Kompetenzzentrum
Öffentliche IT

FORSCHUNG FÜR DEN DIGITALEN STAAT

ANONYMISIERUNG: SCHUTZZIELE UND TECHNIKEN

Jan Dennis Gumz, Mike Weber, Christian Welzel

Gefördert durch:

Bundesministerium des Innern, für Bau und Heimat

Fraunhofer FOKUS

IMPRESSUM

Autoren:
Jan Dennis Gumz, Mike Weber, Christian Welzel

Gestaltung:
Reiko Kammer

Herausgeber:
Kompetenzzentrum Öffentliche IT
Fraunhofer-Institut für Offene Kommunikationssysteme FOKUS
Kaiserin-Augusta-Allee 31, 10589 Berlin
Telefon: +49-30-3463-7173
Telefax: +49-30-3463-99-7173
info@oeffentliche-it.de
www.oeffentliche-it.de
www.fokus.fraunhofer.de

ISBN: 978-3-9819921-2-0

1. Auflage Juni 2019

Bildnachweise:

Seite	Autor	Quelle	Lizenz
1, 6, 15, 19	Stefan W	flickr	CC BY 2.0

VORWORT

In der digitalen Welt hinterlassen wir Nutzer eine umfangreiche Datenspur. Natürlich werden unsere Daten für die ganz konkrete Abwicklung einzelner Vorgänge benötigt, beispielsweise eines Online-Einkaufs. Aber auch darüber hinaus gibt es Bedarf, personenbeziehbare und damit schützenswerte Daten weiter zu nutzen. Sei es für Auswertungen innerhalb einer Organisation, um so aus vergangenen Geschäftsprozessen zu lernen, oder auch in Form von statistischen Daten, die an Dritte oder externe Stellen wie Statistikämter weitergegeben werden (müssen) – um Wirtschaft und Gesellschaft besser zu verstehen und letztlich bessere Entscheidungen zu treffen. Es gibt also wichtige und legitime Interessen an der Nutzung unserer schützenswerten Daten.

Durch die Anonymisierung von Daten kann eine Brücke zwischen den Datenschutzinteressen der Einzelnen und der Nutzung von Daten geschlagen werden. Das spielt schon heute eine wichtige Rolle, nicht nur bei der Bereitstellung von statistischen Daten, sondern für alle Organisationen, die Daten offen zugänglich machen möchten. Anonymisierung wird damit beispielsweise zu einem Thema für jede Verwaltung, die sich den Prinzipien von Transparenz und Open Data verpflichtet fühlt.

Wie schön wäre es doch, wenn es konkrete und leicht umsetzbare Bestimmungen gäbe, um Anonymität rechtssicher herzustellen. Man denke nur an DSGVO-Ausführungsbestimmungen, die Parameter je nach Schutzbedürftigkeit festlegen würden: Für Einkommensdaten reichte die Anonymitätsstufe 5, für Krankheitsbilder sollte es aber schon eine 7 sein. Ganz so einfach ist es natürlich nicht. Zum einen bieten statistische Anonymisierungsverfahren immer nur einen relativen Schutz vor Re-Identifizierbarkeit, zum anderen bedarf das Datenschutzrecht noch weitergehender Konkretisierungen etwa durch die Rechtsprechung. Bleibt also nur, sich auf den ebenso notwendigen wie steinigen Weg des interdiziplinären Diskurses zu begeben. Und der beginnt schon bei abweichenden Begriffen, wenn im Recht etwa von sensiblen, in der Statisitik von sensitiven Daten gesprochen wird.

Anonymisierung bleibt somit bis auf Weiteres ein sensibles Thema – ein sensitives natürlich auch. Wir wünschen eine inspirierende Reise durch die interdisziplinäre Welt der Anonymisierung, wobei wir in diesem Papier dann doch einen Schwerpunkt auf die technische Perspektive legen.

Ihr Kompetenzzentrum Öffentliche IT

ANONYMISIERUNG KANN DIE BRÜCKE ZWISCHEN DEN DATENSCHUTZINTERESSEN BETROFFENER UND DEN NUTZUNGSWÜNSCHEN DRITTER SCHLAGEN.

INHALTSVERZEICHNIS

1.	**Thesen**	**5**
2.	**Bedeutung der Anonymisierung in der datafizierten Gesellschaft**	**7**
3.	**Merkmalstypen der Anonymisierung und Schutzziele**	**8**
4.	**Anonymisierungstechniken, statistische Lösungsansätze**	**10**
4.1	Anonymisierungstechniken nach Wirkungsweise	10
4.2	Kennzahlen für die Anonymisierung	11
4.3	Freiheitsgrade und Korrelationen	13
4.4	Werkzeuge	13
5.	**Formen der Re-Identifizierung (Schutzzielverletzungen)**	**16**
6.	**Juristische Betrachtungen**	**17**
7.	**Jenseits von Anonymisierung**	**18**
8.	**Handlungsempfehlungen**	**20**
	Glossar	**21**

1. THESEN

Anonymisierung ist ein wichtiges Element des Datenschutzes – aber auch des Schutzes von Geschäfts- und Betriebsgeheimnissen.

Eine Weitergabe nicht anonymisierter personenbeziehbarer Daten oder deren Verwendung außerhalb ihrer Zweckbindung verstößt gegen die Datenschutzgesetze, allen voran die Datenschutzgrundverordnung, sofern nicht deren Erlaubnistatbestände greifen. Anonymisierung ist daher ein wichtiges Instrument, um Daten über ihre Zweckbindung hinaus verarbeiten oder weitergeben zu können. Sie ist ebenso notwendig, wenn es beispielsweise darum geht, Wirtschaftsstatistiken zu veröffentlichen und zugleich Betriebsgeheimnisse zu wahren.

Namen und Adressen auszublenden, reicht bei Weitem nicht aus.

Direkte Identifikatoren – wie Namen oder Kundennummern – wegzulassen, ist selbstverständlich notwendig, aber für eine Anonymisierung nicht ausreichend. Auch aus den verbleibenden Werten und Wertekombinationen könnten unerwünschte Rückschlüsse gezogen werden. Rückschlüsse sind zudem indirekt unter Hinzuziehung von Informationen aus anderen Quellen möglich.

Auch die Zusammenfassung von Daten schützt nicht vor Re-Identifizierung.

Gruppenweise aggregierte, also etwa durch Durchschnittsbildung zusammengefasste Daten sind zwar viel weniger durchlässig für Informationen über bestimmbare Einzelpersonen als – niemals perfekt – anonymisierte Einzelfalldaten, dennoch können auch sie in bestimmten Situationen mehr verraten als gewünscht, bspw. wenn Einzelne mit besonderen Merkmalsausprägungen herausstechen.

Anonymisierung ist auch Manipulation.

Einfache Maßnahmen wie Weglassen von direkten Identifikatoren oder Vergröbern von Daten sind aus sich heraus verständlich und arbeiten »mit offenem Visier«. Wo diese jedoch nicht ausreichen, kommen manipulativere Techniken zum Einsatz, bis hin zur Generierung fiktiver Einzeldaten aus statistischen Modellen, die ihrerseits aus den Ursprungsdaten abgeleitet sind. Durch Gestalten dieser Modelle in der Zwischenstufe kann der Anonymisierende weitreichende Manipulationen vornehmen (ggf. auch, ohne den Datennutzer aufzuklären).

Ein einheitliches Verständnis von Anonymisierung fehlt.

Statistik, Rechtswissenschaften und Ethik pflegen verschiedene Sprachen und Denkweisen. Das Recht spricht z. B. von der »Nicht-Bestimmbarkeit einer Person«, Anonymisierungstechniken hingegen liefern ein gestaltbares Anonymitätsmaß, mit dessen Anwachsen die Bestimmung einer Person lediglich graduell immer unplausibler wird. Solche unterschiedlichen Verständnisse aufeinander abzubilden, ist eine zentrale Interpretationsaufgabe, die womöglich erst im Laufe der Zeit durch Gerichtsurteile einer Lösung nähergebracht werden wird.

Die DSGVO setzt Maßstäbe, lässt aber noch Fragen offen.

