Volker Hammer, Michael Knopp

Datenschutzinstrumente Anonymisierung, Pseudonyme und Verschlüsselung

Anonymisierung, Pseudonyme und Verschlüsselung sind wichtige Instrumente für den Schutz personenbezogener Daten. Verantwortliche Stellen, die diese Instrumente einsetzen, erhoffen sich von ihrer Verwendung rechtliche Erleichterungen bei der Verarbeitung von ursprünglich personenbezogenen Daten. Wann aber ist Anonymität gut genug? Wie hängen Anonymität, Pseudonyme und Verschlüsselung zusammen?

1 Identifizierende Merkmale, Kontextwissen und gewonnene Informationen

Das Datenschutzrecht regelt den Umgang mit Daten, die sich auf natürliche Personen beziehen. Davon können Daten unterschieden werden, die mit Personen nichts zu tun haben, also beispielsweise ein Fahrplan für Züge, die Daten einer Prozessüberwachung oder die Größe von Planeten.[1]

Daten, die Einzelangaben über Personen machen, sind nach dem BDSG personenbezogen, wenn sie sich auf *eine bestimmte oder bestimmbare* natürliche Person beziehen (Betroffener, § 3 Abs. 1 BDSG).[2] Andere Datenbestände treffen zwar möglicherweise auch Aussagen über Personen, beziehen sich aber nicht auf eine bestimmte oder auch nur bestimmbare Person. Die einzelne Person kann unbestimmbar sein oder die Daten können sich auf eine Gruppe von Personen beziehen oder auch nur statistische Aussagen treffen. Diese Daten gelten häufig als nicht personenbezogen im Sinne des Datenschutzrechts. Im Einzelfall werden in der Literatur aber auch Beispiele angeführt, in denen auch für Gruppenangaben Personenbezug gegeben ist.[3]

Für nicht personenbezogene Daten gelten die Auflagen des BDSG nicht. Insbesondere können solche Daten ohne Befristung und ohne Zweckbegrenzung durch den jeweiligen Verarbeiter verwendet werden.[4] Die Verarbeitung solcher Daten ist auch durch Dritte zulässig, ohne dass das Datenschutzrecht eine Rechtsgrundlage oder einen Vertrag zur Auftragsdatenverarbeitung fordert – ein wichtiger Aspekt im Zeitalter des Cloud Computing.

Es gibt in vielen Fällen auch gute Gründe, nicht personenbezogene Daten weiter zu verarbeiten, weil dadurch wertvolle Erkenntnisse gewonnen und dokumentiert werden können, z. B. die Unfallschwerpunkte auf Straßen, die Häufigkeit von Nebenwirkungen eines Medikaments oder die Auslastung von Telekommunikationsdiensten. Manchmal will man auch Zeitreihen bestimmen, um Veränderungen festzustellen.

Nicht personenbezogene Daten werden in solchen Zusammenhängen häufig auch als anonyme Daten bezeichnet. Welche Bedingungen aber müssen gelten, damit Daten anonym sind? Um uns dieser Frage zu nähern, verwenden wir einige Begriffe aus der Diskussion um Anonymisierung, um den „Personenbezug von Einzelangaben" etwas techniknäher zu beschreiben. Gelegentlich werden im Zusammenhang mit Anonymisierung außerdem zwei weitere Instrumente genannt, mit denen personenbezogene Daten geschützt werden können: Pseudonyme und Verschlüsselung. In den Abschnitten 3 und 4 wollen wir deshalb das Verhältnis zwischen den drei Instrumenten betrachten.

1 Dammann in [Simitis2014], § 3 Rn. 57 ff grenzt *Sachdaten* ab. [Karg2015] weist auch darauf hin, dass diese in bestimmten Konstellation zeitweise auch Personenbezug aufweisen können, z. B. wenn ein Mobiltelefon von einem Betroffenen mitgeführt wird.

2 Gelegentlich werden auch andere Rechtssubjekte eingeschlossen, beispielsweise Speditionen als Halter oder Mieter von mautpflichtigen LKW nach dem BFstrMG.

3 z. B. Dammann in [Simitis2014] § 3 Rn. 14 und 19.

4 In [A29G2014], 12 wird allerdings darauf hingewiesen, dass für anonymisierte Daten noch Rechtsvorschriften außerhalb des Datenschutzrechts zum Schutz des Betroffenen gelten können.

Dr. Volker Hammer

ist Consultant der Secorvo GmbH. Seit Mitte 2003 unterstützt er die Toll Collect GmbH in verschiedenen Datenschutz-Projekten. Weitere Arbeitsschwerpunkte sind Public Key Infrastrukturen und kritische IT-Infrastrukturen

E-Mail: volker.hammer@secorvo.de

Michael Knopp, Jurist

Berater bei der Secorvo Security Consulting GmbH. Schwerpunkte: Datenschutz und Rechtsfragen im Kontext der IT-Sicherheit.

E-Mail: michael.knopp@secorvo.de

1.1 Merkmale von Betroffenen

Für die folgende Diskussion nehmen wir an, dass ein Rohdatenbestand in Form von Datensätzen dargestellt wird. Jeder Datensatz steht für einen Betroffenen. In einem Datensatz sind mehrere **Merkmale** enthalten; der Begriff Merkmal wird hier als „Container" für **Werte** verstanden. Jedes der Merkmale ist in einem konkreten Datensatz mit einem konkreten Wert belegt. Die Werte stehen für die Einzelangaben über den Betroffenen. [5]

Die Werte mancher Merkmale oder Merkmalskombinationen treten im Datenbestand nur einmal auf. Sie kennzeichnen dann einen Betroffenen eindeutig.[6] Dazu gehören z. B. der Name mit Anschrift und Geburtsdatum, die Telefonnummer, eine IBAN, eine Personalnummer, die Steuernummer oder die Umsatzsteuer-ID, Elemente des Fingerabdrucks oder Teile einer DNA. Merkmale oder Merkmalskombinationen, die im Datenbestand einmalige Werte enthalten, nennen wir **identifizierendes Merkmal**. Diesen Begriff verwenden wir hier umfassend:[7]

