German Record Linkage Center
WORKING PAPER SERIES
NO. WP-GRLC-2018-01 | FEBRUARY 21, 2018

'Big Data' aus wissenschafts-soziologischer Sicht: Warum es kaum sozialwissenschaftliche Studien ohne Befragungen gibt

Rainer Schnell

'Big Data' aus wissenschaftssoziologischer Sicht: Warum es kaum sozialwissenschaftliche Studien ohne Befragungen gibt

Rainer Schnell

21. Februar 2018

1 Surveys als Hauptdatenlieferant der Sozialwissenschaften

Der Anteil an Veröffentlichungen, die auf Daten beruhen, die durch standardisierte Befragungen erhoben wurden, steigt in den Sozialwissenschaften – mit Ausnahme der Ökonomie – monoton weiter (Schnell 2012). Dafür gibt es mehrere Gründe.

Zunächst einmal existieren für quantitative Befragungen methodische Standards, so dass die Umsetzung einer Forschungsfrage in eine Datenerhebung in jedem Schritt einer empirisch bewährten Methodologie folgt.[1] Das entsprechende Fach (Survey Methodology) ist bislang das einzige Beispiel für eine prognosefähige Instrumententheorie in den Sozialwissenschaften, die wissenschaftlichen Ansprüchen (außerhalb der Sozialwissenschaften) genügt.[2] Die Existenz dieser empirisch bewährten Instrumententheorie erleichtert Geldgebern, Gutachtern und Lesern die Beurteilung der Projekte und ermöglicht so einen problemlosen standardisierten Projektablauf von der Antragstellung bis zur Veröffentlichung. Da am Ende der Datenerhebung mit Sicherheit ein Datensatz vorliegen wird, ist ein Surveyprojekt weitgehend risikolos.

Daneben gibt es zwei weitere nicht zu unterschätzende wissenschaftssoziologische Gründe. Erstens existieren schon hypertrophe öffentlich finanzierte Infrastrukturen, die entweder das Management der Datenerhebung oder auch die Datenerhebung selbst den Forschenden abnehmen. Zweitens erfordern Veröffentlichungen – auch methodischer Art – auf der Basis bereits erhobener Daten kaum Aufwand. Durch diese Ersparnis der Datenerhebung können auch von weniger engagierten Betreibern einer wissenschaftlichen Karriere problemlos genügend Publikationen für eine an der Zahl der Publikationen orientierten Einstellungspolitik generiert werden.

In anderen Fächern – auch mit sozialwissenschaftlichem Bezug – gibt es Beispiele, dass sich die Datengrundlagen verändern. In den Wirtschaftswissenschaften ist ein deutlicher Anstieg der Nutzung administrativer Daten gegenüber der Verwendung von Surveydaten seit den 80er Jahren festzustellen (Chetty 2012). In den vier führenden Zeitschriften des Faches (American Economic Review, Econometrica, Journal of Political Economy, Quarterly Journal of Economics) stieg der Anteil der Publikationen auf der Basis administrativer Daten zwischen 1980 und 2010 im Mittel um mehr als 30% (Abbildung 1).[3]

[1] Sieht man von der Analyse von Texten im weiteren Sinne ab, so basiert die qualitative Sozialforschung nahezu ausschließlich auf Befragungen. Von einer prognosefähigen Instrumententheorie kann dort aber keine Rede sein. Da Forschung ohne eine Instrumententheorie nur schwer möglich ist, hält der Autor den Titel des Beitrags für deskriptiv korrekt.

[2] Es ist mehr als erstaunlich, dass Prädiktion als Ziel von Wissenschaft in der Soziologie umstritten ist und es eines Buches (Ekland-Olson und Gibbs 2017) bedarf, darauf hinzuweisen, dass dies in anderen Fächern selbstverständlich ist.

[3] Die Daten der Abbildung 1 wurden einer Abbildung bei Chetty (2012) entnommen.

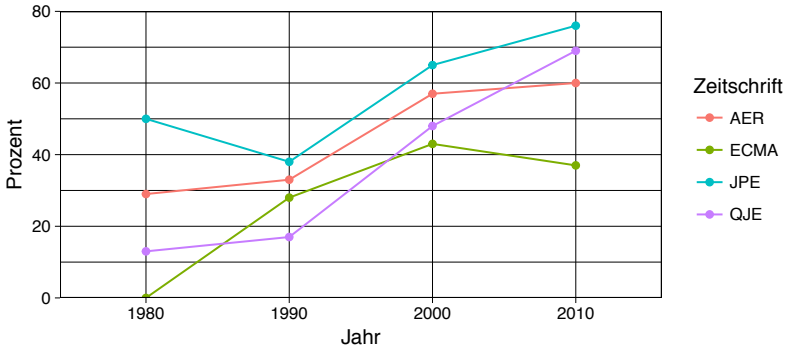

Abbildung 1: Der Anteil von Artikeln, die auf administrativen Daten basieren, in vier führenden ökonomischen Zeitschriften (Chetty 2012)

Damit stellt sich die Frage, warum im Vergleich zur Ökonomie in den anderen Sozialwissenschaften administrative Daten vergleichsweise selten verwendet werden. Für die Beantwortung dieser Frage müssen die verschiedenen Datenformen etwas näher betrachtet werden.

2 Unterschiede zwischen 'Big Data', administrativen Daten und Surveys

Kaum ein anderes Schlagwort im Zusammenhang mit Datenanalysen ist in wenigen Jahren so populär geworden wie der Begriff 'Big Data'. Allerdings ist zumindest in den Sozialwissenschaften kaum klar, was alles dieser Begriff bezeichnen soll und wie mit diesen Daten vor allem im Zusammenhang mit Surveys verfahren werden soll (Japec et al. 2015). Schon Angrist und Krueger (1999) grenzten solche Datenquellen von den traditionellen Datenquellen der Sozialwissenschaften durch die Tatsache ab, dass diese Daten nicht für Forschungszwecke erhoben wurden. Viele Autoren (z.B. Connelly et al. 2016) betrachten vor allem administrative Daten als für die Sozialwissenschaften relevante Form von 'Big Data'.

Betrachtet man die Quellen für 'Big Data' etwas genauer, dann wird rasch deutlich, dass verschiedene Formen von 'Big Data' unterschieden werden müssen (siehe Abbildung 2).[4] Neben den traditionellen Datenformen der Sozialwissenschaften (einschließlich der administrativen Daten) ist eine weitere Unterscheidung zwischen Sensordaten, 'Social Media'-Daten und Transaktionsdaten sinnvoll.[5] Den nicht-administrativen Datenquellen sind einige Merkmale gemeinsam. Dazu gehören zunächst die beiden trivialen Merkmale, dass diese Daten nicht für Forschungszwecke erhoben wurden und im Vergleich zu den traditionellen Datensätzen in der Regel sehr groß sind. Wesentlich bedeutsamer für die Verwendung dieser Daten in der empirischen Sozialforschung sind zwei andere Merkmale: Der Mangel an Kovariaten und die unklare Population.[6]

Um sinnvolle Aussagen über Lebenssituationen und Entscheidungen von Personen für sozialwissenschaftliche Forschungsprojekte machen zu können, sind Informationen über die in den inhaltlichen Modellen der Sozialforschung relevanten Variablen, wie z.B. den demographischen Variablen, Einkommen, Bildung, Familienkonstellation, Beruf etc. unverzichtbar. Diese Variablen sind weder in Sensordaten noch in 'Social Media'-Daten noch in Transaktionsdaten enthalten.[7] Selbst falls in einzelnen Datensätzen einige dieser Merkmale vorhanden sind, ist die Zahl an Kovariaten stark

[4] Die oberste Reihe dieser Abbildung basiert auf einer Abbildung bei Connelly et al. (2016).

[5] Zu den Sensordaten gehören auch die Ergebnisse der klassischen nicht-reaktiven Messverfahren in den Sozialwissenschaften. Daten, die durch Videoüberwachung im öffentlichen Raum entstehen, können ebenfalls als Sensordaten klassifiziert werden. Durch Personen- oder Gestenerkennung können diese zu Transaktionsdaten umgewandelt werden.

[6] Diese beiden Merkmale wurden schon von Angrist und Krueger (1999) für einen Teil der administrativen Daten betont.

[7] Kommerzielle Datenbanken über Konsumenten enthalten zwar gelegentlich demographische Variablen, aber diese sind entweder selbst das Ergebnis der Schätzung durch ein statistisches Modell oder so lückenhaft durch fehlende Angaben, dass die Verwendung für wissenschaftlich belastbare Aussagen derzeit höchst fraglich erscheint.

