Mikrodaten der amtlichen Statistik in einer Open-Data-Welt – Nationale und internationale Herausforderungen*

Manfred Ehling

1 Einleitung

Open Data steht für die Öffnung von Informationen und Daten des öffentlichen Sektors für die Allgemeinheit, eine Forderung, die vor etwa einem Jahrzehnt in der öffentlichen Diskussion aufkam. Einer der frühen Meilensteine dieser Forderung nach offenen Regierungsinformationen war eine Tagung im Jahr 2002 und die anschließende Veröffentlichung der Ergebnisse mit dem Titel „Open Government: Fostering Dialogue with Civil Society“ (Vgl. OECD 2003). Seitdem sind vielfältige nationale und internationale Open Data Initiativen entstanden, die einen freien Zugang und uneingeschränkte Nutzung vor allem von Daten der öffentlichen Verwaltung zum Ziel haben.

Auf der anderen Seite spielen offene Behördendaten im Prozess der Öffnung von Regierung und Verwaltung gegenüber dem Bürger eine wichtige Rolle. Ein partnerschaftliches Verhältnis zum Bürger soll durch Offenheit, Transparenz,

* Für die kritische Diskussion und Anregungen zu dem Beitrag danke ich Dr. Stefan Linz, Heike Habla und Mathias Zenke.

D. Hirschel et al. (Hrsg.), *Daten in der wirtschaftswissenschaftlichen Forschung,*
DOI 10.1007/978-3-658-03456-6_2, © Springer Fachmedien Wiesbaden 2013

Teilhabe und Zusammenarbeit geprägt sein. Diese Entwicklung wird in den angelsächsischen Ländern als „Open Government“ bezeichnet.

Die Ideen zu offenen Daten und hier insbesondere offenes Verwaltungshandeln wurden ab 2009 von den Regierungen - zuerst in den USA (Obama 2009), dann Großbritannien, Kanada und Neuseeland - aufgegriffen und mit der Ankündigung von Initiativen zur Öffnung von Verwaltungsinformationen populär gemacht. In Deutschland ist das Thema offene Daten 2010 auf der politischen Agenda angekommen und hat z.B. zu dem Regierungsprogramm „Vernetzte und Transparente Verwaltung“ und dem Projekt “Open Government” beim Bundesministerium für Inneres (2010) geführt. Auch auf Länderebene gibt es zu diesem Thema vielfältige Arbeiten. In einigen Bundesländern arbeiten Projektgruppen an Vorschlägen für eine Open-Government-Strategie. Ende Mai 2013 hatten sechs Länder ein Open-Data-Portal eröffnet. Auch zahlreiche Kommunen veröffentlichen Datensätze im Internet und haben entsprechende Portale aufgebaut ebenso die EU (http://open-data.europa.eu). Das Datenportal der Europäischen Union bietet einen zentralen Zugang zu dem Datenbestand der Institutionen und anderen Einrichtungen der EU. Ende Mai 2013 waren hier knapp 6000 Datenquellen verzeichnet, die zu 98% von Eurostat, dem Statistikamt der EU, stammten.

Nicht immer klar getrennt vom Open Data-Ansatz ist die Diskussion um die Veröffentlichung administrativer Unterlagen im Zusammenhang mit Anfragen nach dem Informationsfreiheitsgesetz (Gesetz zur Regelung des Zugangs zu Informationen des Bundes), das Anfang 2006 in Kraft getreten ist. Hiernach hat jede Person einen Rechtsanspruch auf Zugang zu amtlichen Informationen von Bundesbehörden. Eine „Amtliche Information“ ist jede amtlichen Zwecken dienende Aufzeichnung, unabhängig von der Art der Speicherung als Akte oder elektronische Information. Die Informationsfreiheit (besser der Informationszugang) hat zahlreiche Ausnahmeregelungen und schließt personenbezogene und betriebsbezogene Daten aus (vgl. Mensching 2006, S. 1ff). Veröffentlichungspflichten sind in dem Gesetz nur geregelt, soweit es sich um Verzeichnisse handelt, aus denen sich die vorhandenen Informationssammlungen und Informationszwecke erkennen lassen, z.B. Organisations- oder Aktenpläne. Noch nicht alle Bundesländer haben für ihren Zuständigkeitsbereich jeweils eigene ähnliche Gesetze erlassen (vgl. Bundesbeauftragter für den Datenschutz und die Informationsfreiheit 2012, S. 33f).