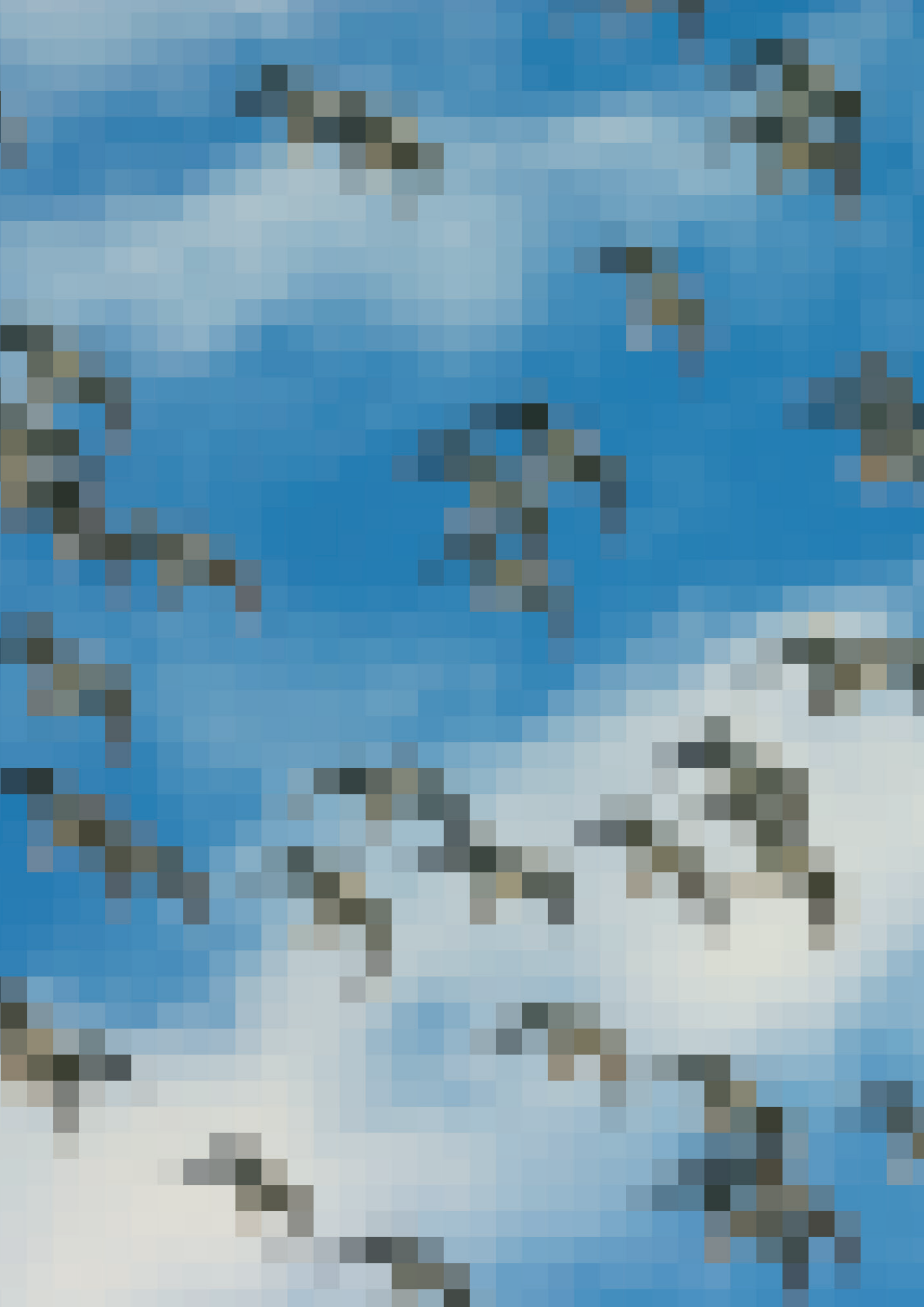
Ordnungsgemäß anonymisierte Daten dürfen auch unter Hinzuziehung externer Zusatzinformationen keine Informationen über bestimmbare Einzelpersonen offenbaren. Trotz Präzisierung in der EU-Datenschutzgrundverordnung (DSGVO) herrscht Unklarheit darüber, welche Zusatzinformationen hier berücksichtigt werden müssen.

Auch anonymisierte Daten können zum Nachteil Einzelner verwendet werden.

Mit nicht-anonymen Daten können die Rechte und Interessen Einzelner unterstützt wie auch beeinträchtigt werden, Auswertungen anonymer Daten können für oder gegen Personengruppen eingesetzt werden. Damit ist lediglich die Zielgenauigkeit geringer. Durch Big Data und Künstliche Intelligenz (KI) werden auf Korrelationen beruhende Gruppenaussagen noch zunehmen.

a. Korrelationen in anonymisierten Daten sind Chancen:
   Big Data und KI ermöglichen nicht nur neue Geschäftsmodelle und Marketingkonzepte, sie erschließen auch neue wissenschaftliche und gesellschaftliche Zusammenhänge und unterstützen so evidenzbasierte politische Entscheidungen.

b. Korrelationen in anonymisierten Daten sind Risiken:
   Beispielsweise das Scoring, bei welchem aus statistischen Daten Schlussfolgerungen auf Individuen und Gruppen gezogen werden, wird ethisch ambivalent beurteilt.



2. BEDEUTUNG DER ANONYMISIERUNG IN DER DATAFIZIERTEN GESELLSCHAFT

Anonymisierung ist ein weit gefasster Begriff. Mal wird darunter das anonyme Surfen im Internet, mal die Verschleierung einer Identität bei Wirtschaftsprozessen verstanden. In diesem White Paper legen wir den Fokus auf den Prozess der Anonymisierung existierender Datenbestände. Anonymisierung wird hier also als Veränderung von personenbeziehbaren oder sensiblen Daten in der Weise verstanden, dass die Bezüge zu einer Person oder Organisation nicht mehr rekonstruierbar sind. Andere Auffassungen des Begriffs Anonymisierung (z. B. »Wann soll man Daten gar nicht erst erheben?« oder »Wie bewege ich mich unerkannt im Netz?«) werden hier nicht behandelt. Auch muss zwischen Anonymisierung und Pseudonymisierung unterschieden werden. Bei einer Pseudonymisierung ist eine kontrollierte Re-Identifizierbarkeit durch die Verwendung unter Verschluss gehaltener Zusatzinformationen möglich. Bei Anonymisierung ist genau dies nicht der Fall.

Anonymisierung wird oftmals fast schon automatisch mit dem Datenschutz und dem Entfernen personenbezogener Merkmale aus Datensätzen gleichgesetzt. Gleichwohl lassen sich darüber hinaus weitere Anwendungsgebiete für Anonymisierung identifizieren. Etwa, wenn es darum geht, Firmeninterna oder gar Betriebsgeheimnisse zu schützen. Auch hierfür können Verfahren der Anonymisierung herangezogen werden. Dieses White Paper nimmt die Anonymisierung personenbezogener Daten in den Fokus, da der Schwerpunkt der öffentlichen und politischen Diskussion sich primär um diesen Aspekt dreht.

In der öffentlichen Debatte spielt die Europäische Datenschutzgrundverordnung (DSGVO) eine wesentliche Rolle, die seit dem 25. Mai 2018 anzuwenden ist. Im Zuge des zunehmenden Einsatzes von Big Data in der Unternehmenswelt wollen die Verantwortlichen Sicherheit haben, nicht gegen die DSGVO zu verstoßen. Auch in der Verwaltung erhalten Anonymisierungstechniken zunehmend Bedeutung. Neben klassischen Aufgaben der Statistik erhält das Thema großes Interesse im Zuge der Open-Data-Bewegung. Gerade bei der Veröffentlichung von Verwaltungsdaten muss sichergestellt werden, dass aus den Daten keine Rückschlüsse auf einzelne Personen oder Organisationen möglich sind. Die Angst vor möglicher Re-Identifizierung wird oftmals als Hindernis für die Veröffentlichung bestimmter Datensätze angeführt.

Daraus ergeben sich typische Herausforderungen, die sich anhand zweier Beispiele vergegenwärtigen lassen:

- Behörde X muss bestimmte Behördendaten anonymisiert veröffentlichen. Dabei will sie sicher sein, weder die DSGVO noch Geschäftsgeheimnisse zu verletzen.
- Firma Y will Kundendaten ohne Einwilligung außerhalb ihrer Zweckbindung nutzen, wobei die Identität der Kunden ohne Belang ist. Die Daten sollen deshalb so anonymisiert werden, dass sie nicht mehr als personenbeziehbar gelten und daher nicht den Einschränkungen der DSGVO unterliegen.

Dieses White Paper nähert sich diesen Herausforderungen aus mehreren Blickwinkeln. Zuerst werden die für die Anonymisierung relevanten Merkmalstypen erläutert und beispielhafte Herausforderungen und formale Schutzziele von Anonymisierung als Teilaufgabe des Datenschutzes vorgestellt. Anschließend werden verschiedene statistische Verfahren zur Anonymisierung untersucht und unterschiedliche Wege von Re-Identifizerung (»Schutzzielverletzungen«) aufgezeigt. Diese Verfahren werden danach zu den juristischen Erfordernissen der Datenschutzgrundverordnung in Beziehung gesetzt. Zum Abschluss werden Handlungsempfehlungen gegeben.

Die streng juristische Sicht steht dabei nicht im Mittelpunkt, es sollen allerdings durchaus praktische Anregungen und Handlungsempfehlungen gegeben werden, die sich aus den Möglichkeiten der Anonymisierung und der aktuellen Rechtslage ergeben. Dabei ist zu berücksichtigen, dass die DSGVO ein junges Gesetz ist. Erfahrungen im Sinne von Gerichtsurteilen mit konkreten Auslegungen und Einzelfällen können in dieses White Paper noch nicht einfließen.