Experimentaldaten

- Beispiele: Psychologische Experimente, klinische Studien
- Datenerhebung zum Test einer spezifischen Hypothese
- Relativ kleine Stichprobe
- Einfache Datenstruktur
- Systematische Datensammlung
- Bekannte Population

Survey-Daten

- Beispiele: BHPS, GSOEP, NHANES
- Datenerhebung für wissenschaftliche Zwecke
- Meist mehr als eine Forschungsfragestellung
- Kann zu großen Datensätzen führen.
- Kann komplexe Datenstrukturen bedingen
- Systematische Datensammlung
- Bekannte Population

Administrative Daten

- Beispiele: Arbeitslosenversicherung, Rentenversicherung, Führerscheinregister
- Datenerhebung nicht für wissenschaftliche Zwecke
- In der Regel sehr große Datensätze
- Kann komplexe Datenstrukturen bedingen
- Meist unkontrollierte Datenerhebung
- Oft erheblicher Datenaufbereitungsaufwand
- Bekannte Population

Sensordaten

- Beispiele: GPS-Daten, Sportuhren, Strassensensoren
- Datenerhebung nicht für wissenschaftliche Zwecke
- Oft in kommerziellem Besitz
- In der Regel sehr große Datensätze
- Meist nur sehr wenige Variablen
- In der Regel keine Kovariaten
- Häufig unbekannte Population

'Social Media'-Daten

- Beispiele: Twitter, Facebook, Instagram
- Datenerhebung nicht für wissenschaftliche Zwecke
- Oft in kommerziellem Besitz
- In der Regel sehr große Datensätze
- Bedarf häufig aufwändiger Codierung
- Meist nur sehr wenige Variablen
- In der Regel keine Kovariaten
- Häufig unbekannte Population

Transaktionsdaten

- Beispiele: Verbindungsdaten, Abrechnungsdaten, Warenkörbe
- Datenerhebung nicht für wissenschaftliche Zwecke
- Oft in kommerziellem Besitz
- In der Regel sehr große Datensätze
- Kann komplexe Datenstrukturen bedingen
- In der Regel keine Kovariaten
- Häufig unbekannte Population

Abbildung 2: Traditionelle und 'Big Data'-Datenquellen der empirischen Sozialforschung

begrenzt und nicht mit der Fülle von Merkmalen aus Surveys oder administrativen Datenbanken zu vergleichen.

In einigen Forschungsgebieten werden solche Unzulänglichkeiten weniger thematisiert als in anderen. Bei vielen Fragestellungen der klassischen Ökonomie enthalten Registerdaten (z.B. Sozialversicherungsdaten) nahezu alle relevanten abhängigen Variablen (Einkommen, Arbeitslosigkeit) und – mit Ausnahme von Bildung – auch die meisten Kovariaten.

In Fächern, bei denen die unklare Abgrenzung der Untersuchungspopulation und der Verzicht auf echte Zufallsstichproben akzeptierte Praxis ist (wie in der Ökonomie und Medizin) erscheint vermutlich die Verwendung von 'Big Data' oder nicht populationsdeckenden administrativen Datensätzen als unproblematischer als in Fächern, bei denen exakt definierten Zufallsstichproben zum Selbstverständnis gehören (wie in der amtlichen Statistik oder der Survey Methodology).

Empirische Projekte, deren Fragestellung aber zusätzliche Kovariate benötigt oder bei denen eine exakte Populationsdefinition erforderlich ist, erfahren bei der Beschränkung auf vorliegende Daten erhebliche Einschränkungen. Beide methodische Probleme lassen sich aber prinzipiell durch eine – notwendigerweise – personenbezogene Verknüpfung der 'Big Data'-Quellen mit administrativen Daten oder Survey-Daten lösen.[8]

3 'Big Data' erfordert fast immer Record Linkage

Als Zwischenergebnis kann festgehalten werden, dass die Verwendung von 'Big Data' für die empirische Sozialforschung fast immer die personenbezogene Verknüpfung mehrerer Datensätze erfordert. Die Verknüpfung der Daten derselben Person in verschiedenen Datenbanken wird als 'Record Linkage' bezeichnet.[9] Es gibt prinzipiell drei Möglichkeiten für die Durchführung von Record Linkage:

1. mit einer eindeutigen Personenkennziffer,
2. mit unverschlüsselten Identifikatoren wie Namen oder Geburtsdatum
3. mit verschlüsselten Identifikatoren.

Falls eine einheitliche Personenkennziffer zur Verfügung steht, ist Record Linkage technisch trivial (Christen 2012). In Deutschland ist eine solche einheitliche Kennziffer – vermutlich auch langfristig – nicht verfügbar.[10]

In Ländern, in denen keine einheitlichen Personenkennziffern verfügbar sind, muss Record Linkage auf Identifikatoren wie Name, Vorname, Geburtsdatum etc. zurückgreifen. Solche Identifikatoren sind häufig instabil, z.B. durch einen Namenswechsel bei Heirat. Weiterhin führen unterschiedliche Schreibweisen desselben Namens sowie Tipp- oder Erfassungsfehler zu vielen verschiedenen Varianten desselben Namens. Das Ausmaß dieser Fehler ist häufig überraschend hoch: Winkler (2009) berichtet für ein Zensus-Experiment in den USA eine Fehlerrate von 25% in Vornamen und 15% in Nachnamen. Daher müssen Record Linkage-Verfahren verwendet werden, die solche Fehler tolerieren können. Das weltweit am weitesten verwendete Verfahren für Record Linkage

[8] Andere Formen der 'Verknüpfung' sind bei Fragestellungen der empirischen Sozialforschung entweder methodisch kaum verwendbar (z.B. aufgrund des Problems des ökologischen Fehlschlusses) oder schlicht undefiniert: In Diskussionen (und Reviews) findet man häufig den Hinweis auf 'Datenverküpfungen' (oder auch 'linked data'), wobei es die Diskutanten aber stets versäumen zu erklären, was denn eine Datenverknüpfung sei, wenn nicht eine personenbezogene Verbindung unabhängiger Datensätze.

[9] Außerhalb der Statistik wird Record Linkage gelegentlich mit Datenfusion (D'Orazio et al. 2006) oder auch mit Propensity-Matching (Guo und Fraser 2010) verwechselt. Obwohl auch dort Mikrodaten über Akteure zusammengeführt werden, ist es das explizite Ziel beider Verfahren, Daten über unterschiedliche, wenn auch sehr ähnliche Fälle, zusammenzuführen. Auch wenn die verwendeten Algorithmen nicht vollkommen überschneidungsfrei sind, unterscheiden sich die Verfahren wesentlich im Ziel der Anwendung.

[10] Aufgrund des Volkszählungsurteils des Bundesverfassungsgerichts (https://openjur.de/u/268440.html) wird eine solche Kennziffer in Deutschland häufig für gesetzeswidrig gehalten. Im Gegensatz zu dieser traditionellen Interpretation (Metschke 2010) wird dies in der neueren Literatur durchaus positiver beurteilt (Martini und Wenzel 2017). Selbst in dem unwahrscheinlichen Fall, dass eine solche Kennziffer neu eingeführt werden sollte (z.B. durch eine für den jeweiligen Vorgang spezifische Verschlüsselung der Nummer des Personalausweises) wird es Jahrzehnte dauern, bis diese in alten Datenbeständen nachträglich ergänzt wurde. In vielen Fällen, wie z.B. im Neonatalregister (Schnell und Borgs 2015), wird dies prinzipiell nicht möglich sein: Verstorbene Neugeborene werden auch nachträglich keine Personalausweisnummer erhalten.

mit unverschlüsselten, fehlerbehafteten Identifikatoren ist das sogenannte probabilistische Record Linkage. Dieses Verfahren ist in vielen frei verfügbaren Programmen (Christen 2012) implementiert und in Lehrbüchern (Herzog et al. 2007) ausführlich beschrieben. Allerdings basiert das Verfahren zunächst auf unverschlüsselten Identifikatoren. Meist wird dieses Verfahren innerhalb besonders geschützter Umgebungen (Vertrauensstellen) so durchgeführt, dass die Vertrauensstelle keinen Zugang zu den inhaltlichen Daten besitzt. Solche Treuhänderlösungen sind in der Medizin weit verbreitet, allerdings kaum in den Sozialwissenschaften. In Deutschland wurde durch die Einrichtung eines Record Linkage-Centers als Kooperationsprojekt zwischen dem Forschungsdatenzentrum der Bundesagentur für Arbeit und dem Verfasser eine entsprechende Infrastruktur geschaffen und zahlreiche Projekte durchgeführt (Antoni und Schnell 2017).

Während für viele Anwendungen in der Medizin Treuhänderlösungen von allen Beteiligten in der Regel akzeptiert werden, finden sich vor allem in der Informatik häufig höhere Anforderungen an Record Linkage-Verfahren. Dort wird das angestrebte höhere Schutzniveau durch die Sicherheit gegenüber prinzipiell ehrlichen, aber neugierigen Mitarbeitern ('honest, but curious': HBC, Vatsalan et al. 2013) operationalisiert. Zu den Verfahren, die diese Sicherheit bieten sollen, gehören vor allem der Gebrauch von Pseudonymen. Diese sind in der medizinischen Forschung weit verbreitet und auch von Datenschützern weitgehend akzeptiert (Schaar 2014).

