezial Band 8.

Fienberg, Stephen E., Margaret E. Martin, Miron L. Straf, 1985: Sharing Research Data: Report of the Commitee on National Statistics. Washington, DC: National Academy Press.

Früh, Werner, 1991: Inhaltsanalyse: Theorie und Praxis, München: Oelschläger.

Hall, Nancy S., 2000: The Key Role of Replication in Science. The Chronicle of Higher Education 47: 14.

King, Gary, 1995: Replication, Replication. Political Science and Politics 28: 443-499.

Krappweis, Hans, Jutta Krappweis, Wilhelm Kirch, 1997: Gesundheitsberichterstattung: Datenschutz im Widerspruch zum Informationsbedarf. Gesundheitswesen und ökonomisches Qualitätsmanagement 2: 157-159.

Meusel, Dirk, Mike Kühne, Wilhelm Kirch, 2005: Zugänglichkeit von Studiendaten kleineren Umfangs. Public Health Forum: 12.

Meier, Friedhelm, 2003: Qualitätsgesichertes Datenmanagement für die Sozialforschung. ZA Information 52: 58-71.

Merton, Robert, 1973: The Sociology of Science: Theoretical and Empirical Investigations. Chicago: University of Chicago Press.

Meusel, Dirk, Peggy Göpfert, Wilhelm Kirch, 2001: Public Use Files Project: Archiving and Disseminating empirical data of the Public Health Research Networks in Germany. In: *Merker, Nora, Peggy Göpfert, Wilhelm Kirch* (Hg.): Public Health Research and Practice: Report of the Public Health Research Association Saxony 2000-2001. Regensburg: Roderer Verlag.

Peto, Julian, Olivia Fletscher, Clare Gilham, 2004: Data protection, informed consent, and research. British Medical Journal 328: 1029-1030.

Polanyi, Michael, 1962: The republic of science: its political and economic theory. Minerva.

Popper, Karl, 1986 (1934): Logik der Forschung. Tübingen: Mohr.

Pötschke, Manuela und Julia Simonson, 2002: Konträr und ungenügend? Ansprüche an Inhalt und Qualität einer sozialwissenschaftlichen Methodenausbildung. ZA Information: 72-92.

Schnell, Rainer, 2002: Anmerkungen zur Publikation "Möglichkeiten und Probleme des Einsatzes postalischer Befragungen" von Karl-Heinz Reuband. Kölner Zeitschrift für Soziologie und Sozialpsychologie: 147-157.

Statistisches Bundesamt, 2004: Demographische Standards. Eine gemeinsame Empfehlung des Arbeitskreises Deutscher Markt und Sozialforschungsinstitute e. V. (ADM), der Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e. V. (ASI) und des Statistischen Bundesamtes. Wiesbaden.