FDZ-Arbeitspapier
Nr. 20

Ulrich Kaiser
Joachim Wagner

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

Neue Möglichkeiten zur Nutzung vertraulicher amtlicher Personen- und Firmendaten

2007

FDZ-Arbeitspapier
Nr. 20

Ulrich Kaiser
Joachim Wagner

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

Neue Möglichkeiten zur Nutzung vertraulicher amtlicher Personen- und Firmendaten

2007

Herausgeber: Statistische Ämter des Bundes und der Länder
Herstellung: Landesamt für Datenverarbeitung und Statistik

Fachliche Informationen
zu dieser Veröffentlichung:

Forschungsdatenzentrum der
Statistischen Landesämter
– Geschäftsstelle –
Tel.: 0211 / 9449-2876
Fax.: 0211 / 9449-8087
forschungsdatenzentrum@lds.nrw.de

Informationen zum Datenangebot:

Statistisches Bundesamt
Forschungsdatenzentrum

Tel.: 0611 / 75-4220
Fax: 0611 / 72-3915
forschungsdatenzentrum@destatis.de

Forschungsdatenzentrum der
Statistischen Landesämter
– Geschäftsstelle –
Tel.: 0211 / 9449 2876
Fax: 0211 / 9449 8087
forschungsdatenzentrum@lds.nrw.de

Erscheinungsfolge: unregelmäßig
Erschienen im Juni 2007

Diese Publikation wird kostenlos als **PDF-Datei** zum Download unter www.forschungsdatenzentrum.de angeboten.

Bei den enthaltenen statistischen Angaben handelt es sich um eigene Arbeitsergebnisse des genannten Autors im Zusammenhang mit der Nutzung der Forschungsdatenzentren. Es handelt sich hierbei ausdrücklich nicht um Ergebnisse der Statistischen Ämter des Bundes und der Länder.

Neue Möglichkeiten zur Nutzung vertraulicher amtlicher Personen- und Firmendaten*

*Ulrich Kaiser[1] und Joachim Wagner[2]**

Abstract:

Researchers in Germany have nowadays access to confidential micro data compiled from official statistics in a way that could not have been dreamt of just a few years ago. This paper describes the new institutions that grant data access - most importantly the research data centers located inside the data producing agencies – and how to access the micro data, and presents information about selected recently released data sets with a high potential for scientific research and policy evaluation. Furthermore, we contrast the German situation with the Danish way of handling research access to confidential micro data. Finally, we take a look at ongoing projects that will further improve data access in Germany.

The times they are a-changin'

Bob Dylan

1. Motivation

Die Arbeit mit Mikrodaten – Daten über einzelne Personen oder Firmen – gehört seit langer Zeit zum Alltagsgeschäft empirisch arbeitender Ökonomen und vieler anderer Sozialwissenschaftler. Hierfür werden vor allem Daten aus Stichproben verwendet, bei denen die Teilnahme freiwillig ist. Hierbei beeinträchtigen oft fehlende Teilnahmebereitschaft und Verweigerung von Auskünften bei als sensibel eingeschätzten Fragen das Analysepotenzial des Datenmaterials. Darüber hinaus sind die Fallzahlen dieser Datensätze aus Kostengründen in der Regel so klein, dass differenzierte Analysen für spezifische Gruppen wie etwa Hochschulabsolventen einer bestimmten Fachrichtung oder Firmen aus einer bestimmten Industrie nicht möglich sind. Prominente Beispiele für solche Stichprobendaten aus Deutschland sind die Personen- und Haushaltsdaten der Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften ALLBUS (Terwey 2000) und

* Für hilfreiche Hinweise danken wir Stefan Bender, Bernhard Bookmann, Michael Fritsch, Katrin Hussinger, Johann Moritz Kuhn, Timm Körting, Anja Münch, Ramona Pohl, Nadine Riedel, Martin Rosemann, Andreas Stephan, Sylvia Zühlke, Markus Zwick und Thomas Zwick. Die Beschreibung der dänischen Daten hat von aktueller und früherer Zusammenarbeit mit Helle Månsson und Rasmus Jørgensen profitiert. Alle verbleibenden Fehler in diesem Aufsatz verantworten wir ganz allein.
Eine weitere Version des Aufsatzes ist als Working Paper Nr. 48 an der Universität Lüneburg (ISSN 1860-5508) erschienen.
[1]Prof. Dr. Ulrich Kaiser, Department of Business and Economics, Universitiy of Southern Denmark at Odense, Campusvej 55, DK-5230 Odense M, e-mail: uka@sam.sdu.dk
[2]Prof. Dr. Joachim Wagner, Leuphana Universität Lüneburg, e-mail: wagner@uni-lueneburg.de

des Sozio-ökonomischen Panels SOEP (Wagner, Frick und Schupp 2007) sowie die Betriebsdaten aus dem IAB Betriebspanel (Kölling 2000).

Neben diesen in verschiedener Hinsicht eingeschränkten Stichproben gibt es eine Vielzahl von Datensätzen, die sich durch eine sehr große Anzahl von Merkmalsträgern (oft in Form der Grundgesamtheit) auszeichnen und die auf der Grundlage gesetzlicher Regelungen erstellt werden, in denen eine Auskunftspflicht der Personen oder Firmen vorgeschrieben ist. Diese Daten, die aus Erhebungen der amtlichen Statistik stammen (wie z.B. aus regelmäßigen Befragungen von Betrieben) oder die als „prozessproduzierte" Daten Ergebnis von Verwaltungshandlungen sind (wie die Statistik der sozialversicherungspflichtig Beschäftigten), sind für umfassende und methodisch angemessene wissenschaftliche Untersuchung zahlreicher Fragestellungen die einzig verlässliche Datenbasis. Ein Zugang zu diesen Mikrodaten ist für Wissenschaftler, die nicht Mitarbeiter der datenproduzierenden Institutionen sind, nicht ohne weiteres möglich. Hierfür gibt es neben den gesetzlichen Regelungen auch weitere gut nachvollziehbare Gründe – kein Unternehmer will z. B. Geschäftsgeheimnisse, die er den statistischen Ämtern mitteilen muss, seiner Konkurrenz zugänglich machen, und niemand will seinen neugierigen Nachbarn Einblick in seine Steuererklärung geben.

Eine Nutzung der vertraulichen Mikrodaten aus der amtlichen Statistik ist aber für externe Wissenschaftler in vielen Fällen durchaus möglich – wenn auch nicht immer ohne eine vorherige Anonymisierung, die eine Re-identifikation von Merkmalsträgern verhindert, und oft verbunden mit einigem (wenn auch geringem) bürokratischen Aufwand. Gegenüber der Situation am Anfang dieses Jahrhunderts, die ausführlich im Gutachten der Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (KVI) aus dem Jahr 2001 dokumentiert ist, haben sich die Zugangsmöglichkeiten zu diesen Daten in den vergangenen Jahren deutlich verbessert. Heute kann jeder Wissenschaftler, der in einer Einrichtung mit der Aufgabe der unabhängigen wissenschaftlichen Forschung arbeitet, mit geringem Aufwand einen umfangreichen und ständig wachsenden Bestand an Mikrodaten aus Erhebungen der amtlichen Statistik und an prozessproduzierten Mikrodaten für Untersuchungen nutzen. Wie dies möglich ist und welches Potential für empirische Untersuchungen damit erschlossen wird, darüber informiert unser Beitrag.

2. Neue Zugangswege zu vertraulichen Mikrodaten in Deutschland

Für externe Wissenschaftler gibt es grundsätzlich drei Möglichkeiten der Arbeit mit vertraulichen amtlich erhobenen bzw. bei Verwaltungsabläufen anfallenden prozessproduzierten Daten für Personen und Haushalte sowie Betriebe und Unternehmen:[3]

- Verwendung von faktisch anonymisierten Datenbeständen, die als so genannte Scientific-Use-Files (SUFs) den Wissenschaftlern im Rahmen von Nutzungsverträgen zur Auswertung an ihren Arbeitsplätzen zur Verfügung gestellt werden. Mikrodaten werden dabei als „faktisch anonym" angesehen, wenn die

[3] Hinzuweisen ist darauf, dass es darüber hinaus für die Verwendung in der Lehre so genannte Campus Files gibt, die absolut anonymisierte Einzeldaten enthalten. Durch eine entsprechende Behandlung des Originalmaterials (z.B. Klassifizierung der Betriebsgröße statt Angabe der Beschäftigtenzahl, Zusammenfassung der Angaben zum Wirtschaftszweigen zu größeren Bereichen, Aussortieren von Unternehmen mit mehr als 250 Beschäftigten, keine Weitergabe der Regionalzuordnung etc.) ist eine Identifikation einzelner Merkmalsträger in diesen Daten nicht mehr möglich. Solche Datensätze werden auch als Public-Use-Files (PUFs) bezeichnet; vgl. hierzu www.forschungsdatenzentrum.de/campus-file.asp sowie Zwick (2007).

Einzelangaben nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft den Merkmalsträgern zugeordnet werden können. Solche SUFs gibt es bisher vor allem für Personen- und Haushaltsdaten; Verfahren zur entsprechenden Anonymisierung von Daten für Betriebe und Unternehmen sind – insbesondere was Paneldaten betrifft – ein aktuelles Forschungsthema.[4] Erste Scientific-Use-Files stehen jedoch auch bereits im Bereich der Unternehmens- und Betriebsdaten zur Verfügung.[5] Der wesentliche Unterschied besteht darin, dass Unternehmens- und Betriebsdaten aufgrund des größeren Reidentifikationsrisikos in der Regel mit datenverändernden Anonymisierungsverfahren bearbeitet werden müssen, während bei Personen- und Haushaltsdaten traditionelle informationsreduzierende Verfahren oder sogar das Entfernen der direkten Identifikatoren (Namen, Adressen) ausreichen („formale Anonymität“).

- Nutzung der kontrollierten Datenfernverarbeitung (KDFV), bei der die Wissenschaftler Auswertungsprogramme an die Datenanbieter senden, die dann dort gerechnet und deren Ergebnislisten anschließend auf geheim zu haltende Angaben überprüft werden, bevor sie an die Wissenschaftler zurück geschickt werden.
- Arbeit mit nur schwach projektbezogen anonymisierten Mikrodaten in den Räumen der Dateneigner an (speziell abgeschotteten) Benutzerarbeitsplätzen. Hierbei ist dies oft der erste Schritt der Arbeit mit einem Datensatz, an den sich dann weitere Arbeiten in Form der KDFV anschließen können. Insbesondere bei der Nutzung von sehr komplexen Datenbeständen ist diese Möglichkeit zu empfehlen.

In begrenztem Umfang gibt es diese Arten des Zugangs zu vertraulichen Mikrodaten bereits seit einigen Jahren. In jüngster Zeit konnten wesentliche Fortschritte beim Ausbau der informationellen Infrastruktur in Deutschland erreicht werden. Hierbei spielt der erstmals im November 2004 auf Empfehlung der KVI-Kommission vom Bundesministerium für Bildung und Forschung (BMBF) berufene Rat für Sozial- und Wirtschaftsdaten (RatSWD) eine Schlüsselrolle. Sein wesentliches Anliegen ist es, den Zugang zu und die Qualität von Mikrodaten nachhaltig zu verbessern (vgl. Rat für Sozial- und WirtschaftsDaten (2007) und www.ratswd.de).