- Wir unterscheiden nicht zwischen identifizierenden Merkmalen, die nur aus einem oder solchen, die aus mehreren Merkmalen bestehen. Schon Name, Straße und Geburtsdatum bestehen ja aus mehreren Merkmalen; die Werte aus mehreren Merkmalen lassen sich technisch immer auch in einem komplexen Wert „zusammenfassen" – beispielsweise könnte man Name, Adresse und Geburtsdatum auch in ein Attribut einer Datenbank schreiben.[8]
- Wenn in einem Datenbestand mehrere Datensätze zu einem Betroffenen enthalten sind, können die Merkmale aus diesen Datensätzen kombiniert werden. Auch eindeutige Werte aus solchen Kombinationen können als identifizierende Merkmale wirken, beispielsweise Muster von Kommunikationsbeziehungen zwischen Betroffenen oder wenige signifikante Orte in einem Bewegungsprofil.
- Für die Eigenschaft „identifizierendes Merkmal" genügt es auch, dass nur in manchen Datensätzen eindeutige Werte enthalten sind. Die Merkmale anderer Datensätze sind dann mehrfach mit identischen Werten belegt.

Die Beispiele für identifizierende Merkmale in den letzten Absätzen dürften im Kontext personenbezogener Daten wenig umstritten sein. Auf den zweiten Blick stellen die identifizierenden Merkmale den Personenbezug aber mehr oder weniger eindeutig und mehr oder weniger direkt her. Eine Personalnummer beispielsweise ist eindeutig, wenn ich auch das Unternehmen kenne, das sie vergeben hat. Eine Festnetz-Telefonnummer ist für einen Betroffenen als Anschlussinhaber eindeutig, wenn ich weiß, dass es sich um einen Einpersonenhaushalt handelt; sonst „verbergen" sich dahinter mehrere Personen. Mobiltelefonnummern sind häufiger eindeutig als Festnetznummern. Name, Adresse und Geburtsdatum sind innerhalb eines Zeitraums eindeutig. Das gilt im Übrigen auch für Telefonnummern, da sich Adresse und Telefonnummer z. B. durch einen Umzug ändern können. Auch eine dynamische IP-Adresse stellt nur im Zusammenhang mit einem eng begrenzten Zeitraum den Bezug zu einem Betroffenen her. Für viele identifizierende Merkmale benötigt ein Akteur deshalb **Kontextwissen**, um den Betroffenen zu identifizieren. Verfügt er nicht über das notwendige Kontextwissen, kann er auch den Personenbezug nicht herstellen.[9]

Ein Teil des Kontextwissens ist der **Referenzwert** eines identifizierenden Merkmals für einen Betroffenen. Den Referenzwert muss der Anwender kennen; er dient dazu, einen Datensatz zu bestimmen. Z. B. kann eine konkrete Telefonnummer verwendet werden, um einen Namen im ‚inversen Telefonbuch' zu finden. In einem Krankenhaus könnte das Datum „27.04.2004" gemeinsam mit dem Unfallbild „Splitterbruch am Handgelenk" eindeutig sein und dazu dienen, im Informationssystem die richtige Krankenakte aufzurufen. Referenzwerte dienen dazu, den richtigen Datensatz zu einem konkreten Betroffenen im Datenbestand zu finden. Über den Zusammenhang im Datensatz kann dann auf andere, vielleicht ebenfalls identifizierende Merkmale geschlossen werden, z. B. den Namen.

Die Steuernummer kann für Privatleute der Finanzbeamte zuordnen. Für das Finanzamt ist es vermutlich sogar das ‚führende' identifizierende Merkmal. Freiberufler ohne Umsatzsteuerpflicht müssen ihre Steuernummer aber auch in ihren Geschäftsbriefen angeben – zumindest Geschäftspartner könnten damit auch eine Zuordnungstabelle aufbauen.

Ähnliches gilt für die IBAN: Die Zuordnungsmöglichkeit zum Betroffenen besteht für die Mitarbeiter der Bank. Sie besteht aber auch für alle Beteiligten, die durch Zahlungstransaktionen die IBAN erfahren haben.

Ob der Betroffene letztlich über Name und Adresse bestimmt wird, oder ob ein anderes identifizierendes Merkmal führend ist, wie die E-Mail-Adresse, der Alias im *Social Network* oder die Personalnummer in einem Unternehmen, hängt vom jeweiligen Kontext ab.[10] Verschiedene identifizierende Merkmale sind in einer verantwortlichen Stelle häufig miteinander verknüpft; durch diese Beziehungen ist der Betroffene bestimmbar. So kann eine Krankenkasse die Versicherungsnummer häufig auch über Namen und Geburtsdatum bestimmen. Und der Finanzbeamte kann eine Steuerzahlung anhand von Betrag und Zahlungsdatum einem Steuerkonto zuordnen. Andere identifizierende Merkmale müssen auch nicht im gleichen Datensatz gespeichert werden wie das führende identifizierende Merkmal. Es genügt, dass geeignete Beziehungen hergestellt werden können.[11]

1.2 Informationen gewinnen

Das BDSG gibt in § 3 Abs. 6 vor: „*Anonymisieren ist das Verändern personenbezogener Daten derart, dass die Einzelangaben über persönliche oder sachliche Verhältnisse nicht mehr oder nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft einer bestimmten oder bestimmbaren natür-*

5 Entsprechende Begriffe verwenden beispielsweise auch [Swee2002] oder [A29G2014].

6 Entsprechend auch die juristische Interpretation z. B. bei [Roßn2013] mwN.

7 [Swee2002], [A29G2014] und andere unterscheiden in der Diskussion von Anonymitätsmaßen „Identifikatoren" und „Quasi-Identifikatoren". Erstere meinen in der Regel „offensichtliche" identifizierende Merkmale, die in jedem Datensatz eindeutig sind, letztere sind identifizierende Merkmale, die nur in einer Teilmenge der Datensätze eindeutig sein müssen. In diesem Beitrag ist diese Unterscheidung nicht notwendig. Um Überscheidungen mit den etablierten Begriffen zu vermeiden, verwenden wir „identifizierende Merkmale".