Im akademischen Bereich spricht man von Open Access als dem freien Zugang zu wissenschaftlichen Informationen. Zuerst bezog sich die Forderung auf den freien Zugang zu wissenschaftlichen Publikationen, jetzt umfasst die Diskussion auch die Bereitstellung von Forschungsprimärdaten und weiterer Materialien (vgl. Fahrenberg 2012). Der offene Zugang zu Forschungsdaten etwa zur Überprüfung veröffentlichter Ergebnisse durch Re-Analyse, für Replikationsstudien oder für die klassische Sekundärdatenforschung soll in diesem Beitrag nicht thematisiert werden (Klump 2012).

Der vorliegende Beitrag diskutiert zuerst die Definition und die Kriterien offener Verwaltungsdaten mit Blick auf die Besonderheiten der Bereitstellung amtlicher Mikrodaten. Dann werden Formen der Bereitstellung dieser Daten vorgestellt. Im nächsten Abschnitt werden ausgewählte internationale und nationale Angebote amtlicher Mikrodaten für die Open-Data-Welt dargelegt. In einem abschließen-den Fazit werden Chancen für das weitere Vorgehen zur Erstellung und die Be-reitstellung von amtlichen Mikrodaten als Open Data angesprochen.

2 Was sind offene Verwaltungsdaten (Open Government Data)

Mit dem Begriff "offene Daten" (Open Data) werden Daten umschrieben, die uneingeschränkt zugänglich sind und ohne Lizenz verwendet werden können. „Laut der Definition der Open Knowledge Foundation ist freies Wissen als ein Gegenstand oder Werk zu verstehen, mit dem Wissen transferiert wird und das verschiedene Kriterien erfüllt. Das Werk sollte u. a. im Ganzen zugänglich sein, einer diskriminierungsfreien Lizenz unterliegen, die eine Weiterverteilung und Wiederverwendung erlaubt, ohne dass der Nutzer dabei technologischen Restriktionen unterliegt. Die Verpflichtung, bei der Wiederverwendung die Urheber zu nennen und ein verändertes Werk als solches zu kennzeichnen, kann Bestandteil dieser Lizenz sein. Zudem muss die Lizenz des Werkes mit diesem weiterverteilt werden, dabei Gültigkeit behalten und gleichzeitig nicht die Weiterverteilung anderer Werke behindern“ (Both, Schieferdecker 2012, S.21).

Im Open Data-Ansatz werden diese Punkte für Datenbestände aus allen Bereichen aufgegriffen und konkretisiert. Offene Daten können demnach durch jedermann und für jegliche Zwecke genutzt, verarbeitet und weiterverbreitet werden. Die freie Verfügbar- und Nutzbarkeit bezieht sich gerade auch auf Informationen, die nicht in Textform vorliegen, wie zum Beispiel Wetterdaten, Karten, Statistiken oder medizinische Daten.

Ob die bereitgestellten Daten als offen bezeichnet werden können, ist abhängig von verschiedenen Faktoren wie der Zugänglichkeit, den Formaten und den rechtlichen Bedingungen, unter denen die Daten genutzt werden dürfen. Open Government Data oder offene Verwaltungsdaten, wie im folgenden Text verwendet, werden in der Studie von Lucke/Geiger (2010, S. 6) wie folgt definiert: „Offene Verwaltungsdaten sind jene Datenbestände des öffentlichen Sektors, die von Staat und Verwaltung im Interesse der Allgemeinheit ohne jedwede Einschränkung zur freien Nutzung, zur Weiterverbreitung und zur freien Weiterverwendung offen zugänglich gemacht werden." Die Grundsätze von Open Data werden dabei auf Daten der öffentlichen Verwaltung angewendet.

