gesis Leibniz-Institut für Sozialwissenschaften

Empfehlungen zur Anonymisierung quantitativer Daten

1 Empfehlungen zur Anonymisierung quantitativer Daten

Verfasst von: Thomas Ebel (GESIS), thomas.ebel@gesis.org[1]

Das vorliegende Dokument soll Forschern und Forscherinnen als eine praktisch orientierte Handreiche dienen, in der aufzeigt wird, wie quantitative Daten so anonymisieren werden können, dass ihrer Archivierung und Weitergabe an Dritte durch das Datenarchiv von GESIS (z. B. im Rahmen von datorium) nichts im Wege steht.

Weiterführende und detailliertere Hinweise zum Thema Datenschutz erhalten Sie in Kinder-Kurlanda/Watteler, 2015: Hinweise zum Datenschutz. Rechtlicher Rahmen und Maßnahmen zur datenschutzgerechten Archivierung sozialwissenschaftlicher Forschungsdaten. GESIS Papers 2015/01.

1.1 Warum sollten Daten anonymisiert werden?

Sozialwissenschaftliche Forschung befasst sich i. d. R. mit Daten von Menschen. Viele dieser Daten betreffen private oder sensible Lebensbereiche bzw. Informationen der Studienteilnehmer. Werden Forschungsergebnisse archiviert oder Dritten zur Verfügung gestellt, müssen kritische, v. a. personenbezogene Daten, das sind „Einzelangaben über persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren natürlichen Person." (§3 Abs. 1, Bundesdatenschutzgesetz BDSG), zuvor anonymisiert werden, um die Studienteilnehmer vor Identifizierung und unerwünschten kommerziellen Interessen zu schützen, die Vorgaben des Datenschutzes zu erfüllen sowie ethisch vertretbare Forschung zu betreiben.

Für die Archivierung im Datenarchiv von GESIS müssen die Daten in mindestens formal anonymisierter Form eingereicht werden. Formale Anonymisierung umfasst das Entfernen aller direkten (personenbezogenen) Identifikatoren, beispielsweise indem die Studienteilnehmer über nichtsprechende, numerische IDs anstelle ihrer Namen repräsentiert werden (sog. Pseudonymisierung). Die Weitergabe der Daten an andere ForscherInnen (bspw. für Sekundäranalysen) macht darüber hinaus eine faktische Anonymisierung erforderlich, die mit strikteren Anforderungen einhergeht. Faktische Anonymisierung bedeutet, dass Einzelangaben über persönliche oder sachliche Verhältnisse nicht mehr oder nur mit unverhältnismäßig hohem Aufwand einer Person zugeordnet werden können (§3 Abs. 6, BDSG). Dazu müssen nicht nur personenbezogene, sondern auch personenbeziehbare Angaben innerhalb der Daten identifiziert und anschließend modifiziert oder gelöscht werden, falls zu befürchten ist, dass diese, ggf. in Kombination miteinander oder unter Zuhilfenahme von Informationen aus externen Quellen, eine Identifikation der Befragten ermöglichen könnten. [1]

Im weiteren Verlauf dieses Dokuments wird erklärt, wie ForscherInnen quantitative Daten faktisch anonymisieren, da dies i. d. R. die Mindestvoraussetzung für die Weitergabe von Forschungsdaten an Dritte darstellt (u. a. auch in datorium). [2]

1.2 Welche Informationen sind zu anonymisieren?

Anonymisierung erfolgt in erster Linie, um die Identifikation von StudienteilnehmerInnen zu verhindern (vgl. §3 Abs. 6, BDSG). Dazu müssen alle Daten, die dieses Risiko erhöhen, gelöscht oder modifiziert werden.

---

[1] Dieses Dokument wurde im Rahmen des Verbunds Forschungsdaten Bildung erstellt und zuerst auf forschungs-daten-bildung.de veröffentlicht.

Direkte (personenbezogene) Identifikatoren umfassen alle Angaben, mit denen eine Person direkt identifiziert werden kann, beispielsweise Name, Anschrift, Telefonnummer, Kfz-Kennzeichen und Email-Adresse. [3] Sie sollten unverzüglich gelöscht werden, wenn die Daten für die Archivierung bereitgestellt werden sollen (formale Anonymisierung).

Indirekte oder personenbeziehbare Identifizierungsmerkmale hingegen erlauben nur in Kombination mit anderen Angaben eine Identifizierung der Studienteilnehmer und müssen dann ebenfalls geprüft und ggf. anonymisiert werden. Dazu zählen insbesondere kleinräumige regionale Informationen (bspw. Anschrift, Name des Wohnortes oder der Gemeinde, Postleitzahl), detaillierte Berufs- und Bildungsangaben (bspw. offene Berufsangaben oder vierstelliger ISCO-Code [4]), spezielle Erhebungskontexte (Nennung von Institutionen, Fachgebieten, Studiengängen etc. und öffentliche Exponiertheit der Befragten bspw. bei Elitenstudien) und alle offenen Angaben, selbst wenn die jeweiligen Fragestellungen an sich unproblematisch sind. [5]

1.3 Anonymisierungsstrategien

Im Folgenden werden in Anlehnung an Kinder-Kurlanda/Watteler 2015 drei Vorgehensweisen für die Anonymisierung quantitativer Daten unterschieden: 1) Einzelne Werte oder Kategorien einer Variable werden vergröbert (aggregiert), 2) alle Werte einer Variable werden vergröbert und 3) die betroffene Variable wird komplett gelöscht. [6] Da alle Anonymisierungsstrategien mit einem Informationsverlust einhergehen, sollte sorgfältig zwischen datenschutzrechtlich notwendigen Maßnahmen und dem verbliebenen Nachnutzungspotential der anonymisierten Daten abgewogen werden.