8 In der Praxis trennt und speichert man die Merkmale zwar aus technischen Gründen in verschiedenen Attributen, aber aus dem Blickwinkel der Informationstheorie verliert oder gewinnt man damit keine Informationen.

9 [RoSc2000] sprechen in diesem Zusammenhang auch von einem relativen Personenbezug.

10 So auch Dammann in [Simitis2014], § 3 Rn. 22.

11 Viele interessante Beispiele für identifizierende Merkmale finden sich z. B. in [Buchm2015], [A29G2014].

lichen Person zugeordnet werden können."[12] Aus anonymisierten Daten sollen also keine Einzelangaben zu Betroffenen mehr abgeleitet werden können. Die Diskussion um faktische, also nur mit unverhältnismäßigem Aufwand, zu überwindende und absolute Anonymität wollen wir hier zurückstellen.[13] Die im Folgenden beschriebenen Grundprobleme bleiben dieselben; lediglich die geforderte Qualität für ein konkretes Verfahren wird geändert.

Um zu entscheiden, ob ein Anonymisierungsverfahren diese Anforderung gut erfüllt, wird in der Literatur häufig ein Angreifer bemüht, der aus anonymisierten Daten Informationen gewinnen will. Wir wollen in der Diskussion lieber einen **Neugierigen** auftreten lassen. Der Neugierige kann sich gegebenenfalls anstrengen und große Mittel aufwenden, dann ist er vergleichbar mit einem Angreifer. Aber in manchen Fällen muss er das vielleicht gar nicht; neugierig ist ja manchmal auch der Nachbar auf der anderen Straßenseite. Der Neugierige will über einen Betroffenen weitere **Informationen gewinnen**. In unserer Terminologie bedeutet das, dass er ein identifizierendes Merkmal verwendet, um einen Datensatz zu identifizieren, und daraus die Werte zu weiteren Merkmalen ableitet. Der Neugierige kann dazu auf alle Datenbestände aus seinem Kontextwissen zurückgreifen. Er kann sie mit den Ergebnisdaten eines Anonymisierungslaufs zusammenführen, beispielsweise wenn gleiche identifizierende Merkmale in beiden Beständen enthalten sind.[14]

2 Anonymisieren

Die personenbezogenen Rohdaten werden durch einen Anonymisierungslauf in **Ergebnisdaten** umgewandelt. Ein Neugieriger sollte aus den Ergebnisdaten keine Informationen über einen Betroffenen mehr gewinnen können.

Ein Beispiel: Nehmen wir an, ein Auftraggeber möchte mit anonymisierten Daten die Korrelationen zwischen Jahresgehalt, Wohnregion und Krankheitstagen berechnen lassen. Im Ergebnisdatenbestand nach dem Anonymisierungslauf sind in jedem Datensatz enthalten: das Jahresnettogehalt, die auf vier Stellen gekürzten Postleitzahl und die Summe der Krankheitstage.

Tabelle 1 | Beispielhafte Ergebnisdaten eines Anonymisierungslaufs

PLZ-Bereich	Jahresgehalt	Krankheitstage
...		
7435x	29.560	4
7435x	18.250	4
7435x	124.940	2
7762x	29.560	4
7762x	48.430	8
7763x	80.300	5
7858x	29.560	4
8288x		17
...		

In den Ergebnisdaten der Tabelle 1 kommen einzelne Werte oder Wertekombinationen genau einmal vor, oft gilt dies für Extremwerte. Im Beispiel oben könnten dies die Jahresgehälter der Geschäftsführer und des Lagerarbeiters sein. In unserem Beispiel könnte auch eine Gehaltssumme zwanzigmal vorkommen (hier € 29560), aber in drei Postleitbereichen je genau einmal. Für Neugierige bilden diese eindeutigen Werte identifizierende Merkmale. Wären die Ergebnisdaten zugänglich, könnten Neugierige mit dem entsprechenden Kontextwissen einzelne Datensätze den Betroffenen zuordnen, z. B.:

- In einer Bank könnten über die Firma als Einzahler die Jahresgehaltssummen berechnet und die Krankheitstage abgeleitet werden.
- Der Nachbar kennt den Arbeitgeber und den Postleitbereich. Stehen nur wenige Datensätze zu Wahl, kann er die vermutliche Höhe des Gehalts schätzen.

2.1 Ein kleiner Rest Betroffener?

Das Beispiel deutet an, dass mit einfachen Anonymisierungsstrategien nicht sichergestellt wird, dass alle identifizierenden Merkmale hinreichend verändert oder gelöscht werden. Natürlich könnte man jetzt die Frage stellen, ob in den Ergebnisdaten nicht doch ein verschwindend kleiner Rest von solchen identifizierenden Wertkombinationen enthalten sein darf. Die Grundrechte gelten aber für alle Betroffenen gleichermaßen – also sind sie auch gleichermaßen zu schützen.

2.2 Chancen für Neugierige

Die Technikentwicklung der vergangenen Jahre hat die Ausgangslage für Neugierige gegenüber der ‚klassischen' Datenverarbeitung um Größenordnungen verbessert:

- Es erheben nicht mehr nur wenige Stellen in begrenztem Umfang Daten. Datenerhebung ist Jedermann-Kultur. Die großen Akteure des Internets, allen voran Konzerne wie Google, Apple, Microsoft, Amazon, Facebook und die Geheimdienste sammeln unbegrenzt Daten und stellen Dritten weitere Plattformen zum Datensammeln zur Verfügung.
- Viele dieser Informationen sind öffentlich zugänglich, beispielsweise in sozialen Netzwerken.
- Die Technik zur Auswertung strukturierter und unstrukturierter Informationen ist wesentlich weiterentwickelt. In der Form von Suchmaschinen und Cloud Services steht diese Technik auch jedem kostenlos oder kostengünstig zur Verfügung. Organisationen mit den entsprechenden Ressourcen setzen eigene Big-Data-Anwendungen ein.