Aus der internationalen Diskussion über offene Daten haben sich zehn Prinzipien herauskristallisiert, die aufzeigen, zu welchem Grad Regierungsdaten offen zugänglich sind. Ob Verwaltungsdaten nun tatsächlich offen sind lässt sich an der Erfüllung dieser zehn Kriterien festmachen (vgl. hierzu und zum Folgenden: Sunlight Foundation 2010 sowie die deutschsprachigen Adaptionen bei Lucke/Geiger (2010, S. 2f) und Klessmann u.a. 2012, S. 58ff).

Vollständigkeit

Datensätze öffentlicher Einrichtungen sollten so vollständig wie möglich veröffentlicht werden, soweit dies die Regelungen zum Datenschutz zulassen. Die Daten sollen dabei in einem möglichst ursprünglichen Format inklusive beschreibender Metadaten veröffentlicht werden. Nutzern soll so ermöglicht werden, die Daten mit einer größtmöglichen Detaillierung zu analysieren.

Primärquellen

Offene Verwaltungsdaten sollten direkt aus den ursprünglichen Quellen (Rohdaten) veröffentlicht und mit Informationen zur Erhebung und Aufbereitung der Daten angereichert werden. Die öffentliche Verbreitung der Daten soll es den Nutzern erlauben zu verifizieren, ob die Daten angemessen erhoben und korrekt erstellt wurden.

Zeitliche Nähe

Die Veröffentlichung der Daten soll möglichst zeitnah erfolgen, d.h. sobald aus den erhobenen Daten ein Datensatz erstellt wurde, sollten die Informationen der Öffentlichkeit zur Verfügung gestellt werden. Daten, deren Nutzen zeitabhängig ist, sollten vorrangig mit hoher Aktualität bereitgestellt werden.

Leichter Zugang

Der Aufwand, um Zugang zu einem oder mehreren Datensätzen zu erhalten, soll möglichst gering sein. Daten sollten einfach aufzufinden und herunterzuladen sein. Dabei sollten keine technischen Restriktionen, wie bei browserorientierten Technologien (z.B. über Eingabemasken) bestehen, sondern sämtliche interessierende Daten sollten auf einmal herunterzuladen sein.

Maschinenlesbarkeit

Viele Potenziale offener Daten lassen sich erst durch die Möglichkeit einer automatisierten Verarbeitung ausschöpfen. Maschineninterpretierbarkeit eines Datensatzes erlaubt die einfache Einbindung in Softwareanwendungen. Große Mengen von Daten lassen sich dementsprechend in neuen Applikationen nutzbar machen.

Diskriminierungsfreiheit

Diskriminierungsfreiheit bezieht sich darauf, wer auf die Daten zugreifen kann und wie dieser Zugriff erfolgt. Ein diskriminierungsfreier Zugang bedeutet, dass jede Person zu jeder Zeit Zugriff auf die Daten hat, ohne Registrierungszwang, zeitliche Restriktionen oder andere einschränkende Voraussetzung einer Datennutzung.

Verwendung offener Standards

Mit dem Einsatz offener Standards ist die Verwendung von Formaten gemeint, die nicht nur von ausgewählten Programmen gelesen und verarbeitet werden können. Offene Standards gewährleisten die Freiheit, mit verschiedenen Programmen auf die Daten zugreifen zu können, ohne dass dafür Lizenzkosten an einzelne Hersteller abgeführt werden müssen.

Lizenzierung

Eine restriktive Lizensierung wirkt als Hürde für die öffentliche Verwendung von Daten und stellt ein Hindernis für deren Nutzung dar. Öffentliche Informationen

sind als solche zu kennzeichnen, sollen aber ohne Nutzungsbeschränkungen frei verfügbar sein, d.h. auch für die kommerzielle Nutzung.

Dauerhaftigkeit

Dauerhaftigkeit bedeutet, dass Daten über eine lange Zeit im Internet zugänglich sind. Offene Verwaltungsdaten sollten dauerhaft für die Öffentlichkeit in Archiven verfügbar sein. Werden Informationen aktualisiert, verändert oder entfernt, sollte dies angemessen dokumentiert werden. Einmal online gestellte Daten sollten online bleiben einschließlich einer Dokumentation der Versionen und der Metadaten.