Gebräuchlich waren Pseudonyme auf der Basis phonetischer Codes: Ähnlich klingende Namen werden durch eine gemeinsame Codegruppe codiert und die resultierende Zahl (mit Passwort) verschlüsselt (Herzog et al. 2007). Diese Art von Pseudonymen ist weltweit sicherlich derzeit die am häufigsten verwendete Form. Es lässt sich aber zeigen, dass diese Pseudonyme weder besonders sicher sind noch das Ausmaß an Fehlertoleranz besitzen, das bei vielen Anwendungen erforderlich ist (Randall et al. 2016). Mittlerweile gelten hingegen die auf die Arbeitsgruppe des Verfassers (Schnell, Bachteler et al. 2009) zurückgehenden Bloom-Filter basierten Verfahren als 'de facto standard' (Smith 2017) in dem Teil der Informatik, der sich mit diesem Problem beschäftigt (Privacy-preserving Record Linkage oder kurz PPRL).

Durch die Verwendung moderner Formen von Pseudonymen lassen sich die meisten Record Linkage-Probleme technisch und datenschutzrechtlich unbedenklich lösen. Daher sind Pseudonyme zentraler Bestandteil der Empfehlungen im Rahmen der EU-Datenschutzrichtlinie (Council of the European Union 2016).

4 Unterschiede zwischen den Fächern in der Verwendung von Record Linkage

Wie gezeigt wurde, sind die technischen Probleme des Record Linkage zwar nicht endgültig gelöst, aber beherrschbar. Die wichtigsten Hindernisse gegen den weit verbreiteten Einsatz von Record Linkage-Techniken sind daher nicht technischer Art. Einige der Ursachen für die Nichtverwendung existierender Datenbestände sollen daher im Folgenden etwas näher erörtert werden.

Betrachtet man die Verwendung von Record Linkage im Vergleich zwischen der Medizin und den Sozialwissenschaften, so sind die Unterschiede auffällig. Während in der Medizin die Anwendung von Record Linkage bei der Verwendung administrativer Daten mittlerweile einer immer populärer werdende Forschungsstrategie darstellt, ist dies in den Sozialwissenschaften weltweit nur sehr begrenzt der Fall (Abbildung 3).[11]

Es stellt sich die Frage, worauf solche Unterschiede selbst bei der Verwendung von Record Linkage zurückzuführen sind. Für die Beantwortung dieser Frage muss weit ausgeholt werden.

5 Ursachen für verzögerte Adaption neuer Technologien am Beispiel des Record Linkage

Obwohl Record Linkage seit mehr als 50 Jahren eingesetzt wird, sind Anwendungen mit großen Datensätzen für Forschungszwecke außerhalb der amtlichen Statistik vergleichsweise neu. Solche

[11] Die Daten der Abbildung stammen aus einer Recherche des Autors im Januar 2018.

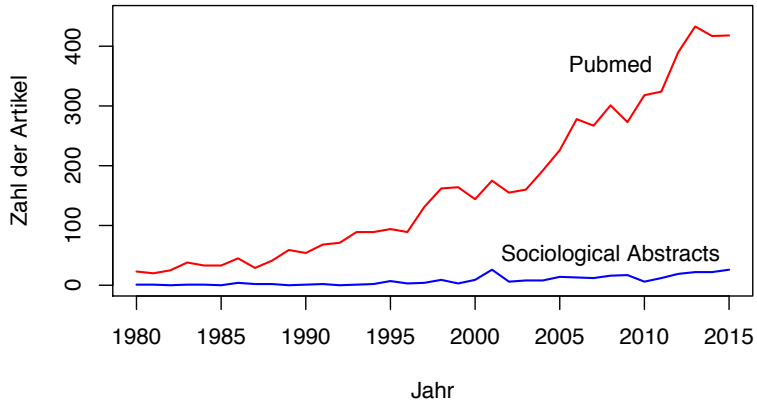

Abbildung 3: Anzahl der Publikationen pro Jahr mit dem Stichwort 'Record Linkage' in den Datenbanken Pubmed und Sociological Abstracts

Prozesse der langsamen Ausbreitung einer technischen Innovation sind Gegenstand der sozialwissenschaftlichen Innovations- und Diffusionsforschung. Auf der Grundlage der dort zumeist identifizierten Einflussfaktoren für die Dauer der Übernahme einer Innovation (Frambach und Schillewaert 2002; Wejnert 2002) lassen sich für die Verwendung administrativer Daten und/oder 'Big Data' für Forschungszwecke einige potentielle Hindernisse für die Nutzung benennen (Abbildung 4).

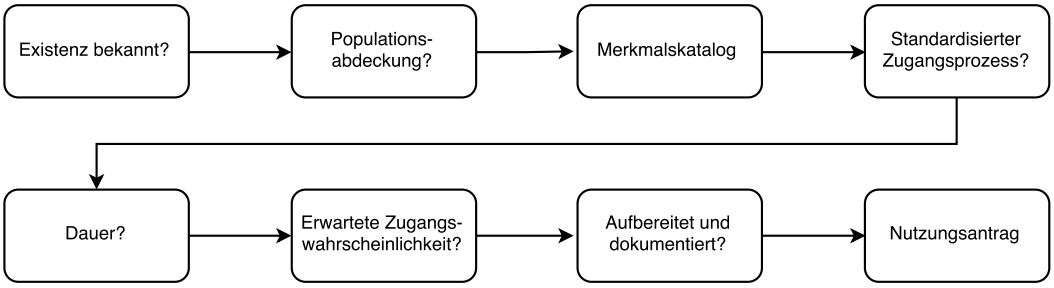

Abbildung 4: Subjektive Hürden für die Nutzung vorhandener administrativer Daten

5.1 Unkenntnis der Existenz potentieller Forschungsdaten

Offensichtlich muss für eine Nutzung die Existenz der Daten bekannt sein. Das Ausmaß der in Ämtern und Organisationen vorhandenen Informationen über Einrichtungen, Personen und Infrastrukturen wird von Außenstehenden meist unterschätzt. Hinzu kommt, dass die informationshaltenden Stellen sich häufig selbst der Tatsache nicht bewusst sind, dass es sich bei ihren Datenbeständen um forschungsrelevante Datenbestände handeln könnte. Ein Beispiel dafür sind Einsatzberichte der Feuerwehr, die Ort, Datum, Uhrzeit und Art des Ereignisses dokumentieren. Über Jahre hinweg lassen sich mit solchen Daten Unfälle, Selbstmordversuche, Brandstiftungen usw. räumlich zuordnen und zur Beschreibung sozialer Räume und damit der Lebensbedingungen in ihnen verwenden. In vielen Fällen sind diese Daten entweder bereits digital vorhanden oder aufgrund ihres hohen Strukturierungsgrades leicht digitalisierbar. Da Außenstehende von diesen Daten nichts wissen und die mit diesen Daten befassten Organisationen kein primäres Forschungsinteresse besitzen, werden diese Daten kaum genutzt.

5.2 Populationsabdeckung

Datenbanken, die vollständige Populationen abdecken, sind außerhalb administrativer Datenbanken (Arbeitslosenversicherung, Rentenversicherung, Zensus) selten. Bei vielen 'Big Data'-Datensätzen (Transaktionsdaten, 'Social Media') ist die Populationsabdeckung unklar (Biemer 2017; Golder und Macy 2014). Die Möglichkeit des Record Linkage zwischen den häufig unscharf definierten 'Big Data'-Datensätzen und administrativen Daten erlaubt prinzipiell die nachträgliche exakte Definition einer Teilpopulation oder zumindest Selektivitätsanalysen. Ohne personenbezogenes Linkage sind die statistischen Analysemöglichkeiten aber sehr begrenzt. Ist ein solches personenbezogenes Linkage nicht möglich, sollten methodische Erwägungen (oder zumindest wissenschaftliche Reviewer im Begutachtungsprozess bei Zeitschriften) die Nutzung undefinierter 'Big Data'-Datensätze einschränken.

Ob die subjektive Erwartung der kritischen Einschätzung der Verallgemeinerbarkeit der Ergebnisse durch potentielle Gutachter die Nutzung solcher Datenbestände einschränkt, ist bislang nur durch Einzelfälle zu belegen. Allerdings ist es auffällig, dass die Nutzung von 'Big Data' für akademische Forschung in der Ökonomie besonders stark verbreitet ist: Hier wurde traditionell der Definition der Untersuchungspopulation wenig Aufmerksamkeit geschenkt. Mit Ausnahme der Literatur zum 'Sample Selection Bias' spielen Stichprobenprobleme und die Definition der Grundgesamtheit in der Ökonomie kaum eine Rolle.[12] Daher ist hier auch kaum mit einem Reviewer-Einwand zu rechnen.

5.3 Merkmalskatalog

In vielen administrativen und 'Big Data'-Datensätzen ist der Kranz der verfügbaren Merkmale stark eingeschränkt. In den meisten dieser Datensätze fehlen alle subjektiven Variablen, häufig auch viele Kovariaten. Trotzdem wird das Ausmaß der vorhandenen Merkmale durch Außenstehende in der Regel unterschätzt. Welche weiteren Möglichkeiten sich für die Forschung durch die Kombination mehrerer Datenquellen (Tokle und Bender 2017) ergeben, ist inhaltlich arbeitenden Sozialwissenschaftlern häufig unklar. Die Unterschätzung der Zahl der vorhandenen Merkmale und das Unwissen der Möglichkeit der Kombination mehrerer Datenquellen führen natürlich zu einer suboptimalen Nutzung vorhandener Daten.