Diese Situation bedeutet, dass nahezu beliebiges Kontextwissen zur Verfügung steht und kaum abzuschätzen ist, wie sich das Kontextwissen für Neugierige entwickeln wird. Unter den beschrieben Umständen fällt es schwer, „unverhältnismäßigen Aufwand" abzugrenzen, wie er in § 3 Abs. 6 BDSG für die De-Anonymisierung gefordert ist.[15]

Können in geschlossenen Umgebungen möglicherweise optimistischere Annahmen getroffen werden, beispielsweise, dass sich berechtigte Benutzer wohlverhalten und kein unzulässiges

12 Zu anderen Definitionen von Anonymisierung siehr z. B. [Karg2015] mwN. oder [A29G2014] mwN.

13 [Karg2015] diskutiert diese Fragen in diesem Heft.

14 [Swee2002] beschreibt dies am Beispiel von amerikanischen Wählerlisten und anonymisierten medizinischen Daten.

15 Aus rechtlicher Perspektive diskutiert [Karg2015] die Faktoren für Aufwand und problematisiert auch Grenzziehungen für ‚unverhältnismäßigen Aufwand'.

Kontextwissen einsetzen? Karg argumentiert, dass die verantwortliche Stelle bei ihrer Abwägung alle Mittel berücksichtigen muss – auch externe –, die sie nach allgemeinem Ermessen aller Voraussicht nach einsetzen könnte.[16] Eine schwächere Veränderung der Daten wäre danach keine Anonymisierung – die Daten unterlägen weiter dem Datenschutzrecht.

Die Artikel 29-Arbeitsgruppe diskutiert allerdings den Fall, dass ein vermeintlich anonymer Ergebnisdatenbestand veröffentlicht wird, der für einen Dritten doch personenbezogen ist.[17] Sie erwartet dann, dass der Dritte wieder die datenschutzrechtlichen Regelungen anwendet. Da aber bei der Veröffentlichung von Daten völlig unklar ist, wer als Dritter mit den Daten weiterarbeitet, ist auch unklar, welchen rechtlichen Rahmenbedingungen er unterliegt; Wohlverhalten der Neugierigen kann jedenfalls nicht angenommen werden. Gerade für Daten, die zur Veröffentlichung bestimmt sind, müssen hohe Qualitätsanforderungen an die Anonymisierung gestellt werden. Der Zielkonflikt zwischen „gut anonymisiert" und „möglichst hoher Informationsgehalt" wird dabei regelmäßig schwer zu lösen sein.

2.3 Ausgangspunkte für gute Anonymisierung

Heute steht das gesamte Internet jedem als Datenquelle zur Verfügung. Es lässt sich nicht vorhersagen, welche Daten ein Neugieriger verwendet, um die Referenzwerte für eine Re-Identifikation zu finden, welches Kontextwissen er hat und für welche weiteren Merkmale eines Betroffenen er sich interessiert. Es wird in vielen Datenbeständen mehr oder weniger sensitive Merkmale geben und solche, die für eine Auswertung mehr oder weniger interessant sind. Diese Unterscheidungen sind aber keine Kriterien dafür, welche Merkmale zu anonymisieren sind oder nicht. Wenn ein Merkmal eindeutige Werte enthält, kann es als identifizierendes Merkmal verwendet werden.

In gut anonymisierten Ergebnisbeständen müssen deshalb alle identifizierenden Merkmale hinreichend gut umgewandelt oder entfernt werden. Einen Überblick zu mathematischen Ansätzen, die die Qualität von Anonymisierung messen und verbessern können, gibt Dr. Erik Buchmann in diesem Heft.[18] Andere Verfahren zielen darauf, die Einzelangaben zu Betroffenen zu verbergen. Dann kann zwar möglicherweise noch der Betroffene identifiziert werden, aber die gewonnene Information ist in einem definierten Maß unscharf. Buchmann stellt auch solche Methoden des *Information Hiding* vor.

Eine wichtige Voraussetzung, um aus einem Datenbestand Informationen über einen Betroffenen gewinnen zu können, ist, dass der Neugierige entscheiden kann, ob Datensätze eines Betroffenen überhaupt im Datenbestand enthalten sind.[19] Nur dann kann er Referenzwerte überhaupt sinnvoll einsetzen. Anonymisierung im Bereich der empirischen Sozialforschung besteht deshalb sowohl darin, identifizierende Merkmale zu vermeiden, als auch Stichproben aus Datenbeständen so zu wählen, dass die Gruppe der Personen, von denen die Daten erhoben wurden, nicht mehr bestimmt werden kann. Oliver Watteler und Dr. Katharina Kinder-Kurlanda stellen, ebenfalls in diesem Heft, Verfahren und Kriterien für die Anonymisierung aus diesem Anwendungsbereich vor.[20]

Eine spannende Frage ist, inwieweit in Ergebnisdaten Aussagen über Gruppen ableitbar sein dürfen. Nach einem Anonymisierungslauf könnten mehrere identische Datensätze vorliegen; Dann gibt es in dieser speziellen Gruppe, einer sogenannten Äquivalenzklasse[21], kein einmaliges, also identifizierendes Merkmal für einen einzelnen Datensatz.

Für den Neugierigen ist dies keine Hürde: Er kann gegebenenfalls seine Informationen aus den einheitlichen Werten der Gruppe gewinnen. In dem Ausschnitt von Tabelle 1 kann der Neugierige alleine aus dem Jahresgehalt darauf schließen, dass jeder Betroffene in der Äquivalenzklasse „7xxxx, 29560" vier Tage krankgeschrieben war. Der Neugierige schließt in diesen Fällen aus den Gruppendaten und nicht aus den Daten einer einzelnen bestimmbaren Person. Seine Ergebnisse bezieht er aber durchaus auf eine bestimmte Person. Liegen in diesem Fall personenbezogene Daten vor? Wie groß muss dann eine Gruppe sein, die eine Äquivalenzklasse bildet, damit keine personenbezogenen Daten mehr gegeben sind?[22]

Letztlich entsprechen auch statistischen Aussagen der Verteilung von Äquivalenzklassen, z. B. in Form einer Aussage wie „alle Teilnehmer der Umfrage hatten die folgende Eigenschaft ...". Der Übergang zwischen Anforderungen an Anonymisierung und an statistische Auswertungen scheint fließend.[23] Deshalb sollten auch im Fall von freiwilligen Erhebungen die Teilnehmer über die Auswertungen und die Verwendung der Auswertungsergebnisse informiert werden – auch wenn die geplanten Statistiken nicht mehr als personenbezogene Daten einzuordnen sind.