Nutzungskosten

Eine der größten Hürden beim Zugriff auf öffentlich verfügbare Informationen sind auferlegte Nutzungskosten, selbst wenn diese gering sind. Die kostenpflichtige Bereitstellung von Verwaltungsdaten behindert die Weiterverwendung und die Erhebung von Gebühren oder sonstigen Abgaben beschränkt generell die Gruppe der Nutzer sowie den Einsatz der Daten zu wirtschaftlichen Zwecken.

Diese zehn Prinzipien offener Verwaltungsdaten werden im Folgenden mit Blick auf die Bereitstellung von Mikrodaten durch die amtliche Statistik diskutiert. Aus den rechtlichen Rahmenbedingungen in denen sich die amtliche Statistik bewegt, ergeben sich Einschränkungen bei der Weitergabe der gewonnenen Einzeldaten. Ein wesentliches Open Data-Prinzip besagt, dass keine Daten veröffentlicht werden dürfen, die einen Rückschluss auf einzelne natürliche Personen zulassen.

Amtlich erhobene Daten unterliegen darüber hinaus der strengen Geheimhaltung nach dem Bundesstatistikgesetz (§16). Einzelangaben über persönliche und sachliche Verhältnisse sind danach geheim zu halten, soweit durch besondere Rechtsvorschriften nichts anderes bestimmt ist. Ausnahmen von dieser Regelung ergeben sich, wenn der Befragte in die (personenbezogene) Veröffentlichung seiner Angaben einwilligt oder die Einzelangaben beziehen sich auf eine Person öffentlichen Rechts, Behörden des Bundes und der Länder sowie Gemeinden und Gemeindeverbände und sind unabhängig von der Statistik öffentlich zugänglich. Wenn Einzelangaben mit denen anderer Befragter zusammengefasst und in statistischen Ergebnissen dargestellt sind oder die Einzelangaben sind dem Befragten

nicht zuzuordnen, also Ergebnisse absolut anonym sind, ist ebenfalls eine Veröffentlichung möglich.

Aus den rechtlichen Gegebenheiten folgt, dass amtliche Einzeldaten nur in einem begrenzten rechtlich klar geregeltem Rahmen übermittelt werden können. Vollständige Mikrodaten werden in der Regel nicht zur Verfügung gestellt, weil dann bei vielen Statistiken über die Information der Teilnahme an einer Erhebung durch wenige Merkmalskombinationen eine Reidentifikation erfolgen könnte. Eine zentrale Maßnahme zur Anonymisierung ist daher unter anderem die Ziehung von Stichproben sowie eine Auswahl von Merkmalen ggf. Vergröbert und damit die Weitergabe eines nicht vollständigen Datensatzes. Die Daten werden bearbeitet, um rechtliche Vorgaben (Datenschutz, Statistikgeheimnis) zu erfüllen.

Die Primärquellen oder Rohdaten werden nicht weitergegeben, weil sie häufig fehlende Werte, Unplausibilitäten oder Fehler aufweisen. In einer sogenannten Plausibilitätskontrolle werden die gewonnenen Daten teils automatisch, teils manuell auf ihre Plausibilität hin geprüft und bei Bedarf entsprechend korrigiert, d.h. im Rahmen der Datenbearbeitung werden die gewonnenen Werte oder die Ergebnisse daraufhin überprüft, ob sie plausibel, also annehmbar, einleuchtend und nachvollziehbar sind. Ist das nicht der Fall werden durch Nachfrage oder auf Basis von Berechnungsverfahren die unstimmigen Werte berichtigt. Die ursprünglichen Daten werden also mit der Absicht bearbeitet, validere Ergebnisse zu erzielen. Die unbearbeiteten Rohdaten weiterzugeben ist deshalb nicht sinnvoll, weil sich erst über die „Veredlung“ der Daten das endgültige publizierte Ergebnis nachvollziehen lässt.

Der Bearbeitungsaufwand nach der Gewinnung der Daten und Maßnahmen zur absoluten Anonymisierung führen in der Regel dazu, dass eine Veröffentlichung der Daten erst mit einigem Abstand zur ersten Ergebnisveröffentlichung erfolgen kann. Welcher Zeitraum hier als zeitnah zu verstehen ist, ist dabei vom Einzelfall abhängig und statistische Daten sind bspw. anders einzuordnen als unsensible Daten wie z.B. zur Auslastung von Verkehrssystemen.