Unterstützt vom RatSWD und finanziell gefördert vom BMBF entstanden Forschungsdatenzentren bei den großen Datenproduzenten, nämlich der Bundesagentur für Arbeit (Kohlmann 2005), dem Statistischen Bundesamt und den Statistischen Ämtern der Länder (Zühlke und Christians 2006; Zühlke et al. 2004) sowie der Deutschen Rentenversicherung (Rehfeld und Mika 2006). Darüber hinaus wurden Datenservicezentren eingerichtet – das German Microdata Lab (GML) als Servicezentrum für Mikrodaten der Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen (GESIS) (Lüttinger et al. 2004) und das Internationale Datenservicezentrum des Forschungsinstituts zur Zukunft der Arbeit (IZA) – zu deren Aufgaben u. a. eine umfassende Dokumentation von Datenbeständen Dritter und eine qualifizierte Betreuung von Nutzern durch Workshops u. ä. gehört.[6]

[4] Vgl. hierzu Ronning u. a. (2005), Gottschalk (2005) oder Drechsler et al. (2007) und die Beiträge in Pohlmeier, Ronning und Wagner (2005) sowie Lenz u. a. (2006).
[5] Vgl. Lenz, Vorgrimler und Rosemann (2005), Scheffler (2005), Sturm und Lenz (2005) und Vorgrimler, Dittrich, Lenz und Rosemann (2005).
[6] Weitere Forschungsdaten- und Datenservicezentren sind im Aufbau bzw. in Planung. Darüber hinaus bieten die Deutsche Bundesbank und die Kreditanstalt für Wiederaufbau Wissenschaftlern Zugangswege zu Datenbeständen. Eine aktuelle Übersicht mit Links zu den Homepages der einzelnen Einrichtungen, auf denen Angaben zu Einzelheiten des Datenzugangs zu finden sind, steht auf der Webseite des RatSWD http://www.ratswd.de/dat/fdz.php

Die Arbeit mit Daten aus den Forschungsdatenzentren setzt voraus, dass potentielle Nutzer in Einrichtungen mit der Aufgabe der unabhängigen wissenschaftlichen Forschung arbeiten, und dass sie einen Antrag auf Nutzung eines genau spezifizierten Datensatzes für ein spezifisches, zeitlich begrenztes Forschungsprojekt genehmigt bekommen. Je nach gesetzlicher Grundlage sieht das Verfahren hierbei unterschiedlich aus. Zwei Beispiele sollen dies erläutern:

Wenn ein Wissenschaftler mit Mikrodaten arbeiten will, die aus Beständen des Statistischen Bundesamtes oder der Statistischen Ämter der Länder stammen und damit die Vorschriften des Bundesstatistikgesetzes (BStatG) zu befolgen sind, dann wendet er sich am besten an das Forschungsdatenzentrum des für ihn räumlich am nächsten liegenden Statistischen Amtes. Das Formular für einen Nutzungsantrag findet man auf der entsprechenden Homepage.[7] Die Erfahrung zeigt, dass eine frühzeitige Kontaktaufnahme mit den Mitarbeitern des Forschungsdatenzentrums schon bei der Antragstellung sehr hilfreich ist. Anträge werden erfahrungsgemäß zügig bearbeitet und genehmigt; allerdings ist wegen des föderalen Aufbaus der amtlichen Statistik in Deutschland und der daraus folgenden Pflicht der Genehmigung jedes Antrags durch die jeweils betroffenen Statistischen Ämter hierbei immer eine gewisse Wartezeit einzuplanen. Die Kosten für die Datenbereitstellung sind nicht Null, betragen aber (dank Subventionierung durch das BMBF) lediglich 65 € pro Datensatz und Jahr.

Für die Arbeit mit vertraulichen Mikrodaten aus den Beständen der Bundesagentur für Arbeit (BA), deren Nutzung im Sozialgesetzbuch (SGB) geregelt ist, gelten andere Vorschriften. Hier greift der Schutz des Sozialgeheimnisses (§35 SGB I). Ein Zugang zu diesen Daten ist für Forscher nur möglich, wenn die Vorgaben des §75 SGB X erfüllt sind. Dies bedeutet, dass die Daten für ein Projekt aus dem Sozialleistungsbereich verwendet werden (hierzu zählen z. B. alle Fragestellungen im Zusammenhang mit Erwerbstätigkeit oder Arbeitslosigkeit), dass das öffentliche Interesse das Geheimhaltungsinteresse der Betroffenen erheblich überwiegen muss, dass das Einholen der Einwilligung aller Betroffenen unzumutbar sein muss, und dass der Zweck des Forschungsvorhabens nicht auf andere Weise erfüllt werden kann. Details zu den Regeln mit Antragsformularen für den Zugang zu Daten bei Gastaufenthalten im FDZ der BA und eine anschließende Nutzung in Form der kontrollierten Datenfernverarbeitung finden sich auf der Homepage des FDZ (http://fdz.iab.de/). Jeder Antrag auf Nutzung pseudoanonymisierter Originaldaten wird auch vom Bundesministerium für Arbeit und Soziales (BMAS) geprüft; nach erfolgreicher Prüfung wird dann ein förmlicher Vertrag mit dem Nutzer geschlossen. Erfahrungsgemäß kann es von der Antragstellung bis zur Genehmigung etwa einen Monat dauern. Kosten für die Datennutzung fallen nicht an.[8]

3. Datens(ch)ätze

Die Datensätze – oder aus der Sicht der (potentiellen) Nutzer besser: die Datenschätze – in den Forschungsdatenzentren und Datenservicezentren umfassen schon heute ein breites Spektrum sehr

[7] Adressen und Ansprechpartner findet man unter http://www.forschungsdatenzentrum.de/ unter dem Link „Kontakt"; unter dem Link „Nutzungsantrag" ist dort auch das Antragsformular verfügbar.

[8] Im Prinzip gelten diese Regelungen auch für die vom Forschungsdatenzentrum der Rentenversicherung bereitgestellten Datenbestände; Einzelheiten findet man auf der Homepage (http://forschung.deutsche-rentenversicherung.de).

unterschiedlicher Bereiche. Ein Stöbern auf den Homepages der Zentren oder ein gezieltes Suchen nach Mikrodaten für die Analyse spezifischer Fragestellungen lohnt sich. Exemplarisch werden hier einige Datensätze vorgestellt, die für Untersuchungen zu zahlreichen ökonomischen Themen besonders geeignet sind, wobei zusätzlich ein besonderer Schwerpunkt auf neue bzw. neu für externe Wissenschaftler zugängliche Daten gelegt wird.[9]

3.1 Personendaten

3.1.1 Mikrozensus und Mikrozensus-Panel

Der Mikrozensus (Schwarz 2001) ist mit einem Auswahlsatz von 1 % der Bevölkerung die größte jährlich stattfindende Haushaltsbefragung in Deutschland. Er wird seit 1957 in Westdeutschland und seit 1991 auch in den neuen Bundesländern erhoben. Durch die sehr hohe Teilnehmerzahl mit gesetzlicher Auskunftspflicht und eine ausgeprägte Kontinuität im Fragenprogramm eignen sich diese Daten sehr gut für Untersuchungen auch spezifischer Teilgruppen von Personen sowie für Vergleiche über einen langen Zeitraum. Das jährliche Grundprogramm umfasst Fragen zu folgenden Themenbereichen: sozio-demographische Angaben, Staatsangehörigkeit, Familien- und Haushaltszusammenhänge, Wohnung, Erwerbsbeteiligung und Arbeitssuche, Schulbesuch und Ausbildungsabschluss, Quellen des Lebensunterhalts, Einkommen und Rentenversicherung. Daneben gibt es ein jährliches Ergänzungsprogramm, in dem für 0.45 % der Bevölkerung Angaben aus den Bereichen frühere und gegenwärtige Erwerbstätigkeit, Aus- und Weiterbildung sowie Situation ein Jahr vor der Befragung erhoben werden; ferner wird ein vierjähriges Zusatzprogramm (z. B. mit Fragen zum Bereich Gesundheit) durchgeführt.[10]

Scientific-Use-Files (faktisch anonymisierte 70%-Stichproben) des Mikrozensus sind zur Zeit für die Jahre 1973, 1976, 1982, 1989, 1991, 1993 sowie 1995 bis 2004 gegen eine Gebühr von je 65 € über die FDZ des Statistischen Bundesamtes bzw. der Statistischen Ämter der Länder zu erhalten. Daten für weitere Jahre – insbesondere auch für das Jahr 2005 – sind über kontrollierte Datenfernverarbeitung und auf Gastwissenschaftler-Arbeitsplätzen verfügbar
(vgl. http://www.forschungsdatenzentrum.de/bestand/mikrozensus/index.asp).

Da der Mikrozensus als eine rotierende Panelstichprobe angelegt ist, bei der jedes Jahr ein Viertel der Auswahlbezirke ausgetauscht wird und die Auswahlbezirke mit den darin wohnenden Haushalte und Personen dabei vier Jahre lang in der Befragung verbleiben, kann dieses Rotationsdesign für die Erstellung von Verlaufsdaten genutzt werden. Die Angaben zu den Merkmalsträgern, die vier Jahre in Folge am Mikrozensus teilgenommen haben, werden dabei zu einem Paneldatensatz verknüpft. Dies ist rechtlich erst seit dem Mikrozensusgesetz 1996 möglich. Das erste Mikrozensus-Panel umfasst daher die Jahre 1996 bis 1999, wobei es sich um eine faktisch anonymisierte 70%-Stichprobe der Auswahlbezirke handelt, die im genannten Zeitraum

[9] Hinweis auf weitere Mikrodatensätze (auch für andere Länder und aus internationalen Erhebungen) enthalten die Beiträge in der Serie *European Data Watch*, die seit 2000 regelmäßig in der Zeitschrift *Schmollers Jahrbuch / Journal of Applied Social Science Studies* erscheint. Bis auf die jeweils neuesten Folgen sind die Beiträge über die Homepage des Rats für Sozial- und Wirtschaftsdaten kostenlos verfügbar unter http://www.ratswd.de/publ/publikationen.php. Ferner ist auf die seit kurzer Zeit ebenfalls institutionalisiert zugänglichen umfangreichen Mikrodaten des ifo Instituts für Wirtschaftsforschung hinzuweisen; vgl. Abberger, Becker, Hofmann und Wohlrabe (2007).