Die folgenden Fragen können helfen, ein Anonymisierungskonzept festzulegen:[24]

- Können die Datensätze im Ergebnisbestand auf eine konkrete Gruppe von Betroffenen bezogen werden?
- Können identifizierende Merkmale im Ergebnisbestand ausgeschlossen werden? Werden auch aussagekräftige Äquivalenzklassen vermieden?
- Ist die Gruppe der Neugierigen begrenzt? Kann angenommen werden, dass sie geringere Mittel einsetzen können, als wenn die Daten veröffentlicht würden?[25]
- Welches Kontextwissen und welche Referenzwerte können für einen Neugierigen verfügbar sein?
- Kann der Neugierige Zuordnungsregeln zwischen scheinbar unabhängigen Datenbeständen erschließen und daraus Informationen gewinnen?

Schließlich muss auch der wirtschaftliche Aufwand herangezogen werden, um das Deanonymisierungsrisiko zu beurteilen, beispielsweise in Form von Kosten, Technikausstattung, Expertenwissen oder Zeitaufwand.[26] Die meisten Faktoren verschieben aber nur den Maßstab für die faktische Anonymität. Insofern

16 [Karg2015]

17 [A29G2014], 11.

18 [Buchm2015]. Die Artikel 29 Datenschutzgruppe stellt in [A29G2014] wichtige Anonymisierungstechniken vor und bewertet sie.

19 Wenn sich die Zugehörigkeit des Betroffenen aus dem Datenbestand direkt ergibt, z. B. über den Namen und die Anschrift, ist das offensichtlich. In anderen Fällen ist es aber nicht offensichtlich. Dann muss es Teil des notwendigen Kontextwissens sein.

20 [WaKK2015]

21 Siehe auch [Buchm2015].

22 Hinweise zu Gruppendaten gibt Dammann in [Simitis2014], § 3 Rn. 14 und 19.

23 Die Autovervollständigung bei Suchmaschinen oder statistische Score-Werte ordnet Dammann in [Simitis2014], § 3 Rn.56, als personenbezogene Daten ein.

24 Siehe dazu auch [Art 29 DSG 2014].

25 Solche Überlegungen dürften den „Secure Data Center" zugrunde liegen, siehe [WaKK2015].

26 Siehe dazu z. B. [RoSc2000], 723f, [Karg2015], [A29G2014] 10.

kann faktische Anonymisierung auch aus verschiedenen Gründen schwach werden, beispielsweise weil anderes Kontextwissen verfügbar ist, andere Neugierige zugreifen können, die Analyseverfahren besser oder die Technik billiger geworden ist, oder deanonymisierte Informationen wertvoller geworden sind. Anonymisierungsprojekte müssen deshalb mögliche Entwicklungen in Kontextwissen, Technik oder dem Wert von Informationen berücksichtigen.[27]

Eine Schlussfolgerung müssen verantwortliche Stellen aus dem Stand der Technik ziehen: Eine Anonymisierung, die diesen Namen verdienen soll, muss sehr gut geplant und geprüft werden.[28] In vielen Fällen kann sie zu einem erheblichen Informationsverlust im Ergebnisbestand führen. Wenn Ergebnisdaten veröffentlicht werden sollen, sind der Kreis der Neugierigen, das ihnen zur Verfügung stehende Kontextwissen und die technischen Mittel kaum begrenzbar. Daher dürften sich die Anforderungen an die faktische Anonymität solcher Daten sehr an die der absoluten Anonymität annähern.[29]

3 Pseudonyme

Im Kontext von Anonymisierung fällt auch immer wieder das Stichwort Pseudonyme.[30] § 3 Abs. 6a BDSG definiert: *„Pseudonymisieren ist das Ersetzen des Namens und anderer Identifikationsmerkmale durch ein Kennzeichen zu dem Zweck, die Bestimmung des Betroffenen auszuschließen oder wesentlich zu erschweren.“* In den Rohdaten werden die Werte von identifizierenden Merkmalen durch geänderte Werte – die Pseudonyme – ersetzt. Dieses „Ersetzen“ erfolgt durch eine Zuordnungsregel zwischen Pseudonym und Betroffenem.

Pseudonyme können danach unterschieden werden, wer sie wählt – danach gibt es drei Konstellationen:[31] In der ersten Variante kann der Betroffene selbst ein Pseudonym wählen und bleibt alleiniger Inhaber des Zuordnungswissens. Beispiele sind pseudonyme Accounts. In der zweiten Konstellation kennt ein Treuhänder den Zusammenhang zwischen einem Pseudonym und dem Betroffenen. Der Treuhänder stellt die Dienstleistung zur Verfügung, das Pseudonym unter definierten Regeln aufzudecken. Schließlich kann die verantwortliche Stelle selbst identifizierende Merkmale in ihrem Datenbestand so umwandeln, dass Pseudonyme vorliegen, die sie bei Bedarf selbst aufdecken kann.

Pseudonyme benötigt man für zwei Zwecke: Entweder möchte man neu entstandene Datensätze der Gruppe bereits vorhandener Datensätze einer Person zuordnen und verwendet dazu die Zuordnungsregel, oder die Zuordnungsregel wird genutzt, um pseudonymen Datensätzen die „real“ identifizierenden Merkmale des Betroffenen zuzuordnen und gegebenenfalls auch offenzulegen. Die Zuordnungsregel ist also eine zentrale Eigenschaft, denn genau dadurch werden die geänderten Werte zu Pseudonymen.[32]

Verfügt eine Stelle selbst über die Zuordnungsregel oder kann sie diese von einem Treuhänder erhalten, sind pseudonyme Datensätze für diese Stelle offensichtlich personenbezogene Daten.[33] Die verantwortliche Stelle benötigt dann z. B. Regeln dafür, wann sie die Zuordnungsregeln aufdeckt und wie sie das Zuordnungswissen schützt.