Zur bestmöglichen Verbildlichung von Problemstellungen und Lösungsansätzen wird in diesem White Paper davon ausgegangen, dass die zu anonymisierenden Daten als Tabellen vorliegen: Die Tabellenzeilen entsprechen dabei den einzelnen Individuen, die Spalten ihren verschiedenen Merkmalen. Durch diese Festlegung auf die Anonymisierung tabellarischer Daten – unabhängig davon, ob die Daten über ein Tabellenkalkulationsprogramm oder ein Datenbank-System verwaltet werden – wird auch das Thema dieses White Papers genauer umrissen: Die Anonymisierung etwa von Fließtexten gehört nicht dazu.

3. MERKMALSTYPEN DER ANONYMISIERUNG UND SCHUTZZIELE

Für die Anonymisierung lassen sich die in einer Tabelle hinterlegten Merkmale nach relevanten Typen unterscheiden[1]: Identifikatoren, Quasi-Identifikatoren und sensitive Merkmale.

Bei einem (direkten) Identifikator handelt es sich um ein Merkmal, dessen Ausprägung in der Regel einer Person entweder eindeutig oder nahezu eindeutig zuordenbar ist. Beispielsweise ist die Ausweisnummer des Personalausweises eindeutig einer Person zuordenbar. Die Kombination aus Vor- und Nachname ist zwar nicht immer eindeutig einer Person zuordenbar, allerdings ist die Anzahl der Menschen mit der gleichen Namenskombination in der Regel sehr gering. Dies gilt auch für eine komplette Postadresse. Auch wenn sich unter Umständen die Ausprägung eines Merkmals eindeutig mit einer Person verbinden lässt, folgt daraus nicht zwangsläufig, dass es sich bei dem Merkmal generell um einen Identifikator handelt. So ist Robert Pershing Wadlow die einzige Person mit einer Körpergröße von 2,72 Metern, generell ist es jedoch schwierig, Menschen anhand ihrer Körpergröße zu identifizieren.

Quasi-Identifikatoren (auch: indirekte Identifikatoren) sind Merkmale, die einzeln keine Identifikatoren darstellen, aber kombiniert und unter Verwendung anderweitig legal erhältlicher Daten die Identifikation ermöglichen. Solche Daten enthalten dann die gleiche Kombination von Quasi-Identifikatoren zusammen mit einem Identifikator. Beispielsweise stellen Postleitzahl, Geburtsdatum und Geschlecht gemeinsam Quasi-Identifikatoren dar. So wurde in den USA gezeigt, dass der Abgleich dieser Merkmalskombination mit Listen registrierter Wähler die Zuordnung von medizinischen Daten zu Personen ermöglicht.[2]

Bei einem sensitiven Merkmal handelt es sich um ein Merkmal, dessen Ausprägung keiner Person zuordenbar sein soll, weil ansonsten die Privatsphäre beeinträchtigt wird oder sogar schwerwiegende Folgen zu befürchten sind. Dabei handelt es sich beispielsweise um Gesundheitsdaten und politische Meinungen. Für derartige Merkmale existieren oft auch gesetzliche Regelungen.

Ein (Quasi-)Identifikator kann gleichzeitig auch ein sensitives Merkmal darstellen. Um die theoretischen Grundlagen der Anonymisierung besser vermitteln zu können, wird im Folgenden davon ausgegangen, dass die Mengen der drei beschriebenen Merkmalstypen disjunkt sind.

Sensible Daten müssen geschützt werden. Doch was heißt »schützen« eigentlich? Um diesen abstrakten Begriff zu konkretisieren, werden in der IT-Sicherheit und im Datenschutz sogenannte Schutzziele definiert. Sie ermöglichen es, den Schutz der Daten zu messen und bewerten. Die Schutzziele im Datenschutz auf politischer und ethischer Ebene bestehen in der Wahrung der informationellen Selbstbestimmung und in der Vermeidung der Einschränkung persönlicher Freiheiten etwa durch Diskrimierung. Die juristische Konkretisierung dieser Schutzziele ist, auf einen Punkt gebracht, die Vermeidung der unzulässigen Nutzung und Weitergabe personenbeziehbarer Daten. Die weitere Interpretation der Schutzziele von Anonymisierung führt zu den folgenden statistischen Schutzzielen:

Vermeidung von *Identity Disclosure* (Aufdeckung der Identität):
Eine *Identity Disclosure* liegt vor, wenn ein Datensatz, also eine Zeile in einer Datentabelle, eindeutig einer Person zugeordnet werden kann. Dies kann über direkte Identifikatoren oder auch über Quasi-Identifikatoren erfolgen.

Vermeidung von *Attribute Disclosure* (Aufdeckung von Merkmalen):
Auch wenn es keine eindeutige Zuordnung von Datensätzen zu Personen gibt, können trotzdem Informationen über eine Person offengelegt sein. Beispielsweise sei einer bestimmten Person zwar keine bestimmte Tabellenzeile zuzuordnen, sondern diese Person sei nur auf 50 Zeilen eingrenzbar. Wenn nun aber diese 50 Zeilen z. B. alle dasselbe Geburtsjahr angeben, ist damit das Geburtsjahr der Person offengelegt, auch ohne *Identity Disclosure*. Allgemein formuliert liegt eine Form von *Attribute Disclosure* dann vor, wenn für alle Personen die in Frage kommenden Datensätze in einem Merkmal a.) denselben Wert aufweisen, b.) ähnliche Werte aufweisen oder c.) eine statistische Werteverteilung aufweisen, die von der Gesamtverteilung dieses Merkmals signifikant abweicht.

[1] Merkmale lassen sich auch auf Basis anderer Kriterien unterscheiden, beispielsweise in kategorische Merkmale (z. B. das Geschlecht einer Person) und quantitative Merkmale (z. B. das Alter einer Person).

[2] Latanya Sweeney: »Simple Demographics Often Identify People Uniquely«, 2000. https://dataprivacylab.org/projects/identifiability/paper1.pdf.

Name	Geburtsdatum	Geschlecht	Befund
Hans Meier	03.01.1968	M	Diabetes Typ I
Peter Müller	05.10.1975	M	Rheuma
Jan Schulze	12.06.1987	M	Diabetes Typ II
Erika Mustermann	**30.04.1961**	**W**	**Migräne**
Maximilian Stein	**06.02.1955**	**M**	**Arthrose**
Anna Schmidt	03.04.1991	W	Ohne Befund
Klaus Hoffmann	**22.12.1951**	**M**	**Arthrose**
Birgit Wagner	**08.09.1962**	**W**	**Arthrose**

Abbildung 1: Beispiel für Merkmalstypen. Bei dem Merkmal »Name« handelt es sich um einen Identifikator, bei den Merkmalen »Geburtsdatum« und »Geschlecht« um Quasi-Identifikatoren und bei dem »Befund« um ein sensitives Merkmal.

Im ersten Fall wird der Merkmalswert trotz fehlender Zuordnung eindeutig offengelegt, im zweiten Fall wird er immerhin noch ungefähr offengelegt und im dritten Fall wird seine abweichende Häufigkeitsverteilung offengelegt. Durch diese Aufzählung wird sichtbar, dass Datenschutzverletzungen in verschiedenen Schweregraden vorkommen können, die darüber hinaus interpretationsbedürftig sind (»ähnliche Werte«, »signifikant abweichen«). Nicht nur das Recht bedient sich also unbestimmter Begriffe. Diese Aspekte müssen in eine Risikobewertung einfließen.

Als Beispiel für (statistisches) *Attribute Disclosure* betrachte man die fett gedruckten Zeilen in Abbildung 1: Selbst wenn dort die Namen gelöscht werden, bleibt es zumindest sehr wahrscheinlich, dass der Befund »Arthrose« für ein beliebiges Mitglied dieser anonymen Gruppe zutrifft, da 75 Prozent der Zeilen diesen Befund aufweisen.

Für das weitere Verständnis des Themas ist es wichtig, das Ineinandergreifen der verschiedenen Abstraktionsebenen zu betrachten: Die abstrakten politischen und ethischen Schutzziele werden im Endeffekt durch die beiden *Disclosure*-Vermeidungen realisiert. Bei der Umsetzung dieser *Disclosure*-Vermeidungen sind Restrisiken unvermeidbar und zur Gestaltung dieser Restrisiken wird wiederum eine Interpretation der abstrakten Schutzziele herangezogen.

Zusätzlich kann es je nach Anwendungsfall weitere ethische Schutzziele geben, die sich gegen die Verwendung bereits anonymisierter Daten richten, z. B. gegen die Verwendung der geografischen Dichteverteilung stigmatisierender Merkmale.