[10] Detaillierte Informationen findet man auf der Homepage des German Microdata Lab unter der Adresse http://www.gesis.org/Dauerbeobachtung/GML/Daten/MZ/index.htm.

in den so genannten „Rotationsvierteln“ enthalten waren. Der Stichprobenumfang dieses faktisch anonymisierten Mikrozensus-Panels 1996 – 1999 beträgt rund 120.000 Personen (bzw. 55.000 Haushalte) pro Jahr mit rund 400 Variablen.[11] Dieser große Stichprobenumfang ermöglicht Verlaufsanalysen auch für spezifische und damit kleine Subpopulationen. Ein Scientific-Use-File des MZ-Panels ist seit kurzer Zeit in den FDZ des Statistischen Bundesamtes und der Statistischen Ämter der Länder verfügbar.

3.1.2 Stichprobe der Integrierten Erwerbsbiographien (IEBS)

Neben Befragungsdaten wie dem eben betrachteten Mikrozensus spielen für die wissenschaftliche Forschung und Politikberatung zunehmend Datensätze eine zentrale Rolle, die auf gleichsam nebenbei im Tagesgeschäft von Behörden anfallenden Informationen – so genannten prozessproduzierten Daten - beruhen. Ein Beispiel hierfür ist die Statistik der sozialversicherungspflichtig Beschäftigten: Beschäftgungsverhältnisse, die der Sozialversicherungspflicht unterliegen, führen zu Meldungen des Arbeitgebers an den für die Arbeitnehmer zuständigen Rentenversicherungsträger, aus denen Beginn und Ende eines Beschäftigungsverhältnisses sowie das gezahlte Entgeld und weitere Informationen zur Person (und dem ihn beschäftigenden Betrieb) hervorgehen. Weitere Beispiele aus dem Bereich von Beschäftigung und Arbeitslosigkeit sind Daten aus Meldungen über Zeiten von Leistungsbezügen wie Arbeitslosengeld oder über die Teilnahme an Maßnahmen der aktiven Arbeitsmarktpolitik wie Arbeitsbeschaffungsmaßnahmen. Diese prozessproduzierten Daten können so aufbereitet werden, dass sie in Form von Paneldatensätzen für einzelne Personen vorliegen. Damit bilden sie eine wertvolle Basis für wissenschaftliche Auswertungen.

Durch die Verknüpfung von verschiedenen bei der Bundesagentur für Arbeit vorhandenen Datenbeständen dieses Typs wurde vor kurzer Zeit ein Datensatz generiert, der weitaus detailliertere Analysen von Erwerbsverläufen zulässt, als dies zuvor möglich war – die Integrierten Erwerbsbiographien (IEB). Hierbei werden für einzelne Personen Angaben für Zeiträume (so genannte „Spells“) aus Datensätzen verknüpft, die Zeiten sozialversicherungspflichtiger Beschäftigung, Zeiten des Empfangs von Arbeitslosengeld, Arbeitslosenhilfe und Unterhaltsgeld, Zeiten der Teilnahme an Maßnahmen der Arbeitsagenturen und Zeiten der Arbeitssuche betreffen. Daraus wird dann die Erwerbsbiographie einer Person rekonstruiert. Diese Verlaufsdaten erlauben z. B. eine Analyse der Wirkungen von Maßnahmen der aktiven Arbeitsmarktpolitik; sie werden intensiv im Rahmen der Evaluation der Hartz-Gesetze genutzt.

Eine 2%-Stichprobe der Integrierten Erwerbsbiographien – die so genannte IEBS (Hummel et al. 2005; Jacobebbinghaus und Seth 2007) – enthält in der Version 1.0 tagesgenaue Angaben für 1.370.031 Personen in Form von 17.049.987 Originalspells aus dem Zeitraum 1990 bis 2004. Verfügbar sind Angaben zu soziodemographischen Merkmalen (Geschlecht, Geburtsjahr, Ausbildung, Staatsangehörigkeit, Wohn- und Arbeitsort, Regionaltyp), zur Beschäftigung (Tagesentgelt, berufliche Stellung, Wirtschaftszweig), zum Leistungsbezug (Arbeitslosengeld, Arbeitslosenhilfe, Unterhaltsgeld), zu Teilnahme an diversen Maßnahmen der aktiven Arbeitsmarktpolitik und zur Arbeitssuche (Status der Arbeitssuche, Erwerbsstatus vor Arbeitssuche,

[11] Vgl. zum Mikrozensus-Panel Basic, Marek und Rendtel (2005). Weitere Informationen findet man unter http://www.forschungsdatenzentrum.de/bestand/mikrozensus_panel/index.asp.

Beginn und Dauer der Arbeitslosigkeit). Die IEBS ist externen Wissenschaftlern seit kurzer Zeit zugänglich. Da sie nur schwach anonymisiert vorliegt, ist eine Nutzung allerdings lediglich im Rahmen von Gastaufenthalten am FDZ der BA möglich.[12]

3.2 Betriebs- und Unternehmensdaten

Während Einzeldaten für Personen und Haushalte aus Beständen der amtlichen Statistik und der Bundesagentur für Arbeit schon länger externen Wissenschaftlern zugänglich sind, ist dies bei Firmendaten erst seit einigen Jahren der Fall. Dies hängt auch damit zusammen, dass bei Personendaten eine faktische Anonymität deutlich leichter zu garantieren ist als bei Daten für Betriebe bzw. Unternehmen. Gerade in der jüngeren Vergangenheit hat sich hier aber viel bewegt – zentrale Beispiele sollen in den folgenden Unterabschnitten vorgestellt werden.

3.2.1 Paneldaten für Industriebetriebe aus der Amtlichen Statistik

Die statistischen Ämter erheben in Betrieben des Verarbeitenden Gewerbes sowie des Bergbaus und der Gewinnung von Steinen und Erden – hier kurz als Industriebetriebe bezeichnet – regelmäßig in einer Reihe von Befragungen Informationen zu Größen wie Beschäftigtenzahl, Inlands- und Auslandsaumsatz, Löhne und Gehälter sowie Investitionen. Deskriptive Auswertungen dieser Querschnittsdaten finden sich in den entsprechenden Publikationen der Ämter. Die so erhobenen Daten lassen sich jedoch darüber hinaus anhand der Betriebsnummern sowohl über die einzelnen Wellen einer Erhebung als auch über die unterschiedlichen Erhebungen hinweg zu Paneldaten mit einer Fülle von Informationen verknüpfen. Diese Art der Steigerung des Analysepotenzials vorhandener Datenbestände und die Bereitstellung entsprechender Mikrodaten für externe Wissenschaftler gibt es in einigen statistischen Landesämtern seit vielen Jahren, wobei das Niedersächsische Landesamt für Statistik eine Pionierrolle gespielt hat. Wagner (2000) stellt diese Datensätze vor und weist auf ausgewählte Publikationen hin, in denen sie genutzt wurden.

Bis vor kurzer Zeit waren diese Daten nur jeweils für einzelne Bundesländer verfügbar und externen Wissenschaftlern nur auf der Basis bilateraler Vereinbarungen mit einem Statistischen Amt zugänglich. Seit 2006 gibt es einen Datensatz, in dem für die Jahre 1995 bis 2004 sämtliche verfügbaren Betriebsdaten aus den Jahresergebnissen des „Monatsberichts für Betriebe des Verarbeitenden Gewerbes sowie des Bergbaus und der Gewinnung von Steinen und Erden“, aus der jährlichen „Investitionserhebung bei Unternehmen und Betrieben des Verarbeitenden Gewerbes sowie des Bergbaus und der Gewinnung von Steinen und Erden“ und aus der jährlichen „Erhebung für industrielle Kleinbetriebe im Bereich Verarbeitendes Gewerbe, Bergbau und Gewinnung von Steinen und Erden“ über alle 16 Bundesländer hinweg zu einem Panel verknüpft vorliegen (vgl. Konold 2007). Dieser Datensatz ist seit Ende 2006 in den FDZ des Statistischen Bundesamtes und der

[12] Hummel et al. (2005, S. 9) diskutieren Gemeinsamkeiten und Unterschiede zwischen der IEBS und zwei weiteren viel genutzten Personendateien aus Beständen der BA, dem IAB Beschäftigtenpanel (Koch und Meinken 2004) und der IAB Beschäftigtenstichprobe IABS (Bender, Haas und Klose 2000).

Statistischen Ämter der Länder für externe Wissenschaftler über die kontrollierte Datenfernverarbeitung und an Benutzerarbeitsplätzen zugänglich. Angesichts des hohen Analysepotenzials dieses Totalerhebungspanels deutscher Industriebetriebe[13] sind dessen Kosten in Höhe von 650 € (für zehn Jahre à 65 €) grundsätzlich relativ gering, können für junge Wissenschaftler jedoch eine substantielle Zugangsbarriere darstellen.

3.2.2 Kostenstrukturerhebungen und Kostenstrukturerhebungs-Panel

Das Statistische Bundesamt erhebt jährlich für eine hochrechnungsfähige Stichprobe von maximal 18.000 Unternehmen mit mehr als 19 Beschäftigten aus dem Bereich des Verarbeitenden Gewerbes sowie des Bergbaus und der Gewinnung von Steinen und Erden umfassende Informationen zur Belegschaft und ihrer Zusammensetzung, zu Umsatz und seinen Bestandteilen, zur Produktion, zum Einsatz von Rohstoffen und Vorprodukten sowie Energie, zu Lohn- und Sozialkosten und weiteren Kosten, zu Wertschöpfung und zum Forschungs- und Entwicklungsaufwand. Externe Wissenschaftler können die Erhebungen der Jahre seit 1992 (zurzeit bis zum Erhebungsjahr 2004) auf Gastwissenschaftler-Arbeitsplätzen in den FDZ des Statistischen Bundesamtes und der Statistischen Ämter der Länder nutzen (vgl. http://www.forschungsdatenzentrum.de/bestand/kse/index.asp). Fritsch u. a. (2004) informieren über das Potenzial dieser Daten für wissenschaftliche Analysen und berichten über bisherige Untersuchungen.
Die für die KSE verwendete Stichprobe ist ein rotierendes Panel, wobei Unternehmen mit 500 und mehr Beschäftigten total und dauerhaft einbezogen werden. Für den 4-Jahres-Zeitraum 1999 bis 2002 wurden die Wellen zu einem Paneldatensatz verknüpft, so dass hiermit Längsschnittstudien möglich sind. Dieser seit Ende 2006 für eine Nutzung durch externe Wissenschaftler in den FDZ des Statistischen Bundesamtes und der Statistischen Ämter der Länder verfügbare Datensatz ist für Forschungszwecke aus zwei Gründen besonders wichtig: Erstens werden in der KSE seit dem Berichtsjahr 1999 erstmals in Deutschland von der Amtlichen Statistik Informationen zur Forschungs- und Entwicklungstätigkeit von Unternehmen erhoben; dies schließt zumindest zum Teil eine schmerzliche Informationslücke über einen für die wirtschaftliche Dynamik zentralen Bereich unternehmerischen Handelns.[14] Zweitens wurden diesen Paneldaten Informationen aus dem Unternehmensregister des Statistischen Bundesamtes (Sturm und Tümmler 2006) zum Gründungsjahr zugespielt, womit eine Basis für empirische Studien in weiteren spannenden Feldern entstanden ist.