Die Pseudonyme müssen eindeutige Werte sein, sonst können sie ihre Funktion nicht erfüllen. Im Sinne unserer oben getroffenen Definition sind sie damit für die verantwortliche Stelle ebenfalls ein identifizierendes Merkmal. Die eigentlichen identifizierenden Merkmale, also die Werte, deren Referenzwerte auch Unberechtigten bekannt sind, werden aber durch die Pseudonyme ersetzt und verborgen. Dadurch sollen Neugierige die Datensätze zu Betroffenen nur wesentlich erschwert bestimmen können. Das BDSG sagt zwar nicht, für wen ausgeschlossen werden soll, dass er den Betroffenen bestimmen kann. Dem Sinn der Vorschrift entspricht aber, dass mittels der Zuordnungsregel zwischen Berechtigten – die die Zuordnungsregel kennen dürfen – und Unberechtigten unterschieden werden soll.

Für das Pseudonym und den Datenbestand ergeben sich aus technischer Sicht mehrere Anforderungen:

- Das Pseudonym selbst darf für Unberechtigte keine Rückschlüsse auf den Betroffenen erlauben. Beispielsweise würden Initiale statt des vollen Namens nicht zu einer ausreichenden Pseudonymisierung führen.
- Die Zuordnungsregeln zwischen den Pseudonymen und den Ursprungswerten müssen vor dem Zugriff Unberechtigter geschützt werden. Das bedeutet: Wenn die Pseudonyme berechnet werden, beispielsweise mit einer Hash-Funktion, dürfen Neugierige keine Möglichkeit haben, sie nachzuberechnen. Eine gute Lösung kann eine Tabelle sein, in der den Ursprungswerten Zufallswerte zugeordnet werden.[34] Dann muss aber diese Tabelle vor dem Zugriff Unberechtigter geschützt werden.
- Die Vorschrift des BDSG unterscheidet nicht zwischen verschiedenen identifizierenden Merkmalen, die zu ersetzen sind. Im Sinne unserer Begriffsbildung wären daher im allgemeinen Fall alle identifizierenden Merkmale zu berücksichtigen. Andernfalls könnte ein Neugieriger die Zuordnungsregeln der Datensätze zu den Betroffenen rekonstruieren – der Neugierige kann nämlich die gleichen Ansätze wählen wie für vermeintlich anonymisierte Daten.

In vielen Fällen sollen Daten in einem pseudonymen Bestand allerdings ähnlich differenziert wie Rohdaten verarbeitet werden; das Pseudonym soll „nur“ gegen eine leichte Zuordnung von genauen Angaben zu einem konkreten Betroffenen schützen. Würden die Daten verändert, wie dies für eine gute Anonymisierung der Fall wäre, entstünde regelmäßig ein erheblicher Informationsverlust. Eine Verarbeitung wie mit den Rohdaten wäre in der Regel nicht mehr möglich.[35] Die Artikel 29-Datenschutzgrup-

27 So z. B. auch [A29G2014], 10; [Karg2015]; oder auch schon [RoSc2000].

28 Siehe dazu auch [Art 29 DSG 2014], 28 ff. Unklar bleibt allerdings, wie eine nachträgliche Kontrolle der veränderter Restrisiken für veröffentlichte Daten erfolgen soll.

29 [RoSc2000], 724, argumentieren, dass ein Restrisiko für die Deanonymisierung grundsätzlich nicht auszuschließen und deshalb nur eine Wahrscheinlichkeitsgrenze zwischen faktischer und absoluter Anonymität möglich sei.

30 Zur juristischen Einordnung siehe [Karg2015]

31 Die drei Fälle unterscheiden z. B. auch [RoSc2000], 725; Scholz in [Simitis2014] § 3 Rn. 220 ff; [Plath2013] § 3 Rn. 63 ff; Buchner in [TaGa2013] § 3 Rn. 48 ff.

32 Das gilt im Grunde auch für selbstgewählte Pseudonyme: Wollte der Betroffene nicht noch einmal unter dem Pseudonym agieren, müsste er gar keines wählen.

33 Siehe auch [Karg2015], [Art 29 DSG 2014], 3, 24 ff. oder [RoSc2000], 724f.; Scholz in [Simitis2014] § 3 Rn. 220c.

34 Dammann in [Simitis2014], § 3 Rn. 221. Verschiedene Möglichkeiten, Pseudonyme zu bestimmen, bewertet [Art 29 DSG 2014], 24 ff.

35 Deshalb wird es auch in den meisten Fällen für einen Treuhänder oder eine verantwortliche Stelle nicht sinnvoll sein, die Zuordnungsregeln zu löschen. Denn entweder könnte die Zuordnung von Datensätzen zu Betroffenen mit an-

pe geht von vorne herein davon aus, dass „in der Regel nur ein einzigartiges Merkmal" durch die Pseudonyme ersetzt wird und der Betroffene aus dem Datenbestand weiter identifiziert werden kann.[36]

Wenn es Unberechtigten (mit vertretbarem Aufwand) möglich ist, die Zuordnungsregeln abzuleiten oder zu rekonstruieren, wird die Schutzfunktion der Pseudonyme unterlaufen. Die Bestimmung des Betroffenen ist nicht mehr mindestens „wesentlich erschwert". Ähnlich wie verratene Geheimnisse keine Geheimnisse mehr sind, sind auch aufgedeckte Pseudonyme keine Pseudonyme mehr. Durch Unberechtigte erschlossene (oder zu leicht erschließbare) Pseudonyme sind nur noch ein Alias.

Datenverarbeitung mit Pseudonymen wird man deshalb häufig eher als eine Verarbeitung von Daten mit – vielleicht starken – Aliassen bewerten müssen. Die Bestände mit Pseudonymen unterliegen aber generell dem Datenschutz. Die verantwortliche Stelle kann die Pseudonyme als eine Schutzmaßnahme einsetzen und wird sie gegebenenfalls durch weitere Maßnahmen ergänzen müssen. Anders als für anonyme Daten – die nicht den Erfordernissen des Datenschutzrechts genügen müssen – könnten dann vielleicht auch Annahmen zum Wohlverhalten der Mitarbeiter getroffen werden. Werden die Daten bei Dritten verarbeitet, wird deshalb regelmäßig eine Bindung des Dienstleisters durch ADV-Verträge oder entsprechende Vereinbarungen bei Übermittlung in Drittstaaten erforderlich sein.