Die einzigen bisher frei zugänglichen Mikrodaten der deutschen amtlichen Statistik sind die für den Einsatz in der akademischen Lehre geschaffenen Campus Files (siehe auch Kapitel 4.2), die einfach über das Internet zugänglich sind und ohne Registrierung kostenfrei herunter geladen werden können. Die Daten werden in den Formaten SPSS, SAS und STATA sowie als ASCII-CSV angeboten. D.h. auf die Daten kann jeder ohne Identifizierung zugreifen und sie werden

in einem hersteller- und plattformunabhängigen Dateiformat angeboten. Die zugrunde liegende Datenstruktur und entsprechende Standards sind öffentlich zugänglich; ebenso sind die Metadaten vollständig publiziert und kostenfrei erhältlich. Eine Weiterverarbeitung der Daten ist so möglich, Vervielfältigung und Verbreitung, auch auszugsweise ist mit Quellenangabe gestattet.

Nach einem kurzen Abschnitt über die Verbreitung von Open Government Data über entsprechende Portale werden die Möglichkeiten der Bereitstellung von amtlichen Mikrodaten in Kapitel 4 detailliert diskutiert.

3 Bereitstellung von offenen Verwaltungsdaten

Weltweit haben bereits viele Staaten Open-Data-Portale eingerichtet, über die man auf staatliche Daten zugreifen kann. Hinter der Einrichtung dieser Portale stehen häufig Erklärungen der jeweiligen Regierungen zu einer offenen und transparenten Verwaltung mit der Verpflichtung, die Daten der öffentlichen Hand in einer leicht verfügbaren Form möglichst kostenlos und ohne Restriktionen bei der Nutzung zugänglich zu machen. Die offenen Regierungsdaten sollen die Verwaltung effizienter und verantwortlicher machen und mehr Mitwirkung bei Entscheidungen sicherstellen. Außerdem sollen sich so die politischen und wirtschaftlichen Initiativen zentraler, regionaler und lokaler Behörden besser aufeinander abstimmen lassen. Auch andere Daten, deren Erstellung aus Steuermitteln bezahlt wurde, sollen im Interesse einer breiten Nutzung in der Forschung, dem Bildungssystem aber auch im kommerziellen Sinn offengelegt werden. Da die Politik, in ihren Entscheidungsprozessen auf hochwertige Statistikdaten angewiesen ist, liefert die amtliche Statistik einen wesentlichen Beitrag zu Transparenz und öffentlicher Nachprüfbarkeit, von der europäischen bis zur kommunalen Ebene (vgl. Both, Schieferdecker 2012, S. 20f).

Das Datenportal für Deutschland (www.govdata.de) bietet einen einheitlichen, zentralen Zugang zu Verwaltungsdaten aus Bund, Ländern und Kommunen und will ein nachhaltiges Angebot an frei zugänglichen Verwaltungsdaten für Bürgerinnen und Bürger, die Wirtschaft und andere Verwaltungseinheiten bereit-

stellen. Die Plattform stellt damit einen zentralen Zugang zu Verwaltungsdaten über alle Verwaltungsebenen hinweg dar und will eine systematische Sammlung, Klassifizierung und Katalogisierung geeigneter und verfügbarer Daten sein. Ziel ist es, diese Daten an einer Stelle auffindbar und so einfacher nutzbar zu machen. Noch nicht an allen Stellen ist die Verwendung offener Lizenzen möglich, aber im Sinne von „Open Data" soll dies gefördert werden, ebenso wie das Angebot an maschinenlesbaren Rohdaten. Deshalb beinhaltet das Portal „GovData“ auch nicht nur offene Daten, sondern auch solche, die eingeschränkt nutzbar sind. Zudem erlaubt das Portal – bisher nur in Teilen – ein systematisches Verlinken und einfaches Auffinden von Informationen und maschinenlesbaren Daten - basierend auf verbindlichen technischen Standards (vgl. Klessmann 2012, S. 277ff).