5.4 Die Folgen der Abwesenheit standardisierter Zugangsprozeduren

Insbesondere in Fächern mit einem hohen Publikationsdruck (Informatik, Ökonomie, Psychologie) sind zahlreiche Publikationen mit hohem Impact-Faktor über große Strecken einer akademischen Karriere von hoher Bedeutung. Die Investition eines Wissenschaftlers in den Zugang zu einem Datensatz wird daher in der Regel eher klein sein. Entsprechend irrational wäre die Investition in einen Datensatz, bei dem der Zugang unwahrscheinlich oder aufwändig wäre.

Dies gilt z.B. für den Fall, dass kein standardisierter Zugang zu einem Datensatz (klare Zuständigkeiten, Verantwortlichkeiten, Antragsprozesse und Antragsformulare) existiert. Muss der Zugang individuell geschaffen werden, wird die Dauer bis zu einem erfolgreichen Zugang entsprechend lang. Für die Genehmigung eines neuen nationalen Verknüpfungsprojekts muss international in den meisten Rechtsordnungen mit einer Dauer von mindestens zwei Jahren gerechnet werden. Für Nachwuchswissenschaftler mit Zeitverträgen sind dies kaum akzeptable Zeiträume. Der hohe Aufwand und die große Dauer lassen dann den Rückgriff auf andere Datenquellen rationaler erscheinen.

Der gleiche Mechanismus greift bei geringen subjektiven Wahrscheinlichkeit eines erfolgreichen Datenzugriffs. In Ermangelung anderer Datenquellen für eine Abschätzung der Wahrscheinlichkeit eines Zugangs dient häufig die Reputation einer Behörde als kooperativ oder unkooperativ.

Einige Behörden – wie z.B. das Kraftfahrzeugbundesamt – haben sich trotz günstiger Gesetzeslage den Ruf vollständiger Verweigerung wissenschaftlicher Kooperation hart aber ehrlich erworben.[13] Der

[12] Zur ungebrochenen Popularität der Sample-Selection-Modelle trotz ihrer Probleme (Kennedy 2008) findet sich eine Übersicht bei Certo et al. (2016).

[13] Die Homepage des KBA (https://www.kba.de/DE/Statistik/Projekte/Forschungsprojekte/forschungsprojekte_node.html) zeigt Anfang 2018 insgesamt 16 abgeschlossene Forschungsprojekte innerhalb von 33 (!) Jahren. Bei keinem Projekt ist erkennbar, dass Daten mit anderen Quellen zusammengeführt werden. Das KBA erscheint weder in der Datenbank der DFG (Gepris) noch in der Medizindatenbank Pubmed. Falls das KBA in der akademischen Forschung auf der Ebene von Mikrodaten kooperieren sollte, wird anscheinend zurückhaltend publiziert.

Versuch eine solche Behörde zu einer Kooperation zu bewegen wird erheblichen Opportunitätskosten für eine wissenschaftliche Karriere bedingen. Aus Sicht einer solchen Behörde ist eine wiederholte massive Demonstration unkooperativen Verhaltens höchst rational. Da die Bereitstellung von Daten in der Regel nicht zu den Dienstaufgaben gehört, können solche Behörden die Taktik der Verschleppung einer Anfrage über Jahre anwenden. Da die meisten akademischen Akteure rational handeln werden Kooperationsbemühungen in der Regel rasch eingestellt.

Aus Sicht der Behörde führt jede wiederholte Demonstration unkooperativen Verhaltens nur zu Gewinnen. Unter solchen rechtlichen Randbedingungen wird kein Verhaltenswandel zu erwarten sein.

Schließlich wirken eine mangelnde Datenaufbereitung und Dokumentation der vorhandenen Datensätze einer Nutzung selbst dann entgegen, wenn der Zugang rechtlich und technisch möglich ist. Bei Rohdaten wird es in diesen Fällen erforderlich, die in der Regel aufwändigen Bereinigungen und Recherchen zur nachträglichen Dokumentation der verwendeten Codes und Eingabeprozeduren selbst durchzuführen. Durch den erhöhten Aufwand für eine Publikation wird auch dies zu einer Vermeidung solcher Datensätze führen.

Zusammenfassend kann festgestellt werden, dass jedes der genannten Hemmnisse ausreicht, um die Nutzungswahrscheinlichkeit vorhandener administrativer Daten oder 'Big Data'-Datensätzen für das Record Linkage geringer werden zu lassen.

6 Organisatorische Widerstände

Neben den Hemmnissen auf der Seite oder in der Wahrnehmung der Forschenden gibt es weiter objektive Widerstände auf der Seite der Datenhalter[14] und der Organisation des Datenschutzes. Die rechtliche Situation in Deutschland bedingt dabei einige Besonderheiten, die sich so kaum in der internationalen Literatur finden. Einige dieser allgemeinen und einige der für Deutschland spezifischen Widerstände sollen kurz diskutiert werden.

6.1 Widerstand gegen Transparenz

Selbst wenn die Existenz forschungsrelevanter Daten allen Beteiligten klar ist, kann es möglich sein, dass Datenhalter deren Weitergabe nicht ermöglichen wollen. Häufig besteht kein Interesse an der Einsicht in bestehende Strukturen oder einem Vergleich zu anderen Organisationen.

Die deutlichsten Beispiele dafür finden sich im Bereich des Gesundheitswesens. Krankenhäuser haben in der Regel kein Interesse daran, dass Daten über die Ergebnisse von Eingriffen bekannt werden, vor allem dann nicht, wenn die tatsächlichen Leistungserbringer getrennt ausgewiesen werden können. Wohl auch aus diesem Grund erfolgen die Dokumentationen in der Regel so früh wie möglich auf Aggregatebene (Krankenhaus statt Arzt, Adresse der Abrechnungsstelle statt OP-Nummer etc.). Selbst wenn die technische Möglichkeit zur getrennten Ausweisung bestünde, wird von solchen Akteuren in der Regel dagegen votiert. Zwar ließen sich eventuelle Unterschiede in der Zusammensetzung der behandelten Fälle ('Casemix', Iezzoni 2003) bei den Analysen berücksichtigen, das Risiko verbleibender Unterschiede zwischen Leistungserbringern erscheint vielen Akteuren im Gesundheitssystem aber offensichtlich zu groß. Sehr ähnliche Verhältnisse finden sich natürlich auch bei Schulen oder Universitäten, wo die Probleme der Notengebung offensichtlich würden, könnte man die Unterschiede zwischen den Leistungserbringern zuordnen. Der Zugang zu Universitätsprüfungsdaten zu Forschungszwecken ist in Deutschland selbst bei ministeriell geförderten Projekten nahezu unmöglich.

Da die Datenhalter bei solchen Konstellationen weder ein Interesse daran haben, dass die Existenz solcher Daten bekannt wird, noch ein Interesse daran haben, die Daten für Analysen zur Verfügung zu stellen, wird sich ohne externen Druck keine Verhaltensänderung einstellen. Diese kann entweder — z.B. nach einer öffentlichen Skandalisierung — durch eine gesetzliche Auflage oder eine Veränderung

[14] Dazu gehören auch direkte finanzielle Interessen kommerzieller Datenhalter. Vor allem im Zusammenhang mit 'Social Media'-Daten muss darauf hingewiesen werden, dass ein definierter Zugang der wissenschaftlichen Forschung zu solchen Daten üblicherweise nur in Ausnahmefällen möglich ist (Kinsley 2014; Innes et al. 2016). Selbst in den Fällen, in denen die kommerziellen Eigentümer beschränkt Zugang gewähren, handelt es sich in der Regel nur um Teilmengen, die nicht in jedem Fall Zufallsstichproben darstellen (Morstatter et al. 2014). Solche Probleme sind aber nicht spezifisch für die Situation in Deutschland, sondern finden sich global.

des Kundenverhaltens erreicht werden. Ohne öffentlichen Druck der einen oder anderen Art wird es bei dieser Ursache nicht zu einer Erschließung neuer Datenquellen kommen.

6.2 Der Wunsch nach Beibehaltung existierender Informationsinfrastrukturen

Es gibt zahlreichen Beispiele, bei denen bereits bestehende Informationsinfrastrukturen in ihrer Existenz im Falle der Bereitstellung von Mikrodatensätzen für die Forschung gefährdet wären.