Telefonnummern, Personalnummern, IP-Adressen und viele andere identifizierende Merkmale sind nach unserem Verständnis regelmäßig keine Pseudonyme. Die Zuordnungsregeln solcher identifizierenden Merkmale werden nicht besonders geschützt. Selbst wenn die Zuordnungsregeln nicht jedem zugänglich sind, können Neugierige Ausschnitte dieser Regeln teilweise selbst erheben. So können beispielsweise Betreiber von Social-Network-Plattformen oder Werbenetzwerken IP-Adressen speichern und teilweise auch selbst Benutzern zuordnen.[37] Die IP-Adressen können dann wiederum dazu dienen, andere Merkmale zu erheben.

Im Zeitverlauf können Pseudonyme schwächer oder zu einem Alias werden, weil der Datenbestand angereichert wird und Neugierige neue identifizierende Merkmale in den Datensätzen finden. Starke Pseudonyme zu wählen und aufrecht zu erhalten, ist schwierig:

- Der Betroffene muss bei *selbstgewählten Pseudonymen* darauf achten, dass er keine identifizierenden Merkmale erzeugt, die auf seine echte Identität schließen lassen. Das ist umso schwieriger, je länger ein Pseudonym verwendet wird und je mehr Daten sich in seinem Kontext ansammeln.[38] In verschiedenen sozialen Netzwerken können sich z. B. ähnliche Kommunikationsbeziehungen herausbilden. Der Betroffene ist aber in vielen Fällen selbst dafür verantwortlich, sein Pseudonym zu schützen. Wäre die verantwortliche Stelle verpflichtet, den Betroffenen zu unterstützen? Wie kann sie das leisten, ohne selbst nach identifizierenden Merkmalen zu suchen? Außerdem sind die einzelnen Merkmale häufig über viele Stellen verteilt.
- Beispiele für Pseudonyme, die bei einem Treuhänder verwaltet werden, sind pseudonyme De-Mail-Adressen (§ 5 Abs. 2 De-Mail-G) oder Zertifikate qualifizierter elektronischer Signaturen nach § 5 Abs. 3 SigG. Solche Pseudonyme sollten vom Treuhänder nur nach definierten Regeln aufgedeckt werden, beispielsweise nach Rechtsverstößen. Leider ist oft nicht geregelt, unter welchen Umständen oder nach welchem Verfahren die Treuhandstelle das Pseudonym aufdecken darf. Der Inhaber des Pseudonyms muss in diesen Verwendungskontexten außerdem darauf achten, dass er an den Einsatzstellen des Pseudonyms keine identifizierenden Merkmale hinterlässt.
- Für Pseudonyme, die durch die verantwortliche Stelle selbst verwaltet werden, muss diese Stelle dafür Sorge tragen, dass das Pseudonym nicht aufgedeckt werden kann. Es dürfen keine Informationen im Datenbestand enthalten sein, die eine (einfache) Re-Identifikation erlauben: Will man „absolute" Pseudonyme erreichen, gelten für die Werteverteilung der anderen Merkmale im Datenbestand letztlich die gleichen Anforderungen wie für anonyme Daten.

Auch Pseudonyme werfen damit spannende Fragen auf. Für selbstgewählte Pseudonyme wäre z. B. zu fragen, ab wann die zugehörigen Datensätze für eine Stelle personenbezogen sind. Solange nur der Betroffene selbst die Zuordnung vornehmen kann, sind die Einzelangaben durch die Stelle ja eben gerade nicht einer Person zuzuordnen. Von großer praktischer Relevanz ist die Frage nach den rechtlichen Wirkungen der Pseudonymisierung. Diese Aspekte werden in diesem Heft ausführlicher in den Beiträgen von Dr. Moritz Karg[39] und Michael Knopp[40] diskutiert.

4 Verschlüsselte Daten

Verschlüsselung ist seit vielen Jahren eine im Datenschutzrecht empfohlene Schutzmaßnahme: Bei guter Verschlüsselung kann ein Unberechtigter zwar die Chiffrate lesen, aber die eigentlichen Inhalte nicht erschließen.

Verschlüsselte personenbezogene Daten sind für die Inhaber des Schlüssels weiter personenbezogene Daten. Wie aber sind verschlüsselte Daten für die Stellen einzuordnen, die nicht über den Schlüssel verfügen? Dazu müssen wir zunächst zwei Fälle unterscheiden:

- Die Werte zu Merkmalen oder Datensätze werden einzeln verschlüsselt. Wenn gleiche Werte oder gleiche Datensätze auf gleiche Chiffrate abgebildet werden, kann der Neugierige unter Umständen mit statistischen Analysen Schlussfolgerungen ziehen und damit Einzelangaben und identifizierende Merkmale rekonstruieren. Das Verschlüsselungsverfahren muss dies verhindern.
- Wenn der Datenbestand als Ganzes gut verschlüsselt wird, sind statistische Analysen nicht mehr möglich.

Im Unterschied zu Pseudonymen sind bei Verschlüsselung alle Werte geschützt, also auch weitere identifizierende Merkmale im Datenbestand. Bei guter Verschlüsselung können Neugierige deshalb keinerlei Referenzwerte auf den Datenbestand abbilden. Aus ihrer Perspektive lassen sich daher weder Betroffene identifizieren noch Einzelangaben zu ihnen ableiten. Aus der Perspektive von Dritten liegen dann keine personenbezogenen Daten vor.

deren identifizierenden Merkmalen rekonstruiert werden oder die beiden oben beschrieben Funktionen der Pseudonyme werden nicht mehr erreicht.

36 [A29G2014], 24 ff.