3.2.3 Das Betriebs-Historik-Panel

Die in den beiden vorigen Abschnitten betrachteten Längsschnittdatensätze aus Erhebungen der Amtlichen Statistik enthalten eine Vielzahl von Informationen über Betriebe und Unternehmen aus der Industrie. Dieser

[13] Der Begriff Totalerhebung trifft nur für die Jahre 1995 bis 2002 zu, denn die jährliche Kleinbetriebserhebung wurde mit dem Berichtsjahr 2003 eingestellt; ferner weist das Material in einigen Jahren bei einigen Bundesländern Lücken aufgrund von Datenverlusten auf. Hinzuweisen ist darauf, dass ab dem Berichtsjahr 2007 für den Monatsbericht nur noch Betriebe mit 50 und mehr tätigen Personen meldepflichtig sind; kleinere Industriebetriebe müssen nur noch jährlich tätige Personen, Lohn- und Gehaltssumme sowie Umsatz melden.
[14] Untersuchungen zur Forschungs- und Entwicklungstätigkeit von Unternehmen in Deutschland mussten bisher auf Daten aus nicht-amtlichen Stichproben ohne Auskunftszwang zurückgreifen; hier sind insbesondere das Mannheimer Innovationspanel (MIP) des Zentrums für Europäische Wirtschaftsforschung (ZEW) zu nennen (vgl. Janz et al. 2001) sowie die Daten der SV-Wissenschaftsstatistik (vgl. http://www.stifterverband.de/site/php/wirtschaft.php?SID=&seite=Statistik), die allerdings nur forschende Unternehmen betreffen.

Bereich der Wirtschaft ist ohne Zweifel von großer Bedeutung, aber es ist eben nur ein Teil der Gesamtwirtschaft, und sein Anteil schrumpft. Über den wichtigen und immer wichtiger werdenden Dienstleistungssektor liegen vergleichbare Paneldaten bisher nicht vor.[15]

Ein neuer Datensatz kann diese Lücke zumindest in einem wichtigen Teilbereich schließen. Seit Anfang 2007 steht im FDZ der BA für Wissenschaftler zur Nutzung auf den Gastwissenschaftler-Arbeitsplätzen und zur anschließenden Arbeit mit kontrolliertem Fernrechnen das Betriebs-Historik-Panel (BHP) zur Verfügung (vgl. Dundler, Stamm und Adler 2006 und http://fdz.iab.de/pageText.asp?PageID=122). Das BHP besteht zurzeit aus Querschnittsdatensätzen für die Jahre von 1993 bis 2003. Enthalten sind alle Betriebe in Deutschland, die zur Jahresmitte mindestens einen sozialversicherungspflichtig Beschäftigten gemeldet haben; ab 1999 werden zusätzlich auch Betriebe mit zumindest einem geringfügig Beschäftigten erfasst. Die Angaben in diesem Datensatz entstanden aus der Aggregation aller Personenmeldungen, die zu einer Betriebsnummer vorliegen. Die einzelnen Querschnittsdatensätze können zu einem Panel verbunden werden. Das BHP enthält neben Informationen zu Wirtschaftszweig und Standort des Betriebes detaillierte Angaben über die Beschäftigten nach verschiedenen Kriterien wie Geschlecht, Alter, Stellung im Beruf, Qualifikationsgrad und Staatsangehörigkeit. Darüber hinaus sind die Mittelwerte und Standardabweichungen des Alters der Beschäftigten eines Betriebes und des Tagesentgelts der Vollzeitbeschäftigten sowie der Beschäftigten insgesamt ausgewiesen. Mit diesem Datensatz sind erstmals umfassende Verlaufsuntersuchungen zur Arbeitsplatzdynamik in Betrieben auch außerhalb der Industrie möglich. Vorteile des BHP sind ferner die Verwendbarkeit für regionale Analysen sowie für die Identifikation von Gründungen und Schließungen; nachteilig für manche Fragestellungen ist die fehlende Möglichkeit einer Aggregation der Angaben zu Unternehmen.

Ein Mangel von Daten der amtlichen Statistik ist, dass eine „Georeferenzierung“, also das räumliche Zuordnen von Unternehmen oder Personen (bisher) nicht möglich ist. Dies macht diese Daten z.B. für finanzwirtschaftliche Analysen wenig attraktiv.

3.2.4 Unternehmensdaten der Deutschen Bundesbank

Die Deutsche Bundesbank erstellt seit 1964 eine Unternehmensbilanzstatistik, die umfangreichste Auswertung von Jahresabschlüssen in Deutschland ansässiger Unternehmen. Sie basiert auf den Bilanzen und Erfolgsrechnungen, die im Zusammenhang mit dem Rediskontgeschäft bei der Deutschen Bundesbank eingegangen sind. Nach dem Wegfall des Rediskontgeschäfts im Zuge der Europäischen Währungsunion hat die Bundesbank begonnen, eine neue Unternehmensbilanzstatistik zu erheben. Beide Statistiken, die „alte“ und die „neue“, auch „Jahresabschlussdatenpool“ genannt, sollen im Folgenden kurz beschrieben werden.

Die alte Unternehmensbilanzstatistik umfasst zwischen 22.000 und 70.000 jährliche Jahresabschlüsse west- und ostdeutscher Unternehmen. Aufgrund der Teilnahme dieser Unternehmen am Rediskontgeschäft handelt es bei diesen Daten nicht um eine originäre Erhebung zur Ertragslage und den Finanzierungsverhältnissen von deutschen Unternehmen, Vergleiche der Daten mit anderen Statistiken weisen – zumindest für die Jahre vor

[15] Vgl. http://www.forschungsdatenzentrum.de/datenangebot.asp#p_wirtschaft zum Datenangebot der amtlichen Statistik hierzu; unter http://www.forschungsdatenzentrum.de/bestand/dienstleistung/index.asp findet man Informationen zu den für Wissenschaftler verfügbaren Daten aus der seit 2000 durchgeführten Stichprobenerhebung im Dienstleistungssektor.

1997 – auf keine systematischen Verzerrungen hin. Nach 1997 ist, aufgrund des Wegfalls des Rediskontgeschäfts, die Repräsentativität eingeschränkt.

Unternehmen aus der Landwirtschaft und, was für den Empiriker sicherlich schwerer wiegt, dem Dienstleistungssektor, ausgenommen Handel und Verkehr, sind aufgrund geringer Fallzahlen nicht Teil der Daten. Allerdings handelt es sich bei der alten Unternehmensbilanzstatistik um einen Datensatz, der bereits seit einem Zeitpunkt erhoben wird, an dem das Dienstleistungsgewerbe noch eine untergeordnete Rolle spielte.

Eine weitere Konsequenz des Ursprungs der Daten aus dem Rediskontgeschäft ist, zumindest potentiell, die systematisch verzerrte Auswahl wirtschaftlich besonders „gesunder“ Unternehmen. Allerdings erfasst die alte Jahresabschlussdatenbank nicht nur die Angaben über den in der Tat meist wirtschaftlich gesunden Hauptzeichner, sondern auch der Nebenzeichner, deren wirtschaftliche Situation oft weniger gut ist als die des Hauptzeichners. Ein Vergleich der alten Jahresabschlussdatenbank mit der Insolvenzstatistik in Deutsche Bundesbank (1998) weist auf keine systematischen Verzerrungen in dieser Hinsicht hin.

In der alten Unternehmensbilanzstatistik sind kleine Unternehmen im Zeitraum nach 1997 unterrepräsentiert. Dieses Problem kann jedoch durch die Anwendung von Hochrechnungsverfahren behoben werden. Problematisch ist an dem Datensatz, dass Unternehmen nicht kontinuierlich am Wechselgeschäft teilgenommen haben, die Daten also sehr stark unbalanciert („unbalanced“) sind, was die Möglichkeiten des Arbeitens mit Paneldatenschätzern, insbesondere mit dynamischen Modellen, einschränkt.

Der Jahresabschlussdatenpool enthält, neben den (wenigen) Jahresabschlüssen, die der Bundesbank im Rahmen des Refinanzierungsgeschäftes weiterhin zugehen, auch Unternehmensbilanzdaten von Geschäftsbanken und Kreditversicherern. Diese stellen der Bundesbank Kundendaten zu Verfügung und erhalten im Gegenzug von der Bundesbank statistische Kennzahlen zur Unternehmensentwicklung. Dem Jahresabschlussdatenpool werden zudem Daten von privaten kommerziellen Anbietern zugefügt. Der Jahresabschlussdatenpool enthält zwischen 100.000 und 120.000 Jahresabschlüsse und reicht zur Zeit bis ins Jahr 2004.

Der Jahresabschlussdatenpool ist aufgrund methodischer Veränderungen und aufgrund von Veränderungen in Hinblick auf die Herkunft der Daten mit der alten Unternehmensbilanzstatistik nur eingeschränkt zu vergleichen. Der Jahresabschlussdatenpool ist wegen der Tatsache, dass das Refinanzierungsgeschäft nun eine untergeordnete Rolle spielt, weitaus weniger verzerrt in Hinblick auf Ertragsstärke und Unternehmensgröße als die alte Unternehmensbilanzstatistik in der Zeit nach 1997. Zudem deckt der Jahresabschlussdatenpool auch deutlich größere Teile des nichtfinanziellen Sektors ab.

Dem Jahresabschlussdatenpool und der alten Unternehmensbilanzstatistik können eingeschränkt externe Daten hinzugefügt werden. Der Jahresabschlussdatenpool enthält beispielsweise Postleitzahlen auf Zweisteller-Ebene. Grundsätzlich ist es möglich, den Bundesbank-Daten noch regionale Daten hinzuzufügen. Harhoff und Ramb (2001) erweitern den Jahresabschlussdatenpool für eine Analyse des Zusammenhangs zwischen Steuern und Investitionen Angaben um Unternehmenssteuerdaten auf der Regionalebene. Insofern sind die Bilanzdatenbestände der Bundesbank eine hervorragende Grundlage für finanzwirtschaftliche Analysen.

Der Jahresabschlussdatenpool enthält keinen Anlagenspiegel und daher keine Angaben zu Investitionen, was für Industrieökonomen schwer wiegt. Kapitalstöcke können jedoch über Buchwerte angenähert werden.

Die alte Unternehmensbilanzstatistik ist in Deutsche Bundesbank (1998) beschrieben, die neue Unternehmensbilanzstatistik in Deutsche Bundesbank (2005). Zugang zu beiden Unternehmensdatensätzen können Wissenschaftler über Gastforscherarbeitsplätze erhalten. Die Bundesbank verfügt z.Z. über zwölf solcher Arbeitsplätze.

Wissenschaftler, die mit den Bundesbankdaten arbeiten möchten, müssen einen Projektantrag schreiben, der von den zuständigen Stellen in der Bundesbank genehmigt werden muss. Nach Aussage der von uns befragten Wissenschaftler ist dies ein unkomplizierter und unbürokratischer Prozess. Aufgrund der Komplexität der Bundesbank-Datensätze empfiehlt es sich, mit Mitarbeitern der Bundesbank zusammen zu arbeiten.