4. ANONYMISIERUNGSTECHNIKEN UND STATISTISCHE LÖSUNGSANSÄTZE

Anonymität ist kein binärer Zustand und der Prozess der Anonymisierung führt nicht immer zwangsläufig zu einer vollständigen Anonymität. Der Grad der Anonymität kann jedoch mit unterschiedlichen Verfahren angehoben werden.

Dass eine Tabelle keine Identifikatoren enthält, ist eine Grundvoraussetzung der Anonymisierung. Bei der formalen Anonymisierung, der schwächsten Form der Anonymisierung, werden Spalten mit direkten Identifikatoren daher entfernt. Die Identifikatoren sind für weitergehende inhaltliche Auswertungen in der Regel uninteressant. Allerdings ist die formale Anonymisierung nur stark genug, wenn für die verbleibenden Merkmale und Merkmalskombinationen jeweils noch eine ausreichend große Variation vorkommt. Die formale Anonymisierung gilt daher allgemein als unzureichend und ist z. B. im Bundesstatistikgesetz[3] auf bestimmte Empfängerkreise beschränkt. Um die Re-Identifizierung auf Basis von Quasi-Identifikatoren und die Aufdeckung von Merkmalen zumindest zu erschweren, sind daher weitere Maßnamen erforderlich.

In diesem Abschnitt werden auf der formalen Anonymisierung aufbauende Techniken vorgestellt, die auf tabellarische Daten angewandt werden können. Es gibt viele Möglichkeiten, solche Techniken zu kategorisieren. Hier werden sie entsprechend ihrer Wirkungsweise unterteilt, um dann Kennzahlen für den Grad der Anonymisierung zu betrachten und schließlich auf die Freiheitsgrade bei der Bewahrung von Korrelationen einzugehen. Abschließend wird ein Blick auf die Funktionsweise gängiger Werkzeuge geworfen.

4.1 ANONYMISIERUNGSTECHNIKEN NACH WIRKUNGSWEISE

Kategorie 1: Verringerung der repräsentierten Personen
Die Anzahl der in den Daten repräsentierten Personen lässt sich durch Weglassen von Datenzeilen verringern. Dafür gibt es folgende Varianten:

– Einzelne Zeilen mit Ausreißern oder seltenen Merkmalswerten und -kombinationen werden weggelassen.
– Ein bestimmter Prozentsatz der Zeilen wird weggelassen, sodass man bei einer gegebenen Person nie sicher wissen kann, ob diese in den Daten überhaupt vorkommt.
– Nur ein kleiner Teil der Zeilen wird herangezogen (Stichprobe).

Üblicherweise achtet man bei solchen Maßnahmen darauf, dass die statistische Aussagekraft insgesamt möglichst wenig beeinträchtigt wird, dass also die verbleibenden Daten repräsentativ sind.

Kategorie 2: Veränderung von Merkmalsausprägungen
Die Techniken dieser Kategorie können in Kombination miteinander sowie zusammen mit den Techniken der ersten Kategorie eingesetzt werden. Statt die Menge der in den Daten repräsentierten Personen zu verkleinern, werden die Merkmalsausprägungen, also die Einträge der Spalten, verändert. Dazu existieren verschiedene Möglichkeiten, z. B.:

– Verrauschen (»*noise addition*«) der einzelnen Einträge, d. h. Hinzufügen zufälliger »künstlicher Messfehler«, wobei die statistische Gesamtaussage gewahrt bleibt.
  Abbildung 2b) zeigt ein Beispiel: Das Merkmal »Geburtsdatum« wurde verrauscht durch zufällige Auswahl eines Datums, das nicht mehr als ein Jahr vom tatsächlichen Geburtsdatum abweicht. Rückschlüsse auf die Identität von Patienten werden so erschwert, während Zusammenhänge zwischen Alter und Befund weitgehend erhalten bleiben.
– Vergröbern (»*generalization*«) von Werten durch Wertebereiche (wie z. B. Altersklassen 0 – 17, 18 – 24 statt Altersangabe in Jahren).
  Abbildung 2c) zeigt ein Beispiel: Das Merkmal »Geburtsdatum« wurde vergröbert auf das Jahrzehnt der Geburt. (Zusätzlich wurde das Merkmal »Geschlecht« weggelassen.)
– Mikro-Aggregation, d. h. Zusammenfassen kleiner Gruppen von Datenzeilen und Ersetzen der Einträge durch die Gruppenmittelwerte.
– Zufälliges Vertauschen (»*data swapping*«) der Einträge einer Spalte, während andere Spalten unverändert bleiben. Dadurch werden natürlich auch statistische Zusammenhänge (Korrelationen) zwischen Merkmalen verfälscht. Deshalb existieren Varianten dieser Technik, bei der nur ähnliche Merkmalswerte miteinander vertauscht werden, wodurch die statistischen Zusammenhänge weitgehend erhalten bleiben.

[3] »Gesetz über die Statistik für Bundeszwecke« ((Bundesstatistikgesetz – BStatG) in der Fassung der Bekanntmachung vom 20. Oktober 2016 (BGBl. I S. 2394), das zuletzt durch Artikel 10 Absatz 5 des Gesetzes vom 30. Oktober 2017 (BGBl. I S. 3618) geändert worden ist.

Geburts-datum	Ge-schlecht	Befund
03.01.1968	M	Diabetes Typ I
05.10.1975	M	Rheuma
12.06.1987	M	Diabetes Typ II
30.04.1961	W	Migräne

Geburts-datum	Ge-schlecht	Befund
12.05.1967	M	Diabetes Typ I
23.09.1976	M	Rheuma
19.06.1987	M	Diabetes Typ II
25.01.1962	W	Migräne

Geburts-datum	Befund
1960 – 1969	Diabetes Typ I
1970 – 1979	Rheuma
1980 – 1989	Diabetes Typ II
1960 – 1969	Migräne

Abbildung 2:
a) formales Anonymisieren b) Verrauschen c) Weglassen und Vergröbern

Abbildung 3 (nächste Seite) zeigt ein Beispiel: Die Befunde für Personen gleichen Geschlechts wurden vertauscht. Mögliche Korrelationen zwischen Geschlecht und Krankheit bleiben dadurch erhalten, während Zusammenhänge zwischen Alter und Krankheit verloren gehen.
- Weglassen weiterer Spalten. (Siehe Abbildung 2c) als Beispiel.)

Kategorie 3: Auflösung von Identitäten (Aggregationen)
Die Techniken dieser Kategorie bestehen aus dem, was man landläufig unter »Statistiken« versteht, nämlich zählenden oder summierenden Zusammenfassungen der Ursprungsdaten. Naturgemäß bilden diese Techniken die besten Anonymisierer, da das Individuum in einer Zählung oder Summe aufgeht. Gleichwohl besteht auch hier noch ein Re-Identifizierungs-Restrisiko, namentlich bei

- zu geringer Anzahl der aggregierten Einzeldaten,
- zu geringer Streuung der aggregierten Einzeldaten,
- wenigen dominanten Einzeldaten.

Diesen Risiken wird durch weiteres Vergröbern der Aggregation oder durch Weglassen von Einzelinformationen begegnet.

Kategorie 4: Erzeugung künstlicher Daten
Die bisher vorgestellten Techniken verändern die realen Daten. Stattdessen lassen sich auch teilweise oder vollständig künstliche Daten auf Basis der realen Daten erzeugen. Dazu wird anhand der Ursprungsdaten ein statistisches Modell erstellt. Mit diesem Modell werden dann künstliche Datenbestände generiert. Die statistischen Aussagen der ausgelieferten Daten sind vorgegeben durch das zugrundeliegende Modell, weshalb alle dort nicht repräsentierten Zusammenhänge gewollt oder ungewollt verloren gehen.

Obwohl es sich um künstliche Daten handelt, besteht trotzdem noch ein Restrisiko der Re-Identifizierung. Erzeugt das Modell etwa künstliche Quasi-Identifikatoren zu realen sensitiven Merkmalen, ist ein Angreifer möglicherweise dazu in der Lage, den künstlichen Quasi-Identifikatoren reale Quasi-Identifikatoren zuzuordnen und so schließlich einen Identifikator mit einem sensitiven Merkmal zu verbinden.[4]

Schritte der Anonymisierung
Die vorgestellten Techniken lassen sich einzeln oder in Kombination miteinander anwenden. Sowohl der Anonymisierungsgrad als auch die statistische Aussagekraft hängt von vielen Faktoren ab, so z. B.:

- den gewählten Anonymisierungstechniken,
- der Reihenfolge, in welcher die Techniken angewandt werden,
- dem Ausmaß der Manipulation durch die Techniken (bspw. kleine oder große Wertebereiche beim Vergröbern),
- den Ursprungsdaten (Existenz und Anzahl von Ausreißern …).