37 In [Hamdy2015] kann man erkennen, wie schwierig es mit IPv6 trotz Privacy Extensions werden wird, Datenschutzanforderungen zu realisieren.

38 [LiSc2011] zeigen in einem Analyse-Beispiel, wie schwer dies ist.

39 [Karg2015]

40 [Knopp2015a].

Voraussetzung für diese Einordnung ist aber die Stärke des Schutzes:

- die Stärke des Verschlüsselungsverfahrens,
- die Wahl guter Schlüssel durch die verantwortliche Stelle,
- der Schutz des Schlüsselmaterials so, dass ausschließlich die verantwortliche Stelle Zugriff darauf hat und auch nur die verantwortliche Stelle ihn verwenden kann,
- und die Eigenschaft der Chiffrate, keinerlei, auch keine statistischen Informationen, über Betroffene preiszugeben.

Wenn Daten nur ausgelagert, aber nicht durch den Diensteanbieter weiter verarbeitet werden sollen, könnten sie so verschlüsselt werden, dass Dritte keine Informationen ableiten können. Dies wäre ein einfaches Einsatzszenario für *Cloud Computing*. Ob die genannten Anforderungen in Lösungen für Office 365 erfüllt werden, untersuchen Christoph Schäfer und Kai Jendrian in ihrem Beitrag.[41] Das Speichern von Daten alleine ist aber nur selten ausreichend – in vielen Fällen wird Verarbeitung gewünscht. Prof. Jörn Müller-Quade, Dr. Matthias Huber und Tobias Nilges stellen in diesem Heft die homomorphe Verschlüsselung und das Konzept verschlüsselter Datenbanken vor, bei denen Dienstleister verschlüsselte Daten erhalten und diese verschlüsselt verarbeiten.[42] In zwei Beiträgen schließlich diskutieren Dr. Roland Steidle gemeinsam mit Dr. Ulrich Pordesch[43] sowie Michael Knopp[44] das Pro und Contra von datenschutzrechtlichen Erleichterungen für verschlüsselte Auftragsdatenverarbeitung.

5 Fazit

Der Kontext der heutigen Informationsverarbeitung und ihre perspektivische Entwicklung stellen hohe Anforderungen an Anonymisierungsverfahren. Hinreichende Anonymisierung wird nur erreicht, wenn sie von der verantwortlichen Stelle gut geplant wird. Ob Spielräume für die Güte von Anonymisierungsverfahren bestehen, wird juristisch diskutiert – je abgegrenzter die Gruppe der Zugriffsberechtigten ist, desto eher könnten Kriterien für faktische Anonymisierung ausreichen.

Interessanterweise müssen solche Kriterien auch für Pseudonyme angewandt werden: Im Datenbestand werden regelmäßig weitere identifizierende Merkmale enthalten sein. Auch für Neugierige können solche Datenbestände deshalb personenbeziehbar sein. Datenschutzrechtlich tragfähige Konzepte für die Verwendung von pseudonymen Daten werden deshalb in der Regel das Wohlverhalten der Zugriffsberechtigten sicherstellen müssen, die die Zuordnungsregeln für die Pseudonyme nicht kennen dürfen.

Anders kann man gut verschlüsselte Daten einordnen: Unberechtigte können aus den Chiffraten keine Angaben ableiten, sie haben keinen Zugriff auf die personenbezogenen Daten.

41 [ScJe2015]
42 [MQHN2015]
43 [StPo2015]
44 [Knopp2015b]

Dank

Wertvolle Hinweise zu diesem Beitrag gaben uns Dr. Erik Buchmann, Max-Planck-Institut für Informatik, und Dirk Fox, Secorvo.

Literatur

[A29G2014] Artikel 29 Datenschutzgruppe (2014): *Stellungnahme 5/2014 zu Anonymisierungstechniken (0829/14/DE WP216)*, http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_de.pdf

[Buchm2015] Buchmann, E.: *Wie kann man Privatheit messen? – Privatheitsmaße aus der Wissenschaft*, in diesem Heft.

[Hamdy2015] Hamdy, S. (2015): *Analyse der Forderungen der Datenschutzbeauftragten zu IPv6*, in: Tagungsband zum 14. Deutschen IT-Sicherheitskongress, Gau-Algesheim, 2015, 479 ff.

[Karg2015] Karg, M.: *Anonymität, Pseudonyme und Personenbezug revisited?*, in diesem Heft.

[Knopp2015a] Knopp, M.: *Pseudonym – Grauzone zwischen Anonymisierung und Personenbezug*, in diesem Heft.

[Knopp2015b] Knopp, M.: *Muss die Wirkung von Verschlüsselung neu gedacht werden?*, in diesem Heft.

[LiSc2011] Lindemann, M., Schneider, J.: *Datenschutz-Fallrückzieher*, c't 1/2011, 108 ff.

[MQHN2015] Müller-Quade, J.; Huber, M., Nilges, T.: *Daten verschlüsselt speichern und verarbeiten in der Cloud*, in diesem Heft.

[Plath2013] Plath, *BDSG*, 1. Aufl. 2013

[RoSc2000] Roßnagel, Scholz: *Datenschutz durch Anonymität und Pseudonymität*, MMR 2000, 721.

[Roßn2013] Roßnagel, A.: *Big Data – Small Privacy?*, ZD, 11/2013, 62 ff.

[ScJe2015] Schäfer, C., Jendrian, K.: *Krypto = Allheilmittel der Cloud?*, in diesem Heft

[Simitis2014] Simitis, S. (Hrsg.): *Bundesdatenschutzgesetz*, 8. Aufl. 2014

[StPo2015] Steidle, R., Pordesch, U.: *Entfernen des Personenbezugs mittels Verschlüsselung durch Cloudnutzer*, in diesem Heft.

[Swee2002] Sweeney, L.: *k-anonymity: a model for protecting privacy*. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

[TaGa2013] Taeger, J., Gabel, D.: *Kommentar zum BDSG*, 2. Aufl. 2013.

[WaKK2015] Watteler, O., Kinder-Kurlanda, K.: *Anonymisierung und sicherer Umgang mit Forschungsdaten in der empirischen Sozialforschung*, in diesem Heft.