Eine weitere von der Bundesbank aufbereitete Datenbank beinhaltet Angaben zu Direktinvestitionen deutscher Unternehmen im Ausland. Sie umfasst Bilanzdaten deutscher Unternehmen und die konsolidierten Bilanzdaten derer Töchter, allerdings keine vollständigen Jahresabschlüsse. Der Datensatz ist in Lipponer (2003) dokumentiert. Für diese Daten gelten dieselben Zugangsregelungen wie für die Unternehmensbilanzdaten. Ein Beispiel für eine auf Grundlage der Direktinvestitionsdaten entstandenen Arbeiten ist der Aufsatz von Arnold und Hussinger (2005).

Die Bundesbank verfügt noch über einen weiteren interessanten Datensatz, nämlich den über Unternehmensfusionen, der in Frey und Hussinger (2006) genutzt wird und der mit den Direktinvestitionsdaten zusammengespielt wurde.

3.3 Kombinierte Personen- und Firmendaten

Die bisher hier betrachteten Datensätze enthielten entweder Informationen über Personen bzw. Haushalte oder über Betriebe bzw. Unternehmen. Für eine umfassende empirische Analyse zahlreicher Fragestellungen sind aber Datensätze erforderlich, die Informationen zu beiden Bereichen – zu den Personen und zu den Betrieben, in denen diese Personen arbeiten – enthalten. Beispiele sind Studien zu Bestimmungsgründen von Unterschieden in der Entlohnung zwischen Personen, wo Merkmale der Person selbst (wie etwa Schulbildung und Berufserfahrung), gleichzeitig aber auch Merkmale des Betriebes (wie z.B. seine Größe oder die Bindung an einen Tarifvertrag) relevant sind. Solche so genannten Linked Employer-Employee-Daten (LEE-Daten) werden seit einigen Jahren verstärkt insbesondere in der Arbeitsmarktökonomik ausgewertet; sie gelten hierbei als Basis für manchen „Sprung nach vorn“ (vgl. Hamermesh 1999, 2007). Zwei solche für Deutschland verfügbare LEE-Datensätze sollen hier kurz vorgestellt werden.

Grundsätzlich ist hier anzumerken, dass die Rechtslage für personenbezogene Daten aufgrund des Volkszählungsurteils wesentlich schwerer zu verändern ist als für firmenbezogene Daten, für die die statistischen Ämter bereits Zusammenführungen vornehmen dürfen.

Anzumerken ist zudem, dass den Wirtschaftsstatistiken der statistischen Ämter öffentlich zugängliche Datenquellen zugespielt werden können. So haben die Forschungsdatenzentren z.B. das Unternehmensregister, die Körperschaftssteuerstatistik und die Gewerbesteuerstatistik mit der Amadeus-Datenbank – einer durch Private erfasste Datenbank standardisierter Jahresabschlüsse – zusammengeführt.

3.3.1 Gehalts – und Lohnstrukturerhebungen

Für die Gehalts- und Lohnstrukturerhebung wird im Allgemeinen alle vier Jahre eine Stichprobe der Betriebe des Verarbeitenden Gewerbes und ausgewählter Dienstleistungsbereiche befragt. Die Besonderheit dieser Erhebung besteht darin, dass gleichzeitig Informationen über den Betrieb und über in dem Betrieb tätige Personen erhoben werden. Informationen zur Person betreffen Geschlecht, Alter, Ausbildung, Steuerklasse und Kinderfreibeträge, ferner Angaben zur Tätigkeit (Berufsschlüssel der Sozialversicherung, Stellung im Beruf, Leistungsgruppe, Arbeitszeit, Dauer der Betriebszugehörigkeit) und zum Verdienst (Brutto, Netto, Zulagen für Schicht-/Nachtarbeit, Sonderzahlungen, Lohnsteuer, Sozialabgaben). Auf Betriebsebene werden zusätzlich Angaben darüber erhoben, ob die öffentliche Hand am Unternehmen beteiligt ist, ob der Betrieb in der Handwerksrolle eingetragen ist, welche Tarifverträge gelten sowie wie hoch die Anzahl der Beschäftigten, differenziert nach Geschlecht und nach Arbeitern und Angestellten, ist.

Die Gehalts- und Lohnstrukturerhebung ist damit ein originärer Linked Employer-Employee-Datensatz. Der Datensatz eignet sich z. B. gut zur Analyse von Lohnunterschieden, für die persönliche und betriebliche Einflussfaktoren bedeutsam sind. Hierfür wird er bereits intensiv genutzt (Stephan 2001). Die Daten für 1995 und 2001 (sowie für ausgewählte Bundesländer auch für 1990 bzw. 1992) sind in den FDZ des Statistischen Bundesamtes und der Statistischen Ämter der Länder verfügbar; für 2001 wurde erstmals ein faktisch anonymisierter SUF erstellt (vgl. Hafner, Lenz und Mischler (2007) sowie http://www.forschungsdatenzentrum.de/bestand/gls/index.asp. Mit Angaben für rund 22.000 Betriebe und über 846.000 Beschäftigte enthält dieser Datensatz das Potenzial für zahlreiche Analysen.

Die Gehalts- und Lohnstrukturerhebung ist ferner ein Beispiel dafür, dass Mikrodaten aus amtlichen Erhebungen mit Informationen aus öffentlich zugänglichen Quellen verknüpft werden können. Über den im Datensatz vorhandenen Tarifschlüssel wurden hierbei Informationen über die geltenden Tarifverträge und deren Ausgestaltung zugespielt, was das Analysepotenzial dieser Daten weiter erhöht hat.

3.3.2 Linked Employer-Employee-Daten aus dem IAB (LIAB)

LEE-Daten müssen nicht – wie im Fall der Gehalts- und Lohnstrukturerhebungen – originär als solche erhoben werden. Wenn es einen eindeutigen Identifikator gibt, der eine Zuordnung von Betrieben und in ihnen tätigen Personen erlaubt, dann lassen sich Mikrodaten für beide Arbeiten von Merkmalsträgern auch dann mit einander verknüpfen, wenn dies bei der Produktion der entsprechenden Datensätze nicht beabsichtigt war. Prominentes Beispiel für einen solchen generierten LEE-Datensatz ist in Deutschland der LIAB, für den die Betriebsdaten aus dem IAB-Betriebspanel (das nicht mit dem oben beschriebenen BHP identisch ist) mit den Personendaten zu in den dort befragten Betrieben beschäftigten Personen aus Datensätzen, die auf prozessproduzierten Daten der Arbeitsverwaltung und der Sozialversicherung beruhen, verknüpft wurden (Alda, Bender und Gartner 2005).

Für den LIAB werden zwei unterschiedliche Datenmodelle angeboten. Im LIAB-Querschnittmodell werden die Personendaten jährlich zu einem bestimmten Stichtag (30. Juni) mit den Daten des IAB-Betriebspanels verknüpft. Im LIAB-Längsschnittmodell sind die Personendaten nicht stichtagsbezogen, sondern umfassen einige Jahre an zeitraumbezogenen Personendaten.

Die LIAB-Daten werden externen Wissenschaftlern auszugsweise ausschließlich im Rahmen des Gastwissenschaftlermodells im Forschungsdatenzentrum der BA im IAB zur Verfügung gestellt. Nach einem Gastaufenthalt, der hauptsächlich dem Kennenlernen der Daten und dem Aufbau eines „individuellen" LIAB dient, ist eine kontrollierte Datenfernverarbeitung möglich.[16] Ausführliche Informationen über diesen sehr komplexen Datensatz sowie Testdaten findet man unter http://fdz.iab.de/pageText.asp?PageID=57.

3.4 Steuerdaten / Finanzwissenschaftliche Daten

Eine wichtige Quelle für finanzwirtschaftlich relevante Daten sind die Finanz- und Steuerstatistiken der Statistischen Ämter. Vier der wichtigsten Finanz- und Steuerstatistiken können über die Forschungsdatenzentren des Statistischen Bundesamtes und der Statistischen Ämter der Länder genutzt werden: die Gewerbesteuerstatistik, die Körperschaftsteuerstatistik, die Lohn- und Einkommensteuerstatistik sowie die Umsatzsteuerstatistik. Alle Datensätze sind auf der Homepage des FDZ (http://www.forschungsdatenzentrum.de/datenangebot.asp) unter den Hinweisen auf „Finanz- und Steuerstatistiken" eingehend beschrieben. Die Gewerbesteuerstatistik liegt für die Jahre 1995, 1998 und 2001 vor, die Körperschaftsteuerstatistik sowie die die Lohn- und Einkommensteuerstatistik für die Jahre 1992, 1995, 1998 und 2001 und die Umsatzsteuerstatistik für die Jahre 1998 und 2000 bis 2004. Die Lohn- und Einkommensteuerstatistik von 1998 steht zudem als Public- und Scientific-Use-File zur Verfügung, die Umsatzsteuerstatistik von 2000 als Scientific-Use-File (vgl. Zwick 2001; sowie Merz et al. 2006). Die Umsatzsteuerstatistik liegt auch als Panel für die Jahre 2000 bis 2004 vor.
Eine ganz neue und sehr viel versprechende Entwicklung auf dem Gebiet der steuerstatistischen Einzeldaten ist die Verknüpfung dieser Daten über die Zeit zu so genannten Taxpayer-Panels (vgl. Kriete-Dodds und Vorgrimler 2007), die weitergehende ökonometrische Analysen ermöglichen werden.

4. Zum Analysepotenzial der neu zugänglichen Mikrodaten

Zahlreiche über die neu errichtete informationelle Infrastruktur heute für externe Wissenschaftler mit geringem Aufwand an Geld und Zeit zugängliche Datensätze sind eine erstklassige Grundlage für qualitativ hochwertige empirische Forschungen. Sie können eine Basis sein sowohl für wissenschaftliche Publikationen, die hohen internationalen Standards genügen, als auch für Evaluationen wirtschaftspolitischer Maßnahmen, die die Anforderungen an eine methodisch dem Stand der Kunst entsprechende Vorgehensweise erfüllen. Zwei aktuelle Beispiele – je eines aus jedem genannten Bereich – sollen dies illustrieren:
Eine große Anzahl Studien mit Betriebsdaten aus vielen Ländern zeigt, dass exportierende Firmen im Durchschnitt höhere Löhne zahlen als nicht exportierende Betriebe aus der selben Industrie und Region; dieses Exporteur-Lohndifferential besteht auch bei Kontrolle für beobachtete und, wie Untersuchungen mit Paneldaten zeigen, unbeobachtete Eigenschaften der Betriebe. In der Literatur zu diesen durchschnittlichen

[16] Ein SUF für einen LIAB befindet sich in Vorbereitung; zu ersten Ergebnissen vgl. Drechsler et al. (2007)

Lohnunterschieden auf Betriebsebene wurde schon sehr früh die Frage aufgeworfen, ob es sich hierbei um ein Artefakt handelt, das darauf zurückzuführen ist, dass die Beschäftigten in den exportierenden Betrieben eine höhere (beobachtbare) Qualifikation und weitere (unbeobachtete) Produktivitätssteigernde Eigenschaften aufweisen. Allein mit Betriebsdaten lässt sich dies nicht überprüfen, hierfür sind LEE-Daten erforderlich, und zwar in Form von Paneldaten, damit für unbeobachtete zeitinvariante Eigenschaften von Betrieben und Personen kontrolliert werden kann. Schank, Schnabel und Wagner (2007) nutzen den oben beschriebenen LIAB für eine erste entsprechende empirische Studie. Sie zeigen, dass Arbeiter und Angestellte in exportierenden Betrieben auch bei Kontrolle für beobachtete und unbeobachtete Betriebs- und Personeneigenschaften mehr verdienen als in nicht exportierenden Betrieben. Dieser überzeugende Beleg für das Vorliegen eines positiven Exporteur-Lohndifferentials wäre den Verfassern ohne die Möglichkeit eines Zugriffs auf die LIAB-Daten im FDZ der BA im IAB nicht möglich gewesen.