Ein Beispiel dafür ist die Erfassung und Analyse der Todesursachen in Deutschland. Mit der Bearbeitung und Kodierung der Todesursachen allein sind sowohl Mitarbeiter in allen kommunalen Gesundheitsämtern als auch in allen Landesämtern für Statistik beschäftigt (Schelhase 2014). Zwar handelt es sich hier zum größten Teil lediglich um wenige angelernte Kräfte, allerdings würden den Landesämtern Kompetenzen entzogen, würde man die fehleranfällige Kodierung (der bislang nicht bundeseinheitlichen) Totenscheine zentralisieren. Ähnliche Konstellationen finden sich in allen Feldern, in denen Informationssysteme auf Aggregatdaten basieren und aus vorgeblichen Datenschutzgründen keine Mikrodatensätze weitergegeben werden, z.B. Ärztestatistik, Prüfungsstatistik, Schulstatistik. Da die Akteure in diesen Feldern von politischen Entscheidern als Spezialisten betrachtet werden sollte man die Veto-Macht der Vertreter existierender Infrastrukturen nicht unterschätzen.

6.3 Unklare gesetzliche Anforderungen

In den meisten Ländern müssen Record Linkage-Projekte viele gesetzliche Anforderungen auf verschiedenen Ebenen erfüllen. In Deutschland sind dies zum Beispiel (vgl. auch Rat für Sozial- und Wirtschaftsdaten 2017): Europäische Datenschutzbestimmungen, das Bundesdatenschutzgesetz, die länderspezifischen Gesundheitsdatenschutz- und Personenstandsgesetze, das Bundesstatistikgesetz, das Sozialgesetzbuch sowie das Bundesmeldegesetz. Je nach Datenbestand können zusätzliche Gesetze gelten. Alle diese Bestimmungen müssen gegen das Verfassungsrecht der Freiheit der Forschung gewichtet werden. Da es kein eindeutiges positives Bundesgesetz für wissenschaftliche Datenverknüpfung gibt, erfordert nahezu jedes Projekt viele Verhandlungen mit Datenschutzbeauftragten auf mehreren Ebenen. Bei einem nationalen Forschungsprojekt dürfte das im Regelfall neben den Bundes- und Landesdatenschützern noch die kommunalen Datenschützer sein. Da auf jeder Verhandlungsebene einzelne Vetospieler ausreichen, um den Prozess um Monate oder Jahre zu verzögern oder endgültig zu blockieren, sinkt die Wahrscheinlichkeit eines erfolgreichen Projekts weiter.[15]

6.4 Unklare Organisationsverfahren

Die Abwesenheit klarer allgemeiner gesetzlichen Regelung des Zugangs zu 'Big Data' für die wissenschaftliche Forschung erschwert die Erschließung dieser Datenbestände erheblich. Da der Zugang zu 'Big Data' ohne individuelle Einwilligung der Merkmalsträger für Forschungszwecke fast immer unklar ist, sind – neben den datenschutzrechtlichen Problemen – entsprechende langwierige organisatorische Klärungsprozesse notwendig.

Es beginnt damit, dass die meisten datenhaltenden Organisationen, deren Zweck eben nicht Forschung ist, keinen klar definierten Ansprechpartner für Wissenschaftler haben. So ist in Deutschland selbst nur an wenigen Universitätskliniken eindeutig geregelt, wer für welche Daten der entsprechende Ansprechpartner ist. Schon die Klärung dieser Frage kann Wochen in Anspruch nehmen. Das nächste Problem ist die Klärung der Frage, wer denn der gesetzliche Eigentümer der Daten ist. Dann muss geklärt werden, wer über eine Anfrage entscheidet, welche Regeln zum Tragen kommen, ob es ein Widerspruchsverfahren gegen eine Entscheid gibt usw. Selbst im Falle eines positiven Entscheides, kann der tatsächliche Austausch von Daten (Format, Speichermedium, Übergabe) zu Problemen führen, da auch hier die Prozeduren und Verantwortlichen nicht zuvor festgelegt wurden. Schließlich müssen Fragen nach den Kosten des Verfahrens geklärt werden.

[15] Die Schwierigkeiten bei der Einrichtung eines deutschen Mortalitätsregisters (Mueller und Werdecker 2014) sind ein exzellentes Beispiel für die resultierenden politischen Implementierungsprobleme trotz positiver Begutachtung auf jeder Ebene.

Die Abwesenheit von Standardprozeduren für Datenanfragen führt dazu, dass jedes einzelne Forschungsprojekt diese Prozeduren immer neu ermitteln, festlegen und durchführen muss. Die Schaffung solcher Prozeduren ist für die beteiligten Wissenschaftler ein außerordentlich zeitraubender Prozess.[16]

Für die datenhaltenden Organisationen ist der Aufwand zwar geringer, kann aber von diesen in Gänze vermieden werden, in dem Anfragen prinzipiell von Anfang an nicht oder nur sehr schleppend beantwortet werden. Hier ist es dann hilfreich wenn die prinzipielle Legitimität des Anliegens kaum bezweifelt werden kann.

6.5 Legitimität der Forschung als Catch-22

Wesentlich aus der Sicht der Datenschützer ist zunächst die Erfüllung formaler gesetzlicher Anforderungen. Bei der Beurteilung des Datenschutzes in der Forschung handelt es sich fast immer um eine Güterabwägung. Es ist in keiner Weise klar, wie eine solche Abwägung zu erfolgen hat.

So fordert zum Beispiel das bundesweit gültige Personenstandsgesetz (§ 66, 3), dass „das öffentliche Interesse an der Durchführung des Forschungsvorhabens die schutzwürdigen Belange des Betroffenen an dem Ausschluss der Benutzung erheblich überwiegt“. Wann das der Fall ist, ist mangels eindeutiger Kriterien (Karaalp 2017) nicht klar.

Da der Gesetzgeber diese Kriterien nicht eindeutig festgelegt hat, bleibt bei der Interpretation erheblicher Spielraum. Das führt zu einem klassischen Catch-22: Da die Kriterien nicht klar definiert sind, ist es schwer, sie zu erfüllen.[17]

Es gibt bislang keine systematischen empirischen Studien zu den Bedingungen, die einen Datenzugang bei einer dezentralen Organisation des Datenschutzes ermöglichen oder verhindern. Vereinzelte Vorträge (Petrila 2015) und Diskussionen auf den Konferenzen des 'International Population Data Linkage Network' (IPDLN) lassen vermuten, dass die dabei stattfindenden Prozesse weltweit ähnlich ablaufen.

Zunächst muss beachtet werden, dass die lokalen Datenhalter auf der operationalen Ebene in der Regel nicht entscheidungsbefugt sind. Zumeist fällt die Entscheidung über eine Nutzung auf zwei Ebenen oberhalb der technischen Datenhaltung: Einer juristischen Ebene und einer organisatorischen Ebene.

Die organisatorische Ebene liegt auf der der Dienststellen- oder Geschäftsführung. Hier erfolgt zumeist eine erste Prüfung auf vermutete Legitimität des Anliegens und die Beurteilung, ob ein solches Projekt die Organisationsziele direkt oder indirekt gefährden könnte.

Die juristische Prüfung erfolgt bei großen Organisationen durch Volljuristen, in der Praxis aber sehr häufig durch lokale Datenschutzbeauftragte, die lediglich eine minimale Schulung erhalten haben.[18] Neben der rechtlichen Unbedenklichkeit im engeren Sinne prüfen die Datenhalter die Legitimität des Anliegens. Die im Grundgesetz verankerte Forschungsfreiheit wird von den meisten Datenhaltern in der Regel zumindest für die universitäre Forschung als ausreichende Begründung für ein Forschungsprojekt angesehen. Bei Projekten des Verfassers hat es sich mehrfach gezeigt, dass es hilfreich ist, wenn des Forschungsprojekts eine für Laien verständliche Beziehung zwischen den Daten und dem zugehörigen Fachbereich der Forschenden besitzen. Im Allgemeinen scheinen medizinische Fragestellungen deutlich problemloser erläuterbar zu sein als sozialwissenschaftliche Projekte. Forschungsprojekte aus der Informatik liegen zwischen den beiden genannten Extremen. Sollte das Forschungsprojekt zu Ergebnissen führen, deren Nutzen den Datenhaltern einsehbar ist, wird dies sicherlich dem Projekt förderlich sein. Eine weitere von den Datenhaltern offensichtlich verwendete Heuristik bei der Einschätzung der Legitimität des Anliegens ist die wahrgenommene Kompetenz der Beteiligten für das Projekt, z.B. durch Verweis auf ähnliche Projekte, sichtbare

[16] Als Beispiel soll ein Projekt des Verfassers zur Verknüpfung eines Einwohnermelderegisters, eines Klinikums und eines privaten Datenhalters in einer Großstadt erwähnt werden: Es waren 70 Verhandlungsschritte über mehr als 9 Monate notwendig, um eine Verknüpfung der Vitaldaten von Verstorbenen zu organisieren.

[17] Als Catch-22 wird im Englischen nach einer Novelle von Joseph Heller eine Situation bezeichnet, aus es der aufgrund widersprechender Regeln kein Entkommen gibt.

[18] Datenschutzbeauftragte bei Behörden und öffentlichen Stellen erhalten in Niedersachsen eine 16-tägige Schulung (Die Landesbeauftragte für den Datenschutz Niedersachsen 2018), behördliche Datenschutzbeauftragte in der Bundesverwaltung mehrere Schulungen mit zusammen ca. 24 Tagen (Bundesakademie für öffentliche Verwaltung im Bundesministerium des Innern 2016) Für die Ausbildung entscheidungsberechtigter Personen erscheinen solche kurze Zeiträume kaum ausreichend.