1.3.1 Strategie 1: Aggregieren einzelner Werte/Kategorien

Wenn einzelne Werte einer Variablen sehr selten vorkommen, insbesondere an den Rändern ihrer Verteilung, kann dies die Reidentifizierung von Studienteilnehmern ermöglichen. Beispielsweise ist ein Teilnehmer einer Studie in einer Kleinstadt in Niedersachsen 110 Jahre alt, oder er hat ein monatliches Nettoeinkommen von 10.200 Euro oder er ist aramäischer Muttersprachler. Die Identität einer solchen Person könnte womöglich unter Zuhilfenahme weiterer (Regional-) Informationen der Studie oder externer Quellen durch Dritte ermittelt werden. In diesem Fall wird empfohlen, die problematischen Werte/Kategorien der Variable zu aggregieren, beispielsweise sehr kleine respektive sehr große Werte in einer nach unten respektive nach oben offenen Kategorien zusammenzufassen.

Sensible Informationen in Variable …	Ursprüngliche Kodierung	Mögliches Datenschutz-Problem	Exemplarische Lösung
Alter	Offene Angaben	Seltene Werte an den Rändern der Verteilung (z. B. 110 Jahre)	Wie zuvor, aber mit einer nach oben offenen Kategorie „Alter größer als 90 Jahre“, die mehrere Fälle zusammenfasst
Einkommen	Offene Angaben	Seltene Werte an den Rändern der Verteilung (z. B. 10.200 Euro)	Wie zuvor, aber mit einer nach oben offenen Kategorie „monatliches Netto-Einkommen größer als 7500 Euro“, die mehrere Fälle zusammenfasst
Muttersprache	Offene Angaben	Seltene Werte (z. B. Aramäisch)	Wie zuvor, aber alle seltenen Werte zu einer gemeinsamen Kategorie umcodiert [7]

Abbildung 1: Aggregieren einzelner Werte

1.3.2 Strategie 2: Aggregieren aller Werte

Wenn nicht nur einzelne Werte, sondern die Informationen einer Variablen generell datenschutzrechtlich problematisch sind, dann sollte die gesamte Variable recodiert werden, sofern dieser Aufwand geleistet werden kann und die recodierte Variable einen tatsächlichen Informationsgehalt besitzt.

Sensible Informationen in Variable ...	Ursprüngliche Kodierung	Exemplarische Lösung
Berufsangabe	Offene Angaben	Kategorisierung der Angaben in (dreistellige) ISCO-Codes
Berufsangabe	Vierstellige ISCO-Codes	Kürzen auf dreistellige Angaben (Löschen der letzten Ziffer)
Postleitzahl	Postleitzahl	Aggregieren in z. B. Bundesland oder BIK-Region

Abbildung 2: Aggregieren aller Werte

1.3.3 Strategie 3: Löschen von Variablen

Als dritte Strategie sollte das Löschen ganzer Variablen aus dem Datensatz ins Auge gefasst werden. Da der Informationsverlust hierbei am größten ist, sollte das Löschen nur durchgeführt werden, wenn das Aggregieren (entweder einzelner Werte oder aller Variablenwerte) zu aufwendig ist oder aus sonstigen Gründen (bspw. Informationsgehalt der recodierten Variable rechtfertigt den Aufwand nicht etc.) nicht geleistet werden kann oder soll.

Sensible Informationen	Exemplarische Lösung
Kleinräumige Regionalangaben wie Postleitzahl, Kreiskennziffern, Orts- oder Stadtteilnamen	Aggregieren in z. B. Bundesland oder BIK-Region u. U. zu aufwendig, daher die Variable löschen
(Fixe) IP-Adressen	Der Informationsgehalt lässt sich nicht erhalten beim Aggregieren der Werte, daher die Variable löschen

Abbildung 3: Löschen von Variablen

Unabhängig von der gewählten Anonymisierungsstrategie sollte auf eine konsistente und sorgfältige Arbeitsweise geachtet werden. Alle Anonymisierungsentscheidungen müssen in einer nachvollziehbaren Weise verschriftlicht werden (bspw. im Codehandbuch).