Eine beispielshafte Abfolge von Schritten, um Daten zu anonymisieren, könnte sein: 1. formales Anonymisieren, 2. Reduzieren der Datenzeilen, 3. Verändern von Datenzellen. Insgesamt sind die Vorgehensweisen zur Anonymisierung von Datensätzen komplexer als hier dargestellt werden kann. Zudem existieren weitere Techniken, die hier nicht erläutert wurden. Eine ausführliche Darstellung findet sich z. B. in Hundepool et al.[5]

4.2 KENNZAHLEN FÜR DIE ANONYMISIERUNG

Zur Bestimmung des Erfolgs von Anonymisierungsmaßnahmen gibt es eine Reihe von Datenschutzmodellen und zugehörige Kennzahlen. Zu den klassischen zählen *k-Anonymity*, *l-Diversity* und *t-Closeness* sowie die (*n*,*k*)-Dominanzregel bei Datenaggregation.

[4] Josep Domingo-Ferrer et. al.: »Re-Identification and Synthetic Data Generators: A Case Study«; https://pdfs.semanticscholar.org/5bf9/74ebdba5df9928729845068aa5c1f860ca10.pdf.

[5] Anco Hundepool et al.: »Handbook on Statistical Disclosure Control, Version 1.0«, CENEX SDC, 2006.

Geburts-datum	Ge-schlecht	Befund
1980 – 1989	M	Diabetes Typ II
1960 – 1969	W	Migräne
1950 – 1959	M	Arthrose
1990 – 1999	W	Ohne Befund

Geburts-datum	Ge-schlecht	Befund
1980 – 1989	M	Arthrose
1960 – 1969	W	Ohne Befund
1950 – 1959	M	Diabetes Typ II
1990 – 1999	W	Migräne

Abbildung 3: Beispiel für das Vertauschen von Merkmalsausprägungen.

Für eine in einer Tabelle vorkommende Wertekombination der Quasi-Identifikatoren wird die Menge aller Zeilen, die diese Wertekombination aufweisen, als Äquivalenzklasse bezeichnet. Anhand der enthaltenen Wertekombinationen der Quasi-Identifikatoren lässt sich eine Tabelle also in Äquivalenzklassen unterteilen.

Eine Äquivalenzklasse heißt *k-anonym*, wenn sie *k* Zeilen enthält. Die gesamte Tabelle heißt *k-anonym*, wenn jede Äquivalenzklasse mindestens *k* Zeilen enthält. Die Kennzahl *k* stellt also eine Untergrenze für die Anzahl der Personen mit der gleichen Wertekombination bezüglich der Quasi-Identifikatoren dar. Je höher der Wert *k* ist, desto größer sind die Gruppen der gemeinsam betrachteten Personen und umso stärker ist die Anonymisierung.

Ein Beispiel zeigt Abbildung 4a): Es existieren drei Äquivalenzklassen: Die erste Zeile mit der Merkmalskombination »1960 – 1969 und M«, die zweite und dritte Zeile mit »1960 – 1969 und W« sowie die vierte und fünfte Zeile mit »1950 – 1959 und M«. Die Äquivalenzklassen »1950 – 1959 und M« sowie »1960 – 1969 und W« sind *2-anonym*. Weil aber »1960 – 1969 und M« nur einmal vorkommt, ist die Tabelle insgesamt *1-anonym*.

Dass jeder Einzelne stets nur zusammen mit anderen in einer Gruppe landet, nützt allerdings dann nur wenig, wenn alle in dieser Gruppe mit stigmatisierenden Angaben beschriftet sind. Auch wenn kein Rückschluss auf die einzelne Person möglich ist, ergibt sich daraus doch das Vorliegen eines stigmatisierenden Merkmals. Das weiterführende Modell der *l-Diversity* addressiert dieses Problem. Eine Äquivalenzklasse heißt *l-divers*, wenn für jedes sensitive Merkmal mindestens *l* »gut repräsentierte« Ausprägungen in der Klasse enthalten sind. Die gesamte Tabelle heißt *l-divers*, wenn alle Äquivalenzklassen zumindest *l-divers* sind. Wie genau die *l-Diversity* definiert ist, hängt vom Verständnis des Begriffs »gut repräsentiert« ab. Beispielhaft sei hier die »*distinct l-Diversity*« genannt, die besagt, dass eine Äquivalenzklasse *l-divers* ist, wenn zumindest *l* verschiedene Ausprägungen von sensitiven Merkmalen in der Äquivalenzklasse vorkommen.

Mithilfe der *l-Diversity* kann man also eine hohe Anzahl von Merkmalsausprägungen pro Äquivalenzklasse fordern. Sinnvollerweise wird man zusätzlich fordern, dass die verschiedenen Merkmalsausprägungen auch hinsichtlich ihrer Bewertungen eine große Spannbreite aufweisen. Beispielsweise hieße das bei Gesundheitsdaten, dass es in den Gruppen entsprechende Anteile Gesunder und harmlos Erkrankter geben muss.

Ein Beispiel zeigt Abbildung 4a): Weil für die Äquivalenzklasse »1960 – 1969 und W« zwei verschiedene Befunde existieren, ist die Äquivalenzklasse (*distinct*) *2-divers*. Die Äquivalenzklasse »1950 – 1959 und M« ist hingegen nur *1-divers*, weshalb die Tabelle insgesamt *1-divers* ist.

Bei *t-Closeness* wird darüber hinaus die statistische Verteilung von sensitiven Merkmalen berücksichtigt. Die Verteilung der einzelnen Äquivalenklassen soll dabei nicht zu sehr von der Verteilung im Gesamtbestand abweichen. Somit sollen aus der Zugehörigkeit zu einer bestimmten Äquivalenzklasse möglichst keine Rückschlüsse gezogen werden können, die nicht auch schon aus der gesamten Tabelle gezogen werden könnten. Eine Äquivalenzklasse besitzt die *t-Closeness*-Eigenschaft, wenn die Verteilung eines sensitiven Merkmals höchstens den Abstand *t* zur Verteilung des sensitiven Merkmals in der gesamten Tabelle besitzt. Die Tabelle besitzt die *t-Closeness*-Eigenschaft mit der Kennzahl *t*, wenn jede Äquivalenzklasse die *t-Closeness*-Eigenschaft besitzt. Die exakte Definition der *t-Closeness*-Eigenschaft hängt von der Wahl des mathematischen Abstandsmaßes ab, generell gilt jedoch, dass kleinere Werte *t* eine größere Ähnlichkeit der Verteilungen und damit einen höheren Grad der Anonymisierung bedeuten. Wegen seiner abstrakten Formulierung ist *t-Closeness* schwerer verständlich und seine Interpretation kann gleichzeitig sehr ambivalent ausfallen, da die Aussagekraft von Korrelationsanalysen gezielt reduziert wird.

Geburts-datum	Ge-schlecht	Befund
1960 – 1969	M	Diabetes Typ I
1960 – 1969	W	Arthrose
1960 – 1969	W	Migräne
1950 – 1959	M	Arthrose
1950 – 1959	M	Arthrose

Geburts-datum	Ge-schlecht	Befund
1960 – 1969	M	Migräne
1960 – 1969	M	Migräne
1960 – 1969	W	Arthrose
1960 – 1969	W	Migräne
1950 – 1959	M	Arthrose
1950 – 1959	M	Arthrose

Abbildung 4:
a) Beispiel für Äquivalenzklassen und *l-Diversity* b) Beispiel für *t-Closeness*

Für den Grad der Anonymisierung von aggregierten Daten kennt die Statistik eine Reihe von Regeln. Z. B. sagt die (*n*,*k*)-Dominanzregel, dass bei einer angegebenen Summe die größten *n* Beiträge nicht mehr als *k* Prozent derselben ausmachen dürfen. Bei einer Anwendung solcher Regeln müssen normalerweise diese Parameter selbst auch geheim gehalten werden, da andernfalls ungewollte Rückschlüsse möglich würden.

Eine wesentliche gemeinsame Eigenschaft der vorgestellten Kriterien ist ihre Parametrisierung. Das heißt, sie liefern Kennzahlen für den Grad der Anonymisierung. Die Auswahl der Kriterien und ihrer Parametrisierungen steuert die Stärke der Anonymisierung.

4.3 FREIHEITSGRADE UND KORRELATIONEN

Alle vorgestellten Techniken verändern potenziell die Spalteninhalte mehr oder weniger stark, um die Anonymisierung zu verstärken. In der Statistik interessieren aber nicht nur die Daten in den einzelnen Tabellenspalten, sondern insbesondere auch die statistischen Zusammenhänge zwischen diesen, die Korrelationen. Für diese gilt in besonderem Maße, dass nicht so sehr die individuellen Wertekombinationen interessieren, sondern vielmehr deren Verteilung in ihrer Gesamtheit von Interesse ist.