Publikationen oder prestigeträchtige Kooperationspartner. Zu diesen Reputationsfaktoren gehört auch die Unterstützung durch die DFG, Ministerien oder Bundeseinrichtungen.

Das Volumen der angeforderten Forschungsinformationen ist nicht ohne Einfluss auf das Entscheidungsverhalten der Datenhalter. Stichproben sind eher zugänglich als vollständige Datensätze, wobei aber der begründete Hinweis auf die Notwendigkeit der vollständige Populationsabdeckung häufig verständlich gemacht werden kann. Trivialerweise machen möglichst schwer angreifbare Pseudonymisierungen und erprobte Datenaustauschprotokolle Verhandlungen mit Datenhaltern einfacher.

Die Antizipation der genannten Heuristiken kann bei den Verhandlungen mit den Datenhaltern hilfreich sein. Die Situation der Forschenden in den meisten Ländern und besonders in Deutschland ist dabei aber die eines armen Bittstellers, ein positives Recht auf einen Datenzugang gibt es nicht. Bei allen Verhandlungen und auch bei der Abschätzung der prinzipiellen Machbarkeit eines Projekts muss beachtet werden, dass die Verhandlungspartner rationale Akteure mit ihren eigenen Interessen sind, die sich fast nie mit den Interessen der Forschung decken.

6.6 Probleme durch die technische Beurteilung der Sicherheit der Verfahren und Abläufe

Die Datenschutzanforderungen, die an ein Projekt gestellt werden, lassen sich nicht absolut allein anhand der betroffenen Daten oder der verwendeten Verschlüsselungsverfahren definieren. Bei der Beurteilung der Sicherheit eines Datensatzes gegenüber Angriffen, Missbrauch oder schlichter Fahrlässigkeit müssen die Daten sowie die damit verbundenen Datenaustauschabläufe insgesamt berücksichtigt werden ('data situation', Elliot et al. 2016). Lokalen Datenschützern und Juristen fehlen in der Regel die Kompetenzen zur Beurteilung der Gesamtheit eines Datenprojekts.

Daher wird die Entscheidung darüber, ob zumindest die technischen Abläufe und Verfahren als ausreichend sicher erscheinen, zumeist nationalen Zertifizierungseinrichtungen wie dem Bundesamt für Sicherheit in der Informationstechnik (BSI) überlassen. Der Beurteilungsprozess durch das BSI kann mehr als ein Jahr beanspruchen.[19] Wird die Record Linkage-Technik und das damit verbundene Datenaustauschprotokoll von den Zertifizierungsbehörden akzeptiert, werden in der Regel die Verfahren auch von Juristen und lokalen Datenschützern akzeptiert. Eine positive Stellungnahme des BSI besitzt in der Regel eine starke Signalwirkung, daher wird bei Großprojekten häufig eine frühe Einbeziehung des BSI angestrebt.

Es gibt aber auch hier Ausnahmen. So sieht z.B. Voitel (2017) sogar im Gebrauch allgemein akzeptierter Verfahren der Kryptographie (Hash-Funktionen für Pseudonyme) ein Problem, da diese nicht mit *absoluter* Sicherheit gegenüber Angriffen geschützt werden können. Es muss betont werden, dass eine solche Interpretation mit der europäischen Datenschutzrichtlinie kaum vereinbar ist. Dass solche Positionen aber von deutschen Datenschützern vereinzelt vertreten werden, illustriert das zentrale Problem einer mangelnden positiven Klärung der Datenverknüpfung zu Forschungszwecken deutlich.

6.7 Lokale Datenschutzbeauftragte als rationale Akteure

In einem älteren aber bemerkenswerten Artikel stellt David Mechanic (1962) fest, dass es nicht ungewöhnlich sei, dass Personen mit niedrigerem Position in Organisationen mehr Macht ausüben, als die formale Definition ihrer Positionen erwarten lässt. In Universitäten ist das Beispiel der Sekretärinnen offensichtlich. Für diese eher rangniedrigen Mitglieder einer Organisation wird ihre Macht umso größer, je mehr sie sich mit Bereichen beschäftigen, die höherrangige Mitglieder der Organisation als nicht lohnend wahrnehmen. Darüber hinaus können diese Personen in komplexen Organisationen ihr Wissen über Normen und Regeln nutzen, um versuchte Veränderungen zu vereiteln. Mechanic betont, dass die Macht dieser Personen weiter wächst, wenn sie schwer zu ersetzen sind. Im Kontext der lokalen Organisation des Datenschutzes in deutschen Verwaltungen gewinnen die Hypothesen von Mechanic besonderes Gewicht. Lokale Datenschützer sind in der Regel schwer zu ersetzen, da die Position und das Thema nicht attraktiv sind und die wahrgenommene

[19] Diese Beurteilung ist zudem meist mit erheblichen Kosten (im Bereich von mehreren Personenjahren) verbunden. Daher wird die überwiegende Mehrheit von Linkageprojekten mit bereits etablierten Lösungen arbeiten – auch wenn diese häufig technisch überholt sind.

Fachkompetenz mit der Zeit steigt. Da das Ansehen in der Hierarchie eher mit der Verweigerung einer Kooperation als mit einer Kooperation steigt, sind die Konsequenzen für routinebrechende Forschungsprojekte offensichtlich.

Durch ihre Erlaubnis, auf die Daten, für die sie verantwortlich sind, zuzugreifen, übernehmen die Datenschützer eine für ihre eigene Position gefährliche Verantwortung. Durch einen Forschungszugriff auf die Daten haben die lokalen Datenschützer absolut nichts zu gewinnen und viel zu verlieren. Daher wird ein rationaler und vorsichtiger lokaler Datenschützer aufgrund seiner persönlichen Konsequenzenbefürchtungen die sichere Alternative wählen und keinen Zugriff erlauben. Das Ergebnis sind die Daten-Silos, die wir in der Praxis sehen.

7 Erfahrungen mit Record Linkage in anderen Ländern

Aufgrund der besonderen rechtlichen Situation in Deutschland (föderale Struktur, Volkszählungsurteil 1983) sind Forschungsprojekte zum Record Linkage für 'Big Data' nur sehr begrenzt mit Projekten in anderen Ländern vergleichbar. Daher sollen exemplarisch drei Länder zum Vergleich herangezogen werden: Norwegen wäre aufgrund hervorragender Infrastrukturen besonders geeignet für solche Forschungsprojekte, nutzt die Infrastruktur aber kaum. Großbritannien erfährt eine außerordentliche politische Unterstützung für die Durchführung solcher Projekte, der Fortschritt verläuft aber eher zäh. Schließlich soll als positives Beispiel die Schweiz erwähnt werden.

7.1 Gründe für die Nichtnutzung vorhandener Datenbanken: Das Beispiel Norwegen

Da die Vorteile der Nutzung administrativer Daten unbestritten sind, stellt sich die Frage, warum selbst in Ländern, in denen die Datenbanken vorhanden sind, die Nutzung für Forschungsfragestellungen anscheinend gering ist.

Ein Beispiel dafür sind die skandinavischen Länder. Die jeweiligen nationalen Identifikationsnummern[20] sind vorhanden (Schweden: 'personnummer' 1947, Finnland: 'SETU' 1964, Norwegen: 'fødselsnummeret' 1967/1968, Dänemark: 'CPR' 1968) und Registerzusammenführungen damit technisch problemlos. Das Ausmaß wissenschaftlicher Publikationen unter Nutzung dieser Register ist aber verblüffend klein: In den Literaturdatenbanken finden sich weniger als 500 Arbeiten, die sich nicht auf Zensusprobleme beziehen. Sucht man nach Studien, die Register und Surveys verlinken, finden sich kaum 50 entsprechende Studien.[21] Die meisten dieser Arbeiten sind nicht rein sozialwissenschaftlichen Fragestellungen gewidmet, sondern behandeln überwiegend gesundheitsbezogene Fragestellungen.

Die Erklärung für die geringe Nutzung der Register für Forschung allgemein und für sozialwissenschaftliche Forschung im Besonderen muss mehrere Faktoren berücksichtigen. Zum einen sind einige Register – wie z.B. in Norwegen – faktisch erst seit 2007 verlinkbar (Maret-Ouda et al. 2017). Zum anderen sind die Forschungsinfrastrukturen vor allem in den Sozialwissenschaften bei weitem nicht so ausgeprägt wie in anderen europäischen Ländern. In Schweden z.B. gibt es nur ca. 30 Lehrstühle in den Sozialwissenschaften an den Universitäten (BDS 2011). Zusammen mit der Tatsache, dass die wenigsten Datensätze in den skandinavischen Ländern über eine englischsprachige Dokumentation verfügen, ist damit die Zahl der potentiellen Interessenten an der Durchführung sozialwissenschaftlicher Analysen auf der Basis der skandinavischen Register trotz der eindrucksvollen Liste[22] solcher Register sehr begrenzt. Berücksichtigt man, dass die Genehmigungsprozesse auch in Skandinavien mindestens ein Jahr benötigen (Maret-Ouda et al. 2017), dann wird erkennbar, dass die oben allgemein als hemmend benannten Faktoren auch hier wirken. Die geringe Zahl an Publikationen in

[20] Obwohl vor allem die deutschsprachige Wikipedia insbesondere im Bereich der Sozialwissenschaften eine höchst zweifelhafte Quelle darstellt, so ist der Artikel zu nationalen Identifikationsnummern der englischsprachigen Wikipedia die umfassendste und aktuellste auffindbare Darstellung (https://en.wikipedia.org/wiki/National_identification_number).