Ein Themenbereich, bei dem das Analysepotenzial der neu zugänglichen Mikrodaten besonders deutlich wird, ist die Evaluation der Wirkungen arbeitsmarktpolitischer Maßnahmen. Die so genannten „Hartz-Gesetze“ zählen hierbei sicherlich zu den größten sozialpolitischen Reformwerken der letzten Jahrzehnte in Deutschland. Der Gesetzgeber hat vorgesehen, die Hartz-Reformen ausführlich evaluieren zu lassen. Das Bundesministerium für Arbeit und Soziales hat hierzu Forschungsaufträge an verschiedenen Konsortien vergeben, die in den Jahren 2004 bis 2006 bearbeitet wurden (Hartz I-III) bzw. bis 2008/2009 bearbeitet werden (Hartz IV).

Die Prozessdaten der Bundesagentur für Arbeit, die durch das Institut für Arbeitsmarkt- und Berufsforschung (IAB) zu Forschungsdaten aufbereitet werden, stellen eine ideale Grundlage für solche Evaluationsvorhaben dar. Um eine optimale Versorgung der Forschungsprojekte gewährleisten zu können, wurde daher das IAB durch das BMAS mit der Datenaufbereitung und –lieferung beauftragt. Insbesondere die speziell aufbereiteten Datenauszüge auf Basis des Verfahrens der „Integrierten Erwerbsbiographien“ (IEB) waren hierbei von zentraler Bedeutung. Die IEB-Daten umfassen neben Charakteristiken von Erwerbsfähigen auch eine Maßnahmenträgerdatei, die Informationen über Art und Zeitpunkt von arbeitspolitischen Maßnahmen enthält.

Da es sich bei den bereitgestellten Daten um Sozialdaten handelt, müssen bei der Arbeit mit den Daten durch die Forschungsinstitute strenge Datenschutzauflagen eingehalten und die Rohdaten nach Abarbeitung der Forschungsaufträge innerhalb bestimmter Fristen gelöscht werden. Aufgrund der zwingend geforderten Löschung der für die Evaluation verwendeten Daten bei den evaluierenden Forschungsinstituten, wurde das IAB durch das BMAS auch mit der Archivierung beauftragt. Dadurch wird gesichert, dass die Ausgangsdaten für Replikationen oder Revisionen wieder zur Verfügung gestellt werden können. Zudem können die Forschungsprojekte auch die von Ihnen generierten Analysefiles und Ergebnisse durch das IAB für zehn Jahre archivieren lassen.

Auch für die verschiedenen Projekte im Rahmen der Hartz-IV-Evaluation wurde von Seiten des IAB wieder Daten aufbereitet und den externen Forschern zur Verfügung gestellt. Um die Daten auch nach dem Ende der Forschungsaufträge weiter für Externe nutzbar zu machen, arbeitet das IAB gegenwärtig an der Erstellung eines Scientific Use Files.

Trotz vieler technischer Probleme und Schwierigkeiten haben die Arbeiten an der Evaluation der Hartz-Gesetze deutlich gemacht, dass durch den Zugriff externer Wissenschaftler auf administrative Mikrodaten die Möglichkeiten für belastbare Wirkungsanalysen arbeitsmarktpolitischer Maßnahmen erheblich gesteigert

werden konnten (vgl. z. B. die Ergebnisse der Hartz-Evaluationen in Brinkmann, Hujer und Koch 2006). Dies verdeutlicht das hohe Analysepotential dieser neu zugänglichen Mikrodaten für wirtschaftspolitisch unmittelbar relevante Fragestellungen.

5. Ein Blick nach Norden …

Die Möglichkeiten des Zugangs zu geheimen Mikrodaten aus amtlichen Statistiken haben sich in Deutschland wie gesehen in den vergangenen Jahren deutlich verbessert. Für eine Einschätzung dessen was erreicht wurde und was vielleicht noch erreicht werden kann ist es hilfreich, einen Blick über den Zaun zu unseren skandinavischen Nachbarn zu werfen. Beispielhaft soll dies hier für Dänemark geschehen.

Dreh- und Angelpunkt aller dänischen Daten ist die eineindeutige Zuordnung von Personen und Unternehmen durch Identifikationsnummern. Allen in Dänemark lebenden Personen im Alter von 18 Jahre oder älter wurde 1980 eine Personennummer zugeordnet, die so genannte „CPR"-Nummer. Die bereits zuvor bestehenden Steuernummern für Unternehmen wurden im Oktober 1999 um so genannte „CVR"-Nummern ergänzt, auf denen die Unternehmensregister aufgebaut sind. Die Zuordnung dieser Identifikationsnummern erlaubt es, verschiedene Statistiken problemlos zusammenzuführen.

Die zentrale Datei für personenbezogene Mikrodaten ist die „IDA"-Datenbank, die zahlreiche Angaben über Personen in einem einzigen Datensatz zusammenfasst. So umfassen die IDA-Daten Angaben über das Einkommen, die Ausbildung, die berufliche Stellung, eigenes Alter, Anzahl und Alter der Kinder, Arbeitsmarkterfahrung, Anzahl der Jahre im gegenwärtigen Beruf („tenure") oder die Teilnahme an arbeitsmarktpolitischen Maßnahmen. Dieselben Angaben für Eltern und Partner können den IDA-Kerndaten problemlos hinzugeführt werden.

Die in der IDA-Datenbank enthaltenen Angaben stammen in erster Linie aus den Lohn- und Einkommenssteuererklärungen sowie Statistiken über Ausbildung und Berufsstand. Über die CPR-Nummer können die IDA-Daten um weitere öffentlich zugängliche Statistiken erweitert werden. Grundsätzlich ist es möglich, den IDA-Daten alle in Dänemark vorhandenen Statistiken zuzuführen. Zu den eher nahe liegenden Statistiken zählen hierbei Angaben zu Gesundheitszustand und Medizingebrauch, zu den etwas exotischeren Statistiken zählen die Intelligenzquotienten, die im Rahmen der Musterung erhoben werden. Eine Übersicht über die zugänglichen Datenbestände gibt die Internetseite von Statistics Denmark, http://www.dst.dk/TilSalg/Forskningsservice.aspx. Diese Internetseite ist auf Dänisch geschrieben und in keine dem internationalen Nutzer leichter zugängliche Sprache übersetzt. Die IDA-Daten sind für die Jahre 1980 bis 2002 erhältlich.

Das Pendant zu den IDA-Daten aus unternehmensbezogener Sicht sind die „FIDA" Daten. Dieser Datensatz ist im Kern eine Kopplung der IDA-Daten mit Daten aus der Unternehmensstatistik, insbesondere der Bilanzdaten dänischer Unternehmen. Die FIDA-Daten enthalten, neben der Verbindung zu den IDA-Daten, Angaben zu Aktiva und Passiva, Umsatz, Gewinn vor und nach Steuern sowie Anzahl der Mitarbeiter. Einige dieser Angaben sind auch auf Betriebsebene erhältlich. Aufgrund der Kopplung mit den personenbezogenen IDA-Daten kann das Humankapital eines Unternehmens leicht nachvollzogen werden; die Erstellung eines LEE

Datensatzes ist also unproblematisch. Die FIDA-Daten sind für die Jahre 1996 bis 2002 erhältlich. Die oben beschriebene Umstellung von Steuernummern auf CVR Nummern hat, neben einigen anderen Veränderungen in der Datenerfassung, zu Folge, dass in den Daten ein Strukturbruch entstanden ist. Grundsätzlich können die „alten“ (1996 bis 1999) und „neuen“ Datensätze (2000 bis 2002) verknüpft werden, allerdings können verschiedene Steuernummern unter einer einzigen CVR-Nummer zusammengefasst sein, was die exakte Definition von Unternehmen erschwert.

Die Verfügbarkeit dänischer Daten geht sogar noch weiter. So kann man zum Beispiel den Daten von Statistics Denmark Daten aus Quellen außerhalb der Bundesstatistik zuführen. Kaiser et al. (2007) spielen z.B. Angaben über Patentanmeldungen dänischer Unternehmen beim Europäischen Patentamt den FIDA-Daten hinzu. Dazu haben die Autoren zunächst dem Bestand an Patentanmeldungen von dänischen Unternehmen CVR-Nummern hinzugefügt und diesen Datenbestand dann mit den FIDA-Daten abgeglichen. In einem weiteren Schritt haben die Autoren dann die gematchten Patent- und FIDA-Daten mit den Mitarbeiterdaten aus der IDA-Datei zusammengeführt.

Ebenso können den dänischen Daten Informationen aus Umfragen hinzugefügt werden, sofern dies wissenschaftlichen Zwecken dient. Umgekehrt können die dänischen Registerdaten dazu verwendet werden, Stichproben für Befragungen zu ziehen. Andersen et al. (im Erscheinen) haben beispielsweise Befragungspersonen aus den IDA-Daten bestimmt.

Zugang zu den dänischen Daten haben grundsätzlich Wissenschaftler und Mitarbeiter staatlicher Institutionen, z.B. des Wirtschaftsministeriums. Datennutzer müssen eine Projektskizze vorlegen, die nicht nur das Projekt beschreibt, sondern auch erklärt, warum welche Variablen im Datensatz enthalten sein müssen. Statistics Denmark legt bei allen Projekten das „need to know“ Prinzip an, d.h. die Benutzung jeder einzelnen Variable muss begründet werden. Dies klingt auf den ersten Blick bürokratisch, ist in der Praxis, aufgrund der hohen wissenschaftlichen Qualifikation der Mitarbeiter von Statistics Denmark, aber kein Problem. Häufig ist eine kurze, intuitive Begründung hinreichend. Beispielsweise fällt es nicht schwer, die Mitarbeiter von Statistics Denmark von der Wichtigkeit von Instrumenten zur Identifikation von kausalen Effekten zu überzeugen.