Die vorgestellten Techniken verändern nicht nur die Spalteninhalte, sondern damit naturgemäß eben auch die statistische Verteilung der Merkmalskombinationen. Nun bestehen bei der Veränderung der Spalteninhalte normalerweise viele Freiheitsgrade. Diese Freiheitsgrade kann man wahlweise dazu nutzen, die statistischen Korrelationen möglichst originalgetreu zu erhalten, aber auch dazu, diese zu unterdrücken oder zu manipulieren. Hierzu ist ein eigener Entscheidungsprozess nötig: Korrelationsinformationen können in manchen Fällen zu einer Re-Identifizierung beitragen und dann unerwünscht sein, in anderen Fällen können sie unproblematisch und zugleich interessant sein, sodass man sie gezielt zusätzlich erarbeitet und mitveröffentlicht.

4.4 WERKZEUGE

Es gibt eine große Vielfalt von Werkzeugen zur Anonymisierung, kommerzielle und freie, mit Benutzeroberfläche oder zur Einbindung in Computerprogramme (APIs) und für diverse Datenformate. Viele dieser Werkzeuge bieten Möglichkeiten zur Anonymisierung personenbeziehbarer Daten als Teil des Data Maskings an. Data Masking bezeichnet generell die Verfremdung von Daten. Personenbeziehbare Daten können dabei ersatzweise zur Anonymisierung auch durch Pseudonymisierung verfremdet werden. Es wird zwischen statischem und dynamischem Data Masking unterschieden und infolgedessen auch zwischen statischer und dynamischer Anonymisierung. Bei der statischen Anonymisierung wird eine Kopie der Datensätze erstellt und anonymisiert. Welche Daten tatsächlich relevant sind und welche Daten wie stark anonymisiert werden, wird hier einmalig festgelegt. Die dynamische Anonymisierung wird erst bei der Abfrage von Datensätzen durch Nutzer mit verschiedenen Berechtigungen durchgeführt. Die Abfrage wird dabei unter Einsatz der Originaldatensätze bearbeitet und die Antwort entsprechend den Berechtigungen des Nutzers anonymisiert. Mitunter besteht dabei trotzdem die Möglichkeit der Re-Identifizierung oder der Offenlegung von Merkmalen durch gezielte Abfragen in Kombination mit den sich aus den Abfragen ergebenden anonymisierten Daten.

Im Kontrast zu dieser breiten Abdeckung nicht-funktionaler Kriterien steht eine weitgehende Einförmigkeit bezüglich der Anonymisierung: Die Werkzeuge arbeiten sich durchweg an leicht formalisierbaren Zielsetzungen ab, wie z. B.:

– Formale Anonymisierung (Entfernung direkter Identifikatoren),

MIT DEN RICHTIGEN ANONYMISIERUNGSTECHNIKEN KÖNNEN DATEN VERÖFFENTLICHT UND INDIVIDUELLE DATENSCHUTZINTERESSEN GEWAHRT WERDEN.

- Verrauschen, Vergröbern oder Mikro-Aggregation von Einträgen,
- Erreichen gewisser Kennzahlen für *k-Anonymity*, *l-Diversity*, *t-Closeness*,
- Analyse von Restrisiken der Re-Identifizierung.

Die Unterschiede liegen im Umfang und in der Tiefe der Behandlung dieser Punkte. Somit dürfte zumindest in komplexeren Anwendungsfällen die Abwägung von Sensibilitäten gegen statistische Restrisiken eine manuelle Lösung erfordern.

5. FORMEN DER RE-IDENTIFIZIERUNG (SCHUTZZIELVERLETZUNGEN)

In Kapitel 3 wurden Schutzziele und damit Zielsetzungen zur Anonymisierung formuliert. Das am weitesten gehende Schutzziel ist die Vermeidung von *Identity Disclosure*, also der Aufdeckung von Identifikationsmerkmalen. Diese Zielformulierung soll hier daher als Maßstab zur Untersuchung möglicher Schutzzielverletzungen herangezogen werden.

Es ist offensichtlich, dass in anonymisierten Daten keine direkten Identifikatoren enthalten sein dürfen, denn diese verraten sogar die Tabellenzeile und damit alle Merkmalsausprägungen der identifizierten Person. In diesem Abschnitt geht es darum, welche Möglichkeiten der Aufdeckung von Merkmalswerten darüber hinaus bestehen.

Bei der Ermittlung möglicher Re-Identifizierungen sind Informationen mit zu berücksichtigen, die zusätzlich zu den anonymisierten Daten vorliegen können. Die DSGVO konstatiert, grob zusammengefasst, dass für die Bewertung »alle objektiven Fakten, wie Kosten der Identifizierung und der dafür erforderliche Zeitaufwand, herangezogen werden [sollten], wobei die zum Zeitpunkt der Verarbeitung verfügbare Technologie und technologische Entwicklungen zu berücksichtigen sind.«[6]

Mögliche Hilfsmittel, die Schlussfolgerungen über Merkmalswerte ermöglichen können, sind:

1. die Tabelle selbst: Sie ermöglicht z. B. »rückwärts« rechnende Maßnahmen, wie sukzessives Herausstreichen, das Herausrechnen aus evtl. vorhandenen Randsummen oder das Auswerten dominanter Einzelwerte;
2. öffentliche oder dem Angreifer bekannte Metadaten über die Tabelle, z. B. solche, die die Bedeutung der Tabelleninhalte präzisieren;
3. öffentliche oder dem Angreifer bekannte externe Zusatzinformationen: Hierunter sind insbesondere fremde Daten zu verstehen, die mit den vorliegenden Daten so verknüpft werden können, dass daraus re-identifizierende Information entsteht.

Punkt 1 subsumiert die Probleme, die auf statistischem Wege zu lösen sind; dieser Problemkreis wurde im vorigen Abschnitt behandelt. Die Punkte 2 und 3 sind wesentlich schwieriger, denn zur Beurteilung dieser Angriffspotenziale ist darüber zu befinden,

– welche externen Informationen dem Angreifer zur Verfügung stehen und
– welche De-Anonymisierungs-Potenziale sich daraus ergeben.

Schon die nicht-automatisierte Ausführung dieser Abschätzungen erscheint problematisch. Umso anspruchsvoller ist es, diese Aufgaben durch eine Formalisierung einer automatisierten Bearbeitung zugänglich zu machen.

Darüber hinaus sind weitere Gesichtspunkte der De-Anonymisierbarkeit veröffentlichter anonymisierter Daten zu berücksichtigen. Zum einen ist nicht nur die Weitergabe genauer, sondern auch die ungefährer oder unvollständiger Informationen über natürliche Personen eine Schutzzielverletzung. So verletzt beispielsweise eine Gehaltsangabe in 500-€-Intervallen die informationelle Selbstbestimmung nahezu ebenso wie eine genaue Angabe. Zum anderen gilt es, nicht nur die Ermöglichung gesicherter Schlussfolgerungen zu verhindern, auch die Formulierung plausibler Vermutungen oder auch nur herleitbarer Verdachtsmomente kann eine Schutzzielverletzung sein oder als Zusatzinformation zu einer solchen beitragen. Dieser Punkt spricht ungesicherte Mutmaßungen an, die sich vor dem Hintergrund des Allgemeinwissens als Zusatzinformation zu Hypothesen mit einem gewissen Wahrscheinlichkeitsgrad verdichten können. Die Beurteilung dieses Wahrscheinlichkeitsgrades und der inhaltlichen Sensibilität der Hypothese ergibt, ob in dieser eine Schutzzielverletzung gesehen werden muss.

[6] Siehe DSGVO Erwägungsgrund 26: http://data.europa.eu/eli/reg/2016/679/oj

6. JURISTISCHE BETRACHTUNGEN

Die Betrachtung der Herausforderungen weckt möglicherweise Wünsche nach eindeutigen Kriterien, bei deren Befolgung eine juristisch korrekte Anonymisierung von Daten garantiert wäre. Dies verbietet sich jedoch aus zwei Gründen: Einerseits kann und will dieses White Paper keine vertiefte juristische Analyse leisten. Andererseits enthält die DSGVO keine direkten Kriterien für die Anonymität von Daten[7]. Für die Prüfung von Re-Identifizierbarkeit besteht die wesentliche Unsicherheit, welche und wessen Zusatzinformationen dazu mit herangezogen werden müssen. Hierzu liefert die DSGVO ein aufwandsbezogenes Kriterium, das einer entsprechend abwägenden Analyse bedarf.

Statt der schwer zu beantwortenden Frage »Sind die Daten anonym genug, um aus dem Geltungsbereich der DSGVO herauszufallen?« wird hier die Frage untersucht »Stellt die Veröffentlichung einer Statistik eine zulässige Datenverarbeitung dar?« Wir gehen also der Frage nach, in welchen Fällen trotz bestehendem Personenbezug die DSGVO eine Verarbeitung für statistische Zwecke legitimiert. Dabei wird davon ausgegangen, dass Statistiken potenziell noch Reste von Personenbezug aufweisen, welche durch Änderungen der Konstruktion der Statistiken justierbar sind.