[21] Diese Angaben basieren auf einer Recherche in den Datenbanken 'Scopus' und 'Sociological Abstracts' im Januar 2018 durch den Autor.

[22] Als Beispiel für Norwegen, siehe https://helsedirektoratet.no/norsk-pasientregister-npr/om-npr/innhold-og-kvalitet.

den skandinavischen Ländern stellt damit kein Gegenargument gegen den Nutzen solcher Register da.

7.2 Zwei gegensätzliche Beispiele für die Nutzung administrativer Daten für die Forschung: Großbritannien und die Schweiz

In Großbritannien wurden an die Einrichtung des durch die akademische Forschungsförderung (ESRC) mit 34 Millionen Pfund unterstützten neuen 'Administrative Data Research Network' (ADRN) große Hoffnungen geknüpft. Nach drei Jahren fällt der Bericht des (Administrative Data Research Network Board 2016) ernüchternd aus. Von ca. 90 Anträgen konnten nur für 11 bis zum Berichtszeitpunkt Daten bereitgestellt werden. Weitere 17 Projekte waren immer noch in Verhandlungen mit den Datenhaltern. Mehr als die Hälfte der Projekte (47) waren noch in der Antragsphase. Das Netzwerk hat nach Aussage des Berichts erhebliche Herausforderungen erfahren bei dem Versuch einige Datenhalter zu einer Kooperation bei der Bereitstellung von Forschungsdaten zu veranlassen.[23] Wohl gemerkt: Bei einer günstigen Rechtslage, der Unterstützung durch die Regierung, der akademischen Forschungsförderung und geregelten und positiven Stellungnahmen der Ethikkommissionen. Insgesamt zeigt das Netzwerk deutlich die Schwierigkeiten der Bereitstellung administrativer Daten für Forschungszwecke.

Es gibt aber auch positive Beispiele für die Nutzung administrativer Daten für die sozialwissenschaftliche Forschung. Vorbildlich aus der Sicht der Forschung scheint hier die Schweiz zu sein. Neben einer günstigen Rechtslage (Verknüpfungsstelle 2017) findet sich hier eine eigene Record Linkage-Stelle innerhalb der amtlichen Statistik mit geregelten Zugangsbedingungen und garantierten Laufzeiten für die Beantragung (Scartazzini und Teichgräber 2017). Aufgrund dieser außerordentlich vorteilhaften Rahmenbedingungen ist zu erwarten, das Studien zum Lebenslauf oder zu Nonresponse und Response-Errors auf der Basis der Daten der Schweiz bald in ihre jeweiligen Gebiete führend sein werden.

8 Ein Lösungsansatz für Deutschland: Schaffung von Data Privacy Boards

Wie gezeigt wurde, stehen der stärkeren Nutzung administrativer und sonstiger 'Big Data' Quellen viele Hemmnisse und Regelungen im Wege. Das größte Problem in Deutschland sind dabei mit großem Abstand die rechtlichen und organisatorischen Regelungen.

Die dezentrale und föderale Datenschutzinfrastruktur mit mehrfachen und unklaren Zuständigkeiten bei Forschungsprojekten zusammen mit der Abwesenheit eines expliziten Forschungsprivilegs für Record Linkage-Projekte lassen die Rechtslage in Deutschland für Forschungsprojekte ohne expliziten Konsens der Merkmalsträger kaum aussichtsreich erscheinen. Ein großer Teil der Record Linkage Projekte in den Sozialwissenschaften konzentriert sich daher auf Fälle, bei denen die Einwilligung der Befragten eingeholt werden kann (Sakshaug et al. 2017). Dieser einfach scheinende Weg ist für Projekte, die eine vollständige Population abdecken sollen, praktisch unmöglich. Selbst bei kleinen Datensätzen erzeugt der Weg über die immer selektive Einwilligung ein prinzipiell unlösbares Nonresponseproblem (Schnell 1997), das man eigentlich durch das Record Linkage vermeiden könnte.

Daher sind andere rechtliche Regelungen erforderlich. Die Erfahrungen bei der Einrichtung eines Mortalitätsregisters und der Umgang des Innenministeriums mit dem Zensus 2021 lassen eine Unwilligkeit der Bundesregierung erkennen, sich trotz einer 2/3-Mehrheit im Bundestag (2013-2017) nicht mit der Lösung der Datenschutzprobleme der Forschung und der amtlichen Statistik zu beschäftigen. Die Umsetzung der EU-Datenschutzrechtlinie (Council of the European Union 2016) in deutsches Recht könnte Abhilfe schaffen, auch wenn die Erfolgswahrscheinlichkeit dafür vermutlich gering ist.

Um trotz der besonderen rechtlichen Rahmenbedingungen in Deutschland zu einer einfachen und schnellen Lösung zu gelangen, erscheint die Schaffung zusätzlicher institutioneller Regelungen geeignet. Diese zusätzlichen Einrichtungen sollen ähnlich wie die Institutional Review Boards

[23] Zwei Beispiele für Begründungen: '. . . is not currently considering any requests for data access for research purpose' und 'no project fulfils requirements of legal gateway'.

(IRBs) in der Medizin allgemein oder den Sozialwissenschaften (wie in Großbritannien) wirken. Entsprechende Institutionen, 'Data Privacy Boards', sollten das Re-Identifikationsrisiko durch die Pseudoidentifikatoren, das statistische Offenlegungsrisiko und den erwarteten Nutzen der aus einem Datenprojekt resultierenden Forschungsdatenbank überprüfen.

Benötigt wird lediglich eine einzige zusätzliche positive gesetzliche Bestimmung: Wenn ein 'Data Privacy Board' einer Verknüpfung zustimmt, sollten alle lokalen, regionalen und föderalen Datenhalter und Datenschutzbeauftragten ein expliziten Rechtsanspruch haben, von allen möglichen Rechtsansprüchen und Konsequenzen eines Linkageprojekts befreit zu werden.

Die Schaffung solcher 'Data Privacy Boards' und vor allem die explizite Befreiung von den rechtlichen Konsequenzen eines Linkage-Projekts für die nachgeordneten Datenhalter und Datenschützer ist ein notwendiger Schritt zur Förderung der Verknüpfung von Forschungsdaten aller Art in Deutschland. Ohne diese Regelung wird in Deutschland eine breite Nutzung von 'Big Data' für Forschungszwecke in den Sozialwissenschaften und der Medizin kaum stattfinden.

Danksagung

Für auch kontroverse Diskussionen im Zusammenhang mit diesem Beitrag danke ich Manfred Antoni, Christian Borgs, Jonas Klingwort, Günther Heller, Johannes Kopp, Winfried Pohlmeier und vor allem Stefan Bender. Die Verantwortung für alle Aussagen in diesem Text trägt nur der Verfasser.

Literatur

Administrative Data Research Network Board. (2016). *Second Annual Report*. UK Statistics Authority. London. Zugriff unter https://www.statisticsauthority.gov.uk/wp-content/uploads/2016/10/ADRN-AR-15-16.pdf

Angrist, J. D. & Krueger, A. B. (1999). Empirical strategies in labor economics. In O. C. Ashenfelter & D. Card (Hrsg.), *Handbook of labor economics* (Bd. 3, S. 1277–1366). Amsterdam: Elsevier.

Antoni, M. & Schnell, R. (2017). The Past, Present and Future of the German Record Linkage Center (GRLC). *Journal of Economics and Statistics*. doi:10.1515/jbnst-2017-1004

BDS. (2011). SoziologInnen in den skandinavischen Ländern. *BDS Newsletter*. Zugriff unter http://bds-soz.de/BDS/PDF/Internationales/skandinavien.pdf

Biemer, P. (2017). Errors and Inference. In I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter & J. Lane (Hrsg.), *Big data and social science* (Kap. 10, S. 265–297). Boca Raton: CRC Press.

Bundesakademie für öffentliche Verwaltung im Bundesministerium des Innern. (2016). Leitfaden Behördliche Datenschutzbeauftragte in der Bundesverwaltung. Zugriff unter http://www.bakoev.bund.de/SharedDocs/Downloads/LG_5/BDSB/Leitfaden.pdf

Certo, S. T., Busenbark, J. R., Woo, H.-s. & Semadeni, M. (2016). Sample selection bias and Heckman models in strategic management research. *Strategic Management Journal*, *37*(13), 2639–2657. doi:10.1002/smj.2475

Chetty, R. (2012). *Time Trends in the Use of Administrative Data for Empirical Research*. NBER Summer Institute, July 2012. Zugriff unter http://www.rajchetty.com/chettyfiles/admin_data_trends.pdf

Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Berlin: Springer.