Der Erfolg des Portals hängt maßgeblich davon ab, dass Datensätze zugänglich gemacht werden, die für potenzielle Nachnutzer interessant und relevant sind. Bis heute sind in Deutschland viele relevante Datensätze nicht als offene Daten zugänglich. Aus der amtlichen Statistik ist das zentrale Informationssystem des Statistischen Bundesamtes GENESIS-Online verfügbar, das ein breites Themenspektrum fachlich tief gegliederter Ergebnisse enthält. Der Tabellenabruf ist kostenfrei und kann durch zeitliche, sachliche und gegebenenfalls regionale Auswahlmöglichkeiten dem individuellen Bedarf angepasst werden.

Die Zensusdatenbank (https://ergebnisse.zensus2011.de), über die einfach, schnell und flexibel die Ergebnisse des Zensus 2011 abgefragt werden können, ist ein zweites Beispiel für den offenen Zugang zu Daten der amtlichen Statistik. Zum einen können hier in vordefinierten Tabellen und Diagrammen Einwohnerzahlen oder zusammengefasste Ergebnisse für Deutschland insgesamt oder nach Regionen und Städten abgerufen werden. Zum anderen können von den Nutzern aus dem Datenmaterial Informationen ausgewählt werden, die sie interessieren, um variable Tabellen und Diagramme zu erstellen, beispielsweise für Gemeinde- oder Regionalvergleiche.

4 Amtliche Mikrodaten für die Open-Data-Welt – Mögliche Angebote

Bisher werden in den Open-Data-Portalen vor allem aggregierte Daten aus amtlichen Statistiken angeboten. Über die Nutzung von aggregierten Daten hinaus ist eine wachsende Nutzergruppe an der Bereitstellung von Einzeldaten für eine freie Nutzung interessiert. Diese bieten eine wesentlich größere Informationsfülle als Daten in aggregierter Form, wie sie auch in den Publikationen der amtlichen Statistik üblicherweise präsentiert werden. Die Wünsche der Nutzer gehen in Richtung Einzeldaten, weil diese ein viel breiteres Analysepotenzial beinhalten oder weil die Daten der amtlichen Statistik als Input für neu erstellte Anwendungen genutzt werden sollen. Werden Einzeldaten weitergeben, sind sie im Sinn des Statistikgeheimnisses und als Maßnahme des Datenschutzes gegenüber den Befragten so zu anonymisieren, dass eine direkte oder indirekte Identifizierung eines konkreten Einzelfalls nicht mehr möglich ist. Dies führt zu Restriktionen im Datenzugang. Nachfolgend sollen als Beispiele für ein Angebot an Einzeldaten aus der amtlichen Statistik zuerst ein internationales Datenportal zu Zensusdaten vorgestellt werden, danach wird einleitend die Weitergabe von Mikrodaten an die Wissenschaft vorgestellt und dann wird ein spezielles Datenangebot zur Unterstützung der akademischen Lehre vorgestellt.

4.1 Volkzählungsdaten aus aller Welt

Die meisten Bevölkerungsdaten – insbesondere Volkszählungsdaten – waren traditionell sowohl für die Allgemeinheit als auch die Wissenschaft lediglich in aggregierter Form zugänglich.

Das Projekt „Integrated Public Use Microdata Series – International“ stellt seinen Nutzerinnen und Nutzern Volkszählungsdaten aus 68 Ländern mit 211 Zensen und 480 Mill. Datensätzen zur Verfügung (Stand Mai 2013). Aktuelle und historische Volkszählungsdaten aus der ganzen Welt der internationalen Wissenschaft in einer Internetdatenbank (http://international.ipums.org/international/) unentgeltlich zur Verfügung zu stellen, ist das Ziel dieses Projektes, das vom

http://www.springer.com/978-3-658-03455-9

Daten in der wirtschaftswissenschaftlichen Forschung
Festschrift zum 65. Geburtstag von Prof. Dr. Joachim Merz
(Eds.)D. Hirschel; P. Paic; M. Zwick
2013, XV, 309 S. 17 Abb., 5 Abb. in Farbe., Softcover
ISBN: 978-3-658-03455-9