1.4 Weitere Schutzmaßnahmen

Für den seltenen Fall, dass keine Anonymisierung quantitativer Daten durchführbar ist, oder als ergänzende Maßnahmen kann der Zugriff auf die Daten beschränkt werden. Bspw. kann die Datennachnutzung über das Datenarchiv von GESIS technisch (per Passwort/Registrierung) und inhaltlich (bspw. Daten ausschließlich für wissenschaftliche Zwecke freigegeben) beschränkt erfolgen. Außerdem können Daten zeitlich befristet unter ein Embargo gesetzt werden, sodass sie erst zu einem späteren Zeitpunkt Dritten zur Verfügung stehen. Den Datengebern wird darüber hinaus auf Wunsch das Recht eingeräumt, der Weitergabe Ihrer Daten im Einzelfall zustimmen zu müssen.

Wenn Ihrerseits weiterhin Fragen zur Anonymisierungsfähigkeit Ihrer Daten bestehen, wenden Sie sich gerne an das Datenarchiv von GESIS bzw. die Kuratoren von datorium. Diese werden Ihnen gerne unterstützend und beratend zur Seite stehen.

1.5 Literatur

BDSG, Bundesdatenschutzgesetz, 1990: Stand: Neugefasst durch Bek. v. 14.1.2003 I 66; zuletzt geändert durch Art. 1 G v. 14.8.2009 I 2814: www.gesetze-im-internet.de/bdsg_1990, abgerufen am 05.03.2015.

Häder, Michael, 2009: Der Datenschutz in den Sozialwissenschaften. Anmerkungen zur Praxis sozialwissenschaftlicher Erhebungen und Datenverarbeitung in Deutschland (RatSWD Working Paper Series. 90). http://www.ratswd.de/download/RatSWD_WP_2009/RatSWD_WP_90.pdf, abgerufen am 06.03.2015.

Jensen, Uwe, 2012: Leitlinien zum Management von Forschungsdaten. Sozialwissenschaftliche Umfragedaten. GESIS-Technical Reports 2012|07. Online-Dokument: www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2012/TechnicalReport_2012-07.pdf, abgerufen am 05.03.2015.

Katharina Kinder-Kurlanda & Oliver Watteler, 2015: Hinweise zum Datenschutz. Rechtlicher Rahmen und Maßnahmen zur datenschutzgerechten Archivierung sozialwissenschaftlicher Forschungsdaten. GESIS Papers 2015|01, Verfügbar unter: http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_papers/GESIS-Papers_2015-01.pdf, abgerufen am 05.03.2015.

Metschke, Rainer & Rita Wellbrock, 2002: Datenschutz in Wissenschaft und Forschung. Materialien zum Datenschutz Nr. 28., 3. Aufl. Berlin, 2002: www.datenschutz-berlin.de/attachments/47/Materialien28.pdf?1166527077, abgerufen am 05.03.2015.

Meyermann, Alexia & Maike Porzelt, 2014: Hinweise zur Anonymisierung von qualitativen Daten. In: forschungsdaten bildung informiert, Nr. 1. Frankfurt am Main: Deutsches Institut für Internationale Pädagogische Forschung. Verfügbar unter: www.forschungsdaten-bildung.de/fdb-informiert, abgerufen am 05.03.2015.

1.6 Fußnoten

[1] Personenbeziehbare Daten sind „Einzelangaben, die eine bestimmte Person zwar nicht eindeutig oder unmittelbar identifizieren, die es aber erlauben, die Identität der Person mit Hilfe anderer Informationen festzustellen." (Metschke/Wellbrock 2002: 19).

Zurück zum Text

[2] Ist die Identifizierung der Studienteilnehmer in jedem Fall ausgeschlossen, spricht man von absoluter Anonymisierung. Absolut anonymisierte Daten können jederzeit ohne Bedenken archiviert und zur Nachnutzung verfügbar gemacht werden, solange die Einwilligungen der Teilnehmer in die Verarbeitung der ursprünglich personenbezogenen Daten vorliegen (Metschke/Wellbrock 2002: 20) und die Datenweitergabe nicht explizit ausgeschlossen wurde.

Zurück zum Text

[3] Für eine umfangreichere Auflistung siehe Kinder-Kurlanda/Watteler 2015.

Zurück zum Text

[4] International Standard Classification of Occupations, ISCO: www.ilo.org/public/english/bureau/stat/isco/isco08/index.htm, abgerufen am: 04.05.2015.

Zurück zum Text

[5] So können Studienteilnehmer bei an sich unbedenklichen Fragen mit offener Antwortmöglichkeit kritische Informationen preisgeben, die zu einer Identifikation führen könnten.

Zurück zum Text

[6] Es gibt weitere Anonymisierungsstrategien, bspw. das Einstreuen von Zufallsfehlern und das Veröffentlichen von Teildatensätzen mit verringerter Stichprobengröße (vgl. Metschke/Wellbrock 2002: 20ff), die aber hier nicht weiter verfolgt werden, da sie v. a. in speziellen Kontexten sinnvoll sind.

Zurück zum Text

[7] Eine Empfehlung hierzu lautet, alle Variablenausprägungen, die jeweils weniger als 5 Personen auf sich vereinen, zu einer gemeinsamen Kategorie zusammenzufassen.

Zurück zum Text