Um mit den Daten von Statistics Denmark arbeiten zu können, mussten Nutzer zunächst in der Zentrale in Kopenhagen arbeiten. Später wurde dann eine Außenstelle in Århus eingerichtet, von der aus ebenfalls mit den Daten gearbeitet werden kann. Seit 2002 dürfen in Dänemark tätige und wohnhafte Wissenschaftler über eine Fernverbindung zu den Großrechnern auf die Daten zugreifen. Diese Fernverbindung ist durch fünf Kennwörter, darunter ein sich 20-sekündlich veränderndes, gesichert. Abgesehen vom völligen Verlust der wissenschaftlichen Reputation sind die Strafen für den Datenmissbrauch drakonisch und reichen von der Verweigerung des Datenzugangs auf Lebenszeit bis hin zu Freiheitsstrafen. Vor diesem Hintergrund ist es sicherlich wenig überraschend, dass in Dänemark kein einziger Fall von Datenmissbrauch bekannt ist.

Angeforderte Daten werden i. d. R. spätestens nach zwei Monaten bereitgestellt. Die Bereitstellung der Daten erfolgt nicht kostenlos, was in der Vergangenheit zu scharfen Diskussionen geführt hat, da die Daten durch Steuergelder bereits bezahlt worden sind und faktisch vorliegen. Statistics Denmark will die Arbeitskosten für einzelne Mitarbeiter, die die Datensätze bereitstellen, gedeckt wissen. Datennutzer müssen also für die Kosten der Bereitstellung aufkommen, die wiederum von der Anzahl der benutzten Variablen abhängen. Berücksichtigt werden müssen noch Kosten für Speicherplatz, der sowohl mit der Anzahl der Variablen als auch der Anzahl

der Beobachtungen zunimmt. Um ein Preisbeispiel zu geben, die Verwendung der Population dänischer Einwohner von 1980 bis 2002 für z.B. eine Lohnregression kostet in der Bereitstellung unter 15.000 €. Damit liegen die Kosten der Nutzung der dänischen Daten zwischen denen für die Verwendung von Daten kommerzieller Anbieter und denen, die das Statistische Bundesamt und die Statistische Landesämter in Deutschland verlangen.

Eines der Kernprobleme bei der Arbeit mit den dänischen Registerdaten ist, dass Schätzergebnisse von Dritten zwar grundsätzlich repliziert werden können – die Projektautoren müssen diese Dritten lediglich als Projektdatennutzer bei Statistics Denmark registrieren lassen. Aufgrund der Notwendigkeit, dass Nutzer sich auf dänischem Boden aufhalten müssen, sind der Replizierbarkeit jedoch deutliche Grenzen gesetzt. Dies ist insbesondere vor dem Hintergrund problematisch, dass immer mehr wissenschaftliche Zeitschriften, wie z.B. das Journal of Human Ressources, das Journal of Business Economics and Statistics oder das Journal of Applied Econometrics, die Bereitstellung der verwendeten Daten und der verwendeten Softwareprogramme verlangen. Aufgrund der Tatsache, dass in allen diesen Zeitschriften in jüngerer Zeit Arbeiten erschienen sind, die auf den dänischen Registerdaten beruhen, besteht jedoch bislang noch kein akutes Problem.

Ein faktisches Problem ist jedoch, dass die Rechenzeiten auf den Servern von Statistics Denmark z. T. sehr lang sind. Dies liegt zum einen an der Geschwindigkeit der Fernverbindung, zum anderen aber auch an der permanenten Überlastung der Rechner. Statistics Denmark kauft zwar regelmäßig neue Server und erweitert den Speicherplatz, allerdings ist die Nachfrage nach den Registerdaten in den vergangenen Jahren sprunghaft gestiegen. Diese Schwierigkeit kann jedoch durch den Kauf eigener Server, die in den Räumlichkeiten von Statistics Denmark platziert werden, umgangen werden. Das Centre for Applied Microeconometrics an der Universität Kopenhagen hat diesen Weg beispielsweise beschritten.

Ein weiteres Problem aus nicht-dänischer Sicht ist, dass Datendokumentationen lediglich in Dänisch vorliegen. Statistics Denmark arbeitet seit einigen Jahren an englischsprachigen Datenbeschreibungen, Ergebnisse hierzu liegen bislang aber nicht vor.

6. … und ein Blick in die Zukunft

Der Zugang zu vertraulichen Mikrodaten aus amtlichen Erhebungen und zu Daten, die auf prozessproduzierten Daten der Arbeitsverwaltung und der Sozialversicherung zurückgehen, ist heute für externe Wissenschaftler sehr viel einfacher als noch vor wenigen Jahren. Wer vor zehn Jahren z. B. behauptet hätte, dass all die Forderungen nach Nutzungsmöglichkeiten der Datenschätze der Bundesanstalt für Arbeit (so hieß die heutige Bundesagentur für Arbeit ja damals), die für eine methodisch dem Stand der Wissenschaften gerecht werdende Evaluation vom Maßnahmen aktiver Arbeitsmarktpolitik erforderlich sind, in absehbarer Zeit einmal erfüllt werden, wäre als Phantast bezeichnet worden. Heute sind solche Möglichkeiten Realität.

Es ist viel erreicht: Nicht nur sind zentrale Datenbestände in Form von Querschnitten zugänglich, sie wurden auch vielfach über die einzelnen Erhebungswellen zu Paneldaten verknüpft aufbereitet und Daten aus allen Bundesländern wurden zusammengeführt. Die Möglichkeiten, die eindeutige und in verschiedenen Datensätzen gleichzeitig vorhandene Identifikatoren bieten, wurden für eine Verknüpfung von Datenbeständen genutzt, um

so kombinierte Datensätze mit einem deutlich gesteigerten Informationsgehalt zu generieren (vgl. die oben vorgestellten Paneldaten für Industriebetriebe und die Integrierten Erwerbsbiographien).

Auch wenn viel erreicht ist, so bleiben doch noch Wünsche offen. Diese lassen sich in zwei große Bereiche unterteilen – die verfügbaren Datensätze einerseits und die Zugangswege zu ihnen andererseits:

Bisher können in Deutschland Daten zu Betrieben aus unterschiedlichen Quellen nur dann verknüpft werden, wenn sie auf einer einheitlichen gesetzlichen Grundlage erfasst wurden. Ohne hier aus Platzgründen auf die juristischen Details eingehen zu können bleibt festzuhalten, dass z. B. eine Verknüpfung von Sozialdaten aus den Beständen der BA mit Daten aus Erhebungen des Statistischen Bundesamtes oder der Statistischen Ämter der Länder nicht möglich ist (es sei denn, jeder Merkmalsträger stimmt dem schriftlich zu – aber wer will auch nur versuchen, die Einwilligung von Millionen Menschen oder Betrieben einzuholen?). Ein zentrales Ziel für den weiteren Ausbau der informationellen Infrastruktur muss sein, diesen Zustand zu ändern. Die Vision eines umfassenden Datenbestandes in Form von LEE-Panels für alle wirtschaftsaktiven Einheiten und alle in ihnen tätigen Personen mit Informationen aus allen Erhebungen und Beständen der amtlichen Statistik, der Arbeitsverwaltung, der Sozialversicherung und aus den Beständen anderer (öffentlicher) Institutionen wie z.B. der Bundesbank sollte Schritt für Schritt Realität werden. Das Beispiel Dänemark aus dem vorigen Abschnitt belegt überzeugend, was hier möglich ist. Die zentrale technische Voraussetzung hierfür ist die fortlaufende Entwicklung des Unternehmensregister-Systems (Sturm und Tümmler 2006), das eine Zusammenführung aller Informationen über wirtschaftsaktive Einheiten (wie Betriebe oder Unternehmen) aus unterschiedlichen Quellen ermöglicht, und zwar auf einem Stand, der Pilotprojekte umsetzbar erscheinen lässt. Erste Ansätze hierfür sollen ab 2007 in dem Projekt *KombiFiD* – Kombinierte Firmendaten für Deutschland – realisiert werden (vgl. Bender, Wagner und Zwick (2007)). An der für eine Umsetzung im großen Stil erforderlichen Reform der rechtlichen Grundlagen bleibt zu arbeiten.

Weitere Erleichterungen beim Datenzugang für externe Wissenschaftler sind in zwei Bereichen zu erwarten. Einerseits sind Methoden zur faktischen Anonymisierung von Paneldaten für Betriebe und Unternehmen (und damit auch für LEE-Panel-Daten) Thema laufender Forschungen; dies wird zu Verfahren führen, die die Weitergabe vertraulicher Mikrodaten in Form von SUFs auch für Paneldaten gestatten. Damit wird die Arbeit am eigenen PC die zeitaufwendige kontrollierte Datenfernverarbeitung und die noch sehr viel teurere Präsenz auf Benutzerarbeitsplätzen in den FDZ in erheblichem Maße ersetzen können. Andererseits werden technische und juristische Voraussetzungen für die Schaffung von abgeschotteten Benutzerarbeitsplätzen in Einrichtungen mit der Aufgabe unabhängiger wissenschaftlicher Forschung (wie etwa Universitäten) geprüft, von denen aus ein direkter Zugriff auf nur On-Site anonymisierte vertrauliche Daten in gleicher Weise möglich ist wie dies auf den Benutzerarbeitsplätzen in den FDZ der Fall ist.

Weitere große Fortschritte beim Zugang zu vertraulichen Mikrodaten in Deutschland in den kommenden Jahren zu erwarten ist damit realistisch.[17] Auch dann bleiben aber noch Wünsche offen: Die Datensätze sollten in der Zukunft über Ländergrenzen hinweg verknüpfbar sein, um Analysen von Verlagerungsprozessen und

[17] Diese Aussage steht für den zentralen Teil der Daten aus dem Statistischen Bundesamt und den Statistischen Ämtern der Länder unter einem Finanzierungsvorbehalt, denn die längerfristige Finanzierung dieser FDZ ist nicht gesichert. Alle Wissenschaftler, die eine gut funktionierende informationelle Infrastruktur für unverzichtbar halten, sind daher aufgefordert, sich an der hier notwendigen Lobbyarbeit zu beteiligen – bei einem erforderlichen Finanzierungsvolumen von ca. 1.8 Millionen Euro pro Jahr muss eine dauerhafte Implementierung dieser FDZ möglich sein!

Vorgängen innerhalb multinationaler Unternehmen sowie von Migrationen zu ermöglichen – und sie sollten (geschützt durch ein Forschungsdatengeheimnis mit harten Strafen bei Missbrauch wie im oben dargestellten Beispiel Dänemark) für alle Wissenschaftler in schwach anonymisierter Form an unseren Arbeitsplätzen zur Verfügung stehen. Heute mag dies als Vision gesehen werden, in zehn Jahren ist es hoffentlich Realität wie heute die FDZ. Und wer jetzt sagt, das klappt doch niemals, dem antworten wir mit dem großen Bert Brecht:

Wer noch lebt, sage nicht: niemals!
Das Sichere ist nicht sicher.
So, wie es ist, bleibt es nicht. (…).
Und aus Niemals wird: Heute noch!
[Bertolt Brecht, Lob der Dialektik]

Literaturverzeichnis

Abberger, K., S. O. Becker, B. Hofmann und K. Wohlrabe (2007), Mikrodaten im ifo Institut für Wirtschaftsforschung – Bestand, Verwendung und Zugang, ifo Working Paper No. 44, März.