Connelly, R., Playford, C. J., Gayle, V. & Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research*, *59*(Supplement C), 1–12. doi:https://doi.org/10.1016/j.ssresearch.2016.04.015

Council of the European Union. (2016). Council regulation (EU) no 679/2016: On the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

D'Orazio, M., Zio, M. D. & Scanu, M. (2006). *Statistical Matching: Theory and Practice*. Chichester: Wiley.

Die Landesbeauftragte für den Datenschutz Niedersachsen. (2018). Bausteinreihe Basiswissen Datenschutz. Zugriff unter https://www.lfd.niedersachsen.de/startseite/fortbildung_service/datenschutzinstitut_niedersachsen_dsin/programm_2018/bausteinreihe_basiswissen_behoerdliche_datenschutzbeauftragte/bausteinreihe-basiswissen-fuer-behoerdliche-datenschutzbeauftragte-138862.html

Ekland-Olson, S. & Gibbs, J. P. (2017). *Science and sociology: predictive power is the name of the game*. Abingdon: Routledge.

Elliot, M., Mackey, E., O'Hara, K. & Tudor, C. (2016). *The Anonymisation Decision-Making Framework*. Manchester: UKAN.

Frambach, R. T. & Schillewaert, N. (2002). Organizational innovation adoption: a multi-level framework of determinants and opportunities for future research. *Journal of Business Research*, *55*(2), 163–176. doi:https://doi.org/10.1016/S0148-2963(00)00152-1

Golder, S. A. & Macy, M. W. (2014). Digital Footprints: Opportunities and Challenges for Online Social Research. *Annual Review of Sociology*, *40*, 129–152.

Guo, S. & Fraser, M. W. (2010). *Propensity Score Analysis: Statistical Methods and Applications*. Advances Quantitative Techniques in the Social Sciences. Los Angeles: Sage.

Herzog, T. N., Scheuren, F. J. & Winkler, W. E. (2007). *Data quality and record linkage techniques*. New York: Springer.

Iezzoni, L. I. (Hrsg.). (2003). *Risk Adjustment for Measuring Health Care Outcomes* (3. Aufl.). Chicago, IL: Health Administration Press.

Innes, M., Roberts, C., Preece, A. & Rogers, D. (2016). Of instruments and data: Social media uses, abuses and analysis. In N. G. Fielding, R. M. Lee & G. Blank (Hrsg.), *The SAGE Handbook of Online Research Methods* (S. 108–124). London: Sage.

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., ... Usher, A. (2015). Big Data in Survey Research: AAPOR Task Force Report. *Public Opinion Quarterly*, *79*(4), 839–880.

Karaalp, R. N. (2017). *Der Schutz von Patientendaten für die medizinische Forschung in Krankenhäusern*. Wiesbaden: Springer.

Kennedy, P. (2008). *A Guide to Econometrics* (6. Aufl.). Malden: Blackwell Publishing.

Kinsley, B. (2014). A political economy of Twitter data? Conducting research with proprietary data is neither easy nor free. Zugriff unter http://bit.ly/1DboYPg

Maret-Ouda, J., Tao, W., Wahlin, K. & Lagergren, J. (2017). Nordic registry-based cohort studies: Possibilities and pitfalls when combining Nordic registry data. *Scandinavian Journal of Public Health*, *45*(17_suppl), 14–19. doi:doi:10.1177/1403494817702336

Martini, M. & Wenzel, M. (2017). *Rechtliche Grenzen einer Personen- bzw. Unternehmenskennziffer in staatlichen Registern*. Universität Speyer. Speyer.

Mechanic, D. (1962). Sources of Power of Lower Participants in Complex Organizations. *Administrative Science Quarterly*, *7*(3), 349–364.

Metschke, R. (2010). Record Linkage from the Perspective of Data Protection. In G. D. F. (RatSWD) (Hrsg.), *Building on progress: Expanding the research infrastructure for the social, economic, and behavioral sciences (Band 1)* (Bd. 2, S. 643–656). Opladen: Budrich UniPress.

Morstatter, F., Pfeffer, J. & Liu, H. (2014). When is it biased?: assessing the representativeness of twitter's streaming API. In *Proceedings of the 23rd International Conference on World Wide Web* (S. 555–556). ACM.

Mueller, U. & Werdecker, A. (2014). Bedeutung von Leichenschau- und Sektionsdaten für ein bundeseinheitliches Mortalitätsregister. In B. Madea (Hrsg.), *Die ärztliche Leichenschau: Rechtsgrundlagen, Praktische Durchführung, Problemlösungen* (S. 227–238). Berlin/Heidelberg: Springer.

Petrila, J. (2015). *Challenges in Using Administrative Data: Legal, Technical and Political*. Presentation OPRE Methods Meeting. Zugriff unter https://opremethodsmeeting.org/docs/2015/Petrila_ChallengesAdministrativeData.pdf

Randall, S. M., Ferrante, A. M., Boyd, J. H., Brown, A. P. & Semmens, J. B. (2016). Limited privacy protection and poor sensitivity: Is it time to move on from the statistical linkage key-581? *Health Informatics Management Journal*, *45*(2), 71–79. doi:10.1177/1833358316647587

Rat für Sozial- und Wirtschaftsdaten. (2017). *Handreichung Datenschutz*. RatSWD Output Series. Berlin.

Sakshaug, J. W., Hülle, S., Schmucker, A. & Liebig, S. (2017). Exploring the Effects of Interviewer- and Self-Administered Survey Modes on Record Linkage Consent Rates and Bias. *Survey Research Methods*, *11*(2), 171–188.

Scartazzini, R. & Teichgräber, M. (2017). *Rechtliche Grundlagen der Verknüpfung von Daten der öffentlichen Statistik und Bewilligungsprozesse im BFS*. Vortrag im Rahmen des Workshops 'Verknüpfung statistischer Daten - Erfahrungen, Möglichkeiten, Grenzen und Perspektiven Workshop FORS - BFS, 11. April 2017', Lausanne.

Schaar, P. (2014). Anonymisieren und Pseudonymisieren als Möglichkeit der Forschung mit sensiblen, personenbezogenen Forschungsdaten. In C. Lenk, G. Duttge & H. Fangerau (Hrsg.), *Handbuch Ethik und Recht der Forschung am Menschen* (S. 95–100). Springer.

Schelhase, T. (2014). Die Todesursachenstatistik der Statistischen Ämter des Bundes und der Länder. In B. Madea (Hrsg.), *Die ärztliche Leichenschau: Rechtsgrundlagen, Praktische Durchführung, Problemlösungen* (S. 217–225). Berlin/Heidelberg: Springer.

Schnell, R. (1997). *Nonresponse in Bevölkerungsumfragen: Ausmaß, Entwicklung und Ursachen*. Opladen: Leske+Budrich.

Schnell, R. (2012). *Survey-Interviews: Methoden standardisierter Befragungen*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Schnell, R., Bachteler, T. & Reiher, J. (2009). Privacy-Preserving Record Linkage Using Bloom Filters. *BMC Medical Informatics and Decision Making*, *9*(41), 1–11.

Schnell, R. & Borgs, C. (2015). Building a National Perinatal Database Without the Use of Unique Personal Identifiers. In *Proceedings of the 2015 IEEE 15th International Conference on Data Mining Workshop* (S. 232–239).

Smith, D. (2017). Secure Pseudonymisation for Privacy-preserving Probabilistic Record Linkage. *Journal of Information Security and Applications*, *34*, 271–279.

Tokle, J. & Bender, S. (2017). Record Linkage. In I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter & J. Lane (Hrsg.), *Big data and social science* (Kap. 3, S. 71–92). Boca Raton: CRC Press.

Vatsalan, D., Christen, P. & Verykios, V. S. (2013). A Taxonomy of Privacy-Preserving Record Linkage Techniques. *Information Systems*, *38*(6), 946–969.

Verknüpfungsstelle. (2017). *Verknüpfungsrichtlinen, Version 1.1*. Bundesamt für Statistik BFS. Bern.

Voitel, B. (2017). Sind Hash-Werte personenbezogene Daten? *Datenschutz und Datensicherheit*, *41*(11), 686–687.

Wejnert, B. (2002). Integrating models of diffusion of innovations: A conceptual framework. *Annual review of sociology*, *28*, 297–396.

Winkler, W. E. (2009). Record linkage. In D. Pfeffermann & C. Rao (Hrsg.), *Handbook of Statistics Band 29a, Sample surveys: Design, methods and applications.* (S. 351–380). Amsterdam: Elsevier, North-Holland.

IMPRINT

Publisher

German Record-Linkage Center
Regensburger Str. 100
D-90478 Nuremberg

Editors

Stefan Bender, Rainer Schnell

Template layout

Christine Weidmann

Download

www.record-linkage.de

The German Record Linkage Center is funded by the German Research Foundation (DFG).