Die Situation in der DSGVO ist folgendermaßen:

- Die DSGVO sieht die Erstellung von Statistiken aus Daten, die ursprünglich für einen legitimen Zweck erhoben wurden, generell als zulässig an (siehe DSGVO Art. 5 I b)[8]).
- Für die Weitergabe bzw. Veröffentlichung personenbeziehbarer Daten verlangt die DSGVO eine Einwilligung der Betroffenen, einen gesetzlichen Auftrag oder ein berechtigtes Interesse.
- Für bestimmte Datenverarbeitungen führt die DSGVO das Instrument der Datenschutz-Folgenabschätzung (auch Datenschutzfolgenanalyse; DSFA) ein und definiert, wann eine solche vorzunehmen und wie auf ihre Ergebnisse zu reagieren ist.

Daraus ergibt sich für die Fragestellung Folgendes:

- Soll die Veröffentlichung erstellter Statistiken ohne Einwilligung der Betroffenen erfolgen, ist ein berechtigtes Interesse notwendig und eine Datenschutz-Folgenabschätzung kann erforderlich sein. (Beim Vorliegen eines gesetzlichen Auftrages oder wirksamer Einwilligungen wären weitere Prüfungen unnötig.)
- Zur Beurteilung der Zulässigkeit der Veröffentlichung werden gemäß Datenschutz-Folgenabschätzung die berechtigten Interessen des Veröffentlichers den Risiken für die Betroffenen und deren schutzwürdige Interessen gegenübergestellt.
- In die Beurteilung dieser Interessen und Risiken fließen das Vorhandensein und die Stärke der Anonymisierung der Daten sowie die Sensibilität der Daten ein.

Für die Datenschutz-Folgenabschätzung lassen sich vier wesentliche Schritte identifizieren:

1. Feststellung der Schutzziele: Hier sind die in Kapitel 3 genannten Schutzziele der Vermeidung der Aufdeckung von Identitäten und Personenmerkmalen zu betrachten.
2. Feststellung und Einordnung der berechtigten Interessen (des Datenverarbeiters).
- Analyse der Risiken einer Verletzung von Schutzzielen, also Einschätzung sowohl der möglichen Folgen wie auch ihrer Eintrittswahrscheinlichkeiten: Die Risiken bestehen in der Verletzung der informationellen Selbstbestimmung mit der entsprechenden Einschränkung persönlicher Freiheiten und den sich daraus ergebenden möglichen wirtschaftlichen Schäden und Gefährdungen. Bei sensiblen Daten bestehen darüber hinaus die Risiken sozialer und politischer Schäden wie Bloßstellung, Stigmatisierung, Ausgrenzung und Repression. Bei der sinngemäßen Anwendung einer Folgenabschätzung für die Daten juristischer Personen bestehen die Risiken in dem Bekanntwerden von Geschäfts- oder Behördengeheimnissen. Die Eintrittswahrscheinlichkeit wird von der Stärke der Anonymität der zu veröffentlichenden Daten wesentlich mitbestimmt.
- Abwägung der gegensätzlichen Interessen. Die Abwägung muss für jeden konkreten Einzelfall neu erfolgen; hier fließen Art und Sensibilität der Daten mit ein.

Die Datenverarbeitung (hier: die Veröffentlichung einer Statistik) ist am Ende zulässig, wenn die berechtigten Interessen daran überwiegen; andernfalls muss die Anonymisierung verstärkt werden oder die Veröffentlichung unterbleiben.

---

[7] Die DSGVO bezieht sich ausschließlich auf personalisierte und pseudonymisierte Daten, nicht auf Daten ohne jeglichen Personenbezug. Siehe dazu Art. 4 I Nrn. 1 u. 5 DSGVO

[8] http://data.europa.eu/eli/reg/2016/679/oj; siehe dazu auch die Erwägungsgründe 162 und 50 der DSGVO.

7. JENSEITS VON ANONYMISIERUNG

Eine wesentliche Motivation zur Anonymisierung ist die Beachtung der informationellen Selbstbestimmung. Anonymisierung schützt insoweit Individualrechte, die Nutzung anonymisierter Daten wird aber nicht weiter beschränkt.

An dieser Stelle soll daher dem Eindruck entgegengewirkt werden, dass mit einer perfekt funktionierenden Anonymisierung alle politischen und ethischen Probleme des Datengebrauchs aus der Welt geschafft wären. Auch mit anonymisierten Daten kann politische, wirtschaftliche und soziale Kontrolle ausgeübt werden. Lediglich die Granularität der Datengrundlage ist gröber; das Individuum kann nur über seine statistischen Merkmale und deren Korrelationen adressiert werden. Gerade diese Einschränkung kann dazu veranlassen, statistische Aussagen ungerechtfertigt auf den Einzelfall zu übertragen: Selbst wenn 99 Prozent der Personen mit der Merkmalswertekombination x_1 bis x_5 auch das Merkmal y aufweisen, bedeutet dies nicht zwangsläufig, dass auch Frau Müller mit der Merkmalswertekombination x_1 bis x_5 das Merkmal y aufweist. Ein solcher Fehlschluss auf Individuen kann dann besonders schwerwiegende Folgen haben, wenn die Entscheidungsprozesse nur schwer überprüfbar sind.

Dies gilt insbesondere für viele Verfahren, die aktuell im Kontext der Künstlichen Intelligenz entwickelt werden. Werden beispielsweise neuronale Netze zur Erkennung oder Zuordnung bestimmter Merkmale eingesetzt, so handelt es sich bei den Ergebnissen immer um einen statistischen Näherungswert. Wann und in welchem Kontext (automatisierte) Entscheidungen auf dieser Grundlage getroffen werden dürfen, erfordert daher eine gewissenhafte Abwägung.

Eine gegenwärtig bereits ausgeübte Praxis dieser Art ist das Scoring. Hierbei werden von privatwirtschaftlicher Seite anhand statistischer Daten zu Personengruppen Rückschlüsse auf die wahrscheinliche Bonität von Einzelpersonen gezogen und daraufhin Kreditvergabeentscheidungen getroffen.

Ähnliche Vorgehensweisen sind auch von hoheitlicher Seite denkbar. So werden beispielsweise in den USA bereits Rückfallwahrscheinlichkeiten von Straftätern bestimmt. Die Gefahr von Diskriminierung und Repression erhöht sich hierbei mit zunehmender Automatisierbarkeit und – paradoxerweise – bei Qualitätssteigerung der zugrundeliegenden Datenbeschaffung und -interpretation: Je umfassender die Datenbasis und je valider die statistischen Modelle, desto größer scheint die Versuchung, Statistiken für Entscheidungen auf individueller Ebene heranzuziehen.

8. HANDLUNGSEMPFEHLUNGEN

Der Blick auf die Möglichkeiten und Grenzen der Anonymisierung hat eine abgestufte Palette von Maßnahmen mit ebenso abgestuften Nebenwirkungen aufgezeigt. Dabei verbleibt ein (gestaltbares) Restrisiko einer Re-Identifizierung, dessen Minimierung eine druchaus komplexe Angelegenheit sein kann. Werden dabei Datenmanipulationen und Korrelationen in den Blick genommen, kann die Anonymisierung sogar eigene Entwurfsüberlegungen erforderlich machen. Je nach Sensibilität der Daten ist eine Anonymisierungstechnik mit hinreichend geringem statistischem Re-Identifizierungs-Risiko einzusetzen. Die nachfolgenden Handlungsempfehlungen können bei der Bewältigung der Herausforderungen Orientierung geben.

Von den Statistikämtern lernen
Statistikämter bearbeiten seit jeher einschlägige Anonymisierungsprobleme. Die Statistiker verfügen daher vielfach über eine Palette bewährter Vorgehensweisen und Kriterien.

Die Unzulänglichkeit formaler Anonymisierung verstehen

Formale Anonymisierung adressiert nur die unmittelbar sichtbaren Risiken einer Re-Identifizierung. Die eigentliche Herausforderung ist, die anderen Risiken, die oft erst aus dem Kontext heraus entstehen, überhaupt zu erkennen.

Anonymisierungsstärke an die Datensensibilität anpassen

Sensible Daten müssen mit erhöhter Sorgfalt behandelt werden, andererseits geht eine stärkere Anonymisierung auch mit einer stärkeren Verfälschung bzw. Vergröberung einher. Um aus den Ergebnissen keine falschen Schlüsse zu ziehen, muss der Datennutzer über Art und Stärke der verwendeten Anonymisierungstechniken aufgeklärt werden, soweit dies die Anonymisierung selbst nicht gefährdet.

Im Zweifel nur Daten fiktiver Identitäten (oder ausreichend starke Aggregationen) veröffentlichen

Man muss damit rechnen, dass personenbezogene Daten, die nach heutigen Maßstäben anonymisiert wurden, durch zukünftige Informationen und Techniken wieder de-anonymisiert werden. Hier sind insbesondere Big Data und die Methoden des maschinellen Lernens zu nennen. Einer Re-Identifizierung kann durch Verfälschung oder Gruppierung von Identitäten vorgebeugt werden.