Alda, H., S. Bender und H. Gartner (2005), The linked employer-employee dataset created from the IAB establishment panel and the process-produced data of the IAB (LIAB), *Schmollers Jahrbuch / Journal of Applied Social Science Studies* 125, 327-336.

Andersen, S., G.W. Harrison, M.I. Lau und E.E. Rutström (im Erscheinen), Lost in State Space: Are Preferences Stable?, erscheint in: *Econometrica*.

Arnold, J.M. und K. Hussinger (2005), Exports versus FDI in German Manufacturing: Firm Performance and Participation in International Markets, ZEW Discussion Paper No. 05-73, Mannheim.

Basic, E., I. Marek und U. Rendtel (2005), The German Microcensus as a Tool for Longitudinal Data Analysis: An Evaluation Using SOEP Data, *Schmollers Jahrbuch / Journal of Applied Social Science Studies* 125, 167-181.

Bender, S., A. Haas und C. Klose (2000), The IAB Employment Subsample 1975-1995, *Schmollers Jahrbuch / Journal of Applied Social Science Studies* 120, 649-662.

Bender, S., J. Wagner und M. Zwick (2007), Machbarkeitsstudie: Zusammenführung von Mikrodaten der Statistischen Ämter, des Instituts für Arbeitsmarkt- und Berufsforschung und weiterer Datenproduzenten – Kombinierte Firmendaten für Deutschland (KombiFiD), mimeo.

Brinkmann, C., R. Hujer und S. Koch (Hrsg.) (2006), Evaluation aktiver Arbeitsmarktpolitik in Deutschland, Themenheft der Zeitschrift für ArbeitsmarktForschung, 39 (Heft 3 und 4), 317-617.

Deutsche Bundesbank (1998), Methodische Grundlagen der Unternehmensbilanzstatistik der Deutschen Bundesbank, *Monatsbericht* Oktober 1998, 51-57.

Deutsche Bundesbank (2005), Ertragslage und Finanzierungsverhältnisse deutscher Unternehmen – eine Untersuchung auf neuer Datenbasis, *Monatsbericht* Oktober 2005, 33-71.

Drechsler, J., A. Dundler, S. Bender, S. Rässler und T. Zwick (2007), A new approach for disclosure control in the IAB Establishment Panel – Multiple imputation for a better data access, IAB Discussion Paper Nr. 11/2007, Nürnberg.

Dundler, A., M. Stamm und S. Adler (2006), *Das Betriebs-Historik-Panel – BHP 1.0.* FDZ-Datenreport Nr. 3/2006, Bundesagentur für Arbeit, Nürnberg.

Harhoff, D. und F., Ramb (2001), Investment and Taxation in Germany - Evidence from Firm-Level Panel Data, in: Deutsche Bundesbank (Hrsg.), *Investing Today for the World of Tomorrow*, Springer, Heidelberg.

Frey, R. und K. Hussinger (2006), The Role of Technology in M&As: A Firm Level Comparison of Cross-Border and Domestic Deals, ZEW Discussion Paper No. 06-069, Mannheim.

Fritsch, M., B. Görtzig, O. Hennchen und A. Stephan (2004), Cost Structure Surveys for Germany, *Schmollers Jahrbuch / Journal of Applied Social Science Studies* 124, 557-566.

Gottschalk, Sandra (2005), Unternehmensdaten zwischen Datenschutz und Analysepotenzial, Zentrum für Europäische Wirtschaftsforschung Wirtschaftsanalysen, Bd. 76, Baden-Baden.

Hafner, H.-P., R. Lenz und F. Mischler (2007), Einzeldaten der Gehalts- und Lohnstrukturerhebung 2001 als Scientific-Use-File, *Wirtschaft und Statistik*, Heft 2, 144-149.

Hamermesh, D. (1999), LEEping Into The Future of Labor Economics: The Research Potential of Linking Employer and Employee Data, *Labour Economics* 6, 25-41.

Hamermesh, D. (2007), Fun With Matched Firm-Employee Data: Progress and Road Maps, IZA Discussion Paper 2580, January.

Hummel, E. et al. (2005), *Stichprobe der Integrierten Erwerbsbiographien – IEBS 1.0*. FDZ-Datenreport Nr. 6/2005, Bundesagentur für Arbeit, Nürnberg.

Jacobebbinghaus, P. und S. Seth (2007), The German Integrated Employment Biographies Sample IEBS, *Schmollers Jahrbuch / Journal of Applied Social Science Studies* 127 (forthcoming).

Janz, N., G. Ebling, S. Gottschalk und H. Niggemann (2001), The Mannheim Innovation Panels (MIP and MIP-S) of the Centre for European Economic Research (ZEW), *Schmollers Jahrbuch / Journal of Applied Social Science Studies* 121, 123-129.

Kaiser, U., H.C. Kongsted und T. Rønde (2007), Human Capital, Patenting and Knowledge Spillovers, Centre for Economic and Business Research at Copenhagen Business School Mimeo, Kopenhagen.

Koch, I. und H. Meinken (2004), The Employment Panel of the German Federal Employment Agency. *Schmollers Jahrbuch / Journal of Applied Social Science Studies* 124, 315-325.

Kölling, A. (2000), The IAB-Establishment Panel, *Schmollers Jahrbuch / Journal of Applied Social Science Studies* 120, 291-300.

Kohlmann, A. (2005), The Research Data Centre of the Federal Employment Service in the Institute for Employment Research, *Schmollers Jahrbuch / Journal of Applied Social Science Studies* 125, 437-447.

Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (2001), *Wege zu einer besseren informationellen Infrastruktur.* Nomos Verlagsgesellschaft, Baden-Baden.

Konold, M. (2007), New possibilities for economic research through integration of establishment-level panel data of German official statistics, *Schmollers Jahrbuch / Journal of Applied Social Science Studies* 127 (in press).

Kriete-Dodds, S. und D. Vorgrimler (2007), Das Taxpayer-Panel der jährlichen Einkommenssteuerstatistik, *Wirtschaft und Statistik*, Heft 1, 77-85.

Lenz, R., Vorgrimler, D. und Rosemann, M. (2005), Ein Scientific-Use-File der Kostenstrukturerhebung im Verarbeitenden Gewerbe, *Wirtschaft und Statistik* 2/2005, 91-96.

Lenz, R., M. Rosemann, D. Vorgrimler und R. Sturm (2006), Anonymizing business micro data – results of a German project, *Schmollers Jahrbuch / Journal of Applied Social Science Studies* 126 (in press).

Lipponer, A. (2003), Deutsche Bundesbank's FDI micro database, *Schmollers Jahrbuch / Journal of Applied Social Science Studies* 123, 593-600.

Lüttinger, P., B. Schimpl-Neimanns, H. Wirth und G. Papastefanou (2004), The German Microdata Lab at ZUMA: Services Provided to the Scientific Community, *Schmollers Jahrbuch / Journal of Applied Social Science Studies* 124, 455-467.

Merz, J., D. Vorgrimler und M. Zwick (2006), De Facto Anonymised Microdata File on Income Tax Statistics 1998, *Schmollers Jahrbuch / Journal of Applied Social Science Studies* 126, 313-327.

Pohlmeier, W., G.Ronning und J. Wagner (Hrsg.) (2005), *Econometrics of Anonymized Micro Data*. Lucius & Lucius, Stuttgart.

Rat für Sozial- und WirtschaftsDaten (Hrsg.) (2007), Eine moderne Dateninfrastruktur für eine exzellente Forschung und Politikberatung, Berlin.

Rehfeld, U. G. und T. Mika (2006), The Research Data Centre of the German Statutory Pension Insurance (FDZ-RV), *Schmollers Jahrbuch / Journal of Applied Social Science Studies* 126, 121-127.

Ronning, G. u.a. (2005), *Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten*. Statistisches Bundesamt, Wiesbaden.

Schank, T., C. Schnabel und J. Wagner (2007), Do exporters really pay higher wages? First evidence from German linked employer-employee data, *Journal of International Economics* 72, 52-74.

Scheffler, M. (2005), Ein Scientific-Use-File der Einzelhandelsstatistik 1999, *Wirtschaft und Statistik* 3/2005, 197-200.

Schwarz, N. (2001), The German Microzensus, *Schmollers Jahrbuch / Journal of Applied Social Science Studies* 121, 649-654.

Stephan, G. (2001), The Lower Saxonian Salary and Wage Structure Survey – Linked Employer-Employee Data from Official Statistics, *Schmollers Jahrbuch/ Journal of Applied Social Science Studies* 121, 267-274.

Sturm, R. und Lenz, R. (2005), Erste Scientific-Use-Files aus den Wirtschaftsstatistiken, Proceedings der Nutzerkonferenz am Institut für Weltwirtschaft der Universität Kiel, 19. Mai 2005, 191-207.

Sturm, R. und T. Tümmler (2006), Das statistische Unternehmensregister – Entwicklungsstand und Perspektiven, *Wirtschaft und Statistik*, Heft 10, 1021-1036.

Terwey, M. (2000), ALLBUS: A German General Social Survey, *Schmollers Jahrbuch/ Journal of Applied Social Science Studies* 120, 151-158.

Vorgrimler, D., Dittrich, S., Lenz, R. und Rosemann, M (2005), Ein Scientific-Use-File der Umsatzsteuerstatistik, *Wirtschaft und Statistik* 3/2005, 201-210.

Wagner, G. G., J. R. Frick und J. Schupp (2007), The German Socio-Economic Panel Study (SOEP) – Evolution, Scope and Enhancements, *Schmollers Jahrbuch* / *Journal of Applied Social Science Studies* 127 (im Druck).

Wagner, J. (2000), Firm Panel Data from German Official Statistics, *Schmollers Jahrbuch* / *Journal of Applied Social Science Studies* 120, 143-150.

Zühlke, S. und H. Christians (2006), Datenangebot und Datenzugang im Forschungsdatenzentrum der Statistischen Landesämter, Statistische Analysen und Studien NRW (29), 3-11.

Zühlke, S., M. Zwick, S. Scharnhorst und T. Wende (2004), The research data centres of the Federal Statistical Office and the statistical offices of the *Länder*, *Schmollers Jahrbuch* / *Journal of Applied Social Science Studies* 124, 567-578.

Zwick, M. (2001), Individual tax statistics data and their evaluation possibilities for the scientific community, *Schmollers Jahrbuch* / *Journal of Applied Social Science Studies* 121, 639-648.

Zwick, M. (2007), CAMPUS-Files – Kostenfreie Public Use Files für die Lehre, *AStA – Wirtschafts- und Sozialstatistisches Archiv*, Heft 4/2007 (im Druck).