Entscheidungsgründe dokumentieren

Beim Erstellen eines Anonymisierungskonzepts bewertet die datenverarbeitende Stelle die Berechtigung eigener Interessen und die Schutzbedürftigkeit fremder Daten und nimmt entsprechende Abwägungsentscheidungen vor. Naturgemäß liegt in dieser Konstruktion die Gefahr, dass die erforderliche Neutralität durch eine interessengetrieben eingefärbte Sichtweise aufgeweicht wird. Um dem entgegenzuwirken und die Bewertungs- und Abwägungsvorgänge nachvollziehbar zu machen, ist es hilfreich, eine externe Perspektive einzubinden und alle Entscheidungen nachvollziehbar zu dokumentieren.

Automatisierungswerkzeuge einsetzen

Zumindest für die gut formalisierbaren Teile der Anonymisierung sollten etablierte Werkzeuge eingesetzt werden. Derlei gibt es viele auf dem Markt. Sie reduzieren nicht nur den Aufwand, sondern vermeiden auch unnötige Flüchtigkeitsfehler. Bewertungs- und Abwägungsvorgänge werden allerdings im Wesentlichen beim Anwender verbleiben.

Augenmerk auf Merkmalskopplungen legen

Korrelationen machen viele Statistiken überhaupt erst interessant. Die durch sie übermittelten Informationen können legitime, aber auch »unberechtigte« Interessen bedienen, insbesondere bei einer Re-Identifizierung. Diese Interessenlagen sind im Rahmen einer Risikoabschätzung zu hinterfragen und zu unterscheiden.

Geschäftsgeheimnisse durch Anonymisierung schützen

Auch Firmendaten können »sensibel« sein, z. B. solche zu wirtschaftlichen Verhältnissen oder Geschäftsbeziehungen. Hierbei ist zu beachten, dass der Begriff der »sensiblen Daten« für Firmen anders besetzt ist als für natürliche Personen. Die Re-Identifizierungs-Potenziale sind hier anders gelagert, ebenso die gesetzlichen Grundlagen, da für Firmendaten die DSGVO nicht gilt. Die Verfahren der Anonymisierung können jedoch auch für diese Fälle angewendet werden.

GLOSSAR

Die folgende Auflistung enthält die in diesem Papier verwendeten Fachtermini ergänzt um weitere wichtige Begriffe im Kontext der Anonymisierung.

Anonymisierung:
Bearbeitung vorhandener Daten, sodass daraus keine Informationen mehr über bestimmbare Personen (oder auch Firmen usw.) herausgelesen werden können, auch nicht mit Hilfe von Zusatzinformationen.

Aggregierte Daten:
Daten, bei denen mehrere oder alle Einzelfalldaten durch Summen oder Zählungen zusammengefasst wurden (Gegensatz: Mikrodaten). Aggregation ist naturgemäß der Anonymisierung dienlich.

Äquivalenzklasse:
Für eine in einer Tabelle vorkommende Wertekombination der Quasi-Identifikatoren wird die Menge aller Zeilen, die diese Wertekombination aufweist, als Äquivalenzklasse bezeichnet. Anhand der enthaltenen Wertekombinationen der Quasi-Identifikatoren lässt sich eine Tabelle also in Äquivalenzklassen unterteilen.

Data Masking:
Überbegriff für Techniken zur Verfremdung von Daten, beispielsweise mit dem Ziel der Anonymisierung.

Datenschutz-Folgenabschätzung (auch Datenschutzfolgenanalyse; DSFA):
Vorgang, zu dem ein Datenverarbeiter durch die DSGVO in bestimmten Fällen verpflichtet ist; sie umfasst u. a. Bewertungen der Notwendigkeit, der Verhältnismäßigkeit und der Risiken von Datenverarbeitungsmaßnahmen und Bewertungen der eingesetzten Schutzmaßnahmen für die berechtigten Interessen der Betroffenen.

Datenschutzgrundverordnung (DSGVO):
Seit dem 25.05.2018 gültige EU-Verordnung mit Gesetzesrang in den Mitgliedstaaten:
http://data.europa.eu/eli/reg/2016/679/oj.

Re-Identifizierung:
Nachträgliche Wiederherstellung eines oder mehrerer Personenbezüge in Datensätzen, bspw. durch Hinzunahme von Zusatzinformationen.

Direkte Identifikatoren:
Bei einem (direkten) Identifikator handelt es sich um ein Merkmal, dessen Ausprägung in der Regel einer Person entweder eindeutig oder nahezu eindeutig zuordenbar ist, z. B. Name, Telefon-, Konto-, Kundennummer.

(n,k)-Dominanzregel:
Aggregationsregel für Daten, die vorgibt, dass bei einer angegebenen Summe die größten *n* Beiträge nicht mehr als *k* Prozent der Summe ausmachen dürfen.

Formale Anonymisierung:
Einfache, aber unzulängliche Form der Anonymisierung, bei welcher lediglich die direkten Identifikatoren ausgeblendet, ansonsten aber keine Vorkehrungen gegen eine Re-Identifizierung getroffen werden.

k-Anonymity:
Datenschutzmodell in der Statistik, das die Zusammenfassung von Mikrodaten zu Subgruppen mit gleichen Merkmalskombinationen bewertet. Eine Tabelle heißt *k-anonym*, wenn jede Äquivalenzklasse mindestens *k* Zeilen enthält. Die Kennzahl *k* stellt also eine Untergrenze für die Anzahl der Personen mit der gleichen Wertekombination bezüglich der Quasi-Identifikatoren dar. Je höher der Wert *k* ist, desto größer sind die Gruppen der gemeinsam betrachteten Personen und umso stärker ist die Anonymisierung.

Korrelation:
Korrelationen sind statistische Zusammenhänge zwischen Tabellenspalten, Messgrößen, Wertereihen, Dichtefunktionen, o. Ä.

l-Diversity:
Datenschutzmodell in der Statistik, das die Streuung eines sensitiven Merkmals in Subgruppen berücksichtigt. Das Kriterium *l-Diversity* bestimmt die Anzahl der Merkmalsausprägungen, die mit einem »fairen« Anteil des Gesamtbestandes in den Gruppen repräsentiert sind. Zur Bestimmung des fairen Anteils gibt es verschiedene formale Kriterien.

ANONYMITÄT IST KEIN BINÄRER ZUSTAND. DER GRAD DER ANONYMITÄT KANN JEDOCH MIT UNTERSCHIEDLICHEN VERFAHREN ANGEHOBEN WERDEN.

Mikrodaten:
Einzelfalldaten, auch als Individualdaten bezeichnet (Gegensatz: aggregierte Daten)

Pseudonymisierung:
Unter Pseudonymisierung versteht man das Ersetzen des Namens und anderer Identifikationsmerkmale durch ein Kennzeichen zu dem Zweck, die Bestimmung des Betroffenen auszuschließen oder wesentlich zu erschweren. Eine kontrollierte Re-Identifizierbarkeit unter Verwendung von (unter Verschluss gehaltener) Zusatzinformation ist möglich.

Quasi-Identifikatoren:
Merkmale, deren Kombination durch Abgleich mit zusätzlichen Informationen zu einer Identifikation eines Individuums führen oder dazu beitragen kann; typischerweise z. B. Alter, Geschlecht, Körpergröße

Sensible/sensitive Daten:
Merkmale (wie z. B. Gesundheitsdaten, politische Meinungen u. a.), die bei einem Missbrauch besonders schwerwiegende Folgen haben und für die deshalb besondere Regelungen gelten. Diese Merkmale werden in der DSGVO mit dem Oberbegriff »besondere Kategorien personenbezogener Daten« bezeichnet. Die entsprechenden Regelungen sind in Art. 9 I DSGVO ivm. Art. 4 I Nr. 13-15 DSGVO niedergelegt.

t-Closeness:
Datenschutzmodell in der Statistik, das die Verteilung sensitiver Merkmale in Subgruppen berücksichtigt. Die Verteilung in identifizierbaren Teilmengen, also in den einzelnen Gruppen, soll dabei nicht zu sehr von der Verteilung im Gesamtbestand abweichen. Das Kriterium *t-Closeness* erfasst die Nähe der Verteilungen, wobei kleinere Werte eine größere Ähnlichkeit der Verteilungen und damit einen höheren Grad der Anonymisierung anzeigen.

KONTAKT

Christian Welzel
Kompetenzzentrum Öffentliche IT (ÖFIT)
Tel.: +49 30 3463-7173
Fax: +49 30 3463-99-7173
info@oeffentliche-it.de

Fraunhofer-Institut für
Offene Kommunikationssysteme FOKUS
Kaiserin-Augusta-Allee 31
10589 Berlin

www.fokus.fraunhofer.de
www.oeffentliche-it.de
Twitter: @OeffentlicheIT

ISBN: 978-3-9819921-2-0

kre | 1906 (Fotos: https://www.flickr.com/)