/open

RheinlandPfalz
MINISTERIUM FÜR ARBEIT, SOZIALES, TRANSFORMATION UND DIGITALISIERUNG

HANDREICHUNG ANONYMISIERUNG

Warum Anonymisierung von Daten im Kontext von Open Data?

Die Bereitstellung von Open Data birgt ein hohes Potenzial, gesellschaftlichen Nutzen zu erzeugen – sei es in Forschung, Wirtschaft oder Verwaltung. Offene Daten sind dabei in der Regel nicht personenbezogen. In einzelnen Fällen können Daten auf Individualebene dennoch Bestandteil offener Datensätze sein, beispielsweise wenn der gesellschaftliche Nutzen höher zu bewerten ist als der Personenbezug der Daten. In solchen Ausnahmefällen muss durch effektive Anonymisierung sichergestellt werden, dass die Veröffentlichung der Daten den geltenden datenschutzrechtlichen Bestimmungen entspricht, insbesondere denen der Europäischen Datenschutz-Grundverordnung (DSGVO). Die Anonymisierung der Daten soll sicherstellen, dass diese weiterhin für Analysen und andere Zwecke nützlich bleiben. Gleichzeitig soll sie das Risiko senken oder gar eliminieren, dass Personen anhand der öffentlichen Daten identifiziert werden können.

Grundlagen der Anonymisierung und Abgrenzung zur Pseudonymisierung

Die europäische DSGVO definiert personenbezogene Daten als alle Informationen, die sich auf eine identifizierte oder identifizierbare Person beziehen. Dazu gehören nicht nur offensichtliche Identifikatoren wie Name oder Adresse, sondern auch weniger direkt erkennbare Daten wie IP-Adressen oder spezifische Kombinationen von Merkmalen, die eine Re-Identifizierung ermöglichen könnten. Entsprechend gilt eine Information als anonymisiert, wenn die betroffene Person nicht mehr identifiziert werden kann und dieser Zustand irreversibel ist. Eine Re-Identifizierung darf auch mit zusätzlichen Informationen oder zukünftigen technischen Mitteln nicht möglich sein. Re-Identifizierung ist der entscheidende Unterschied zwischen pseudonymisierten und anonymisierten Daten: Während pseudonymisierte Daten weiterhin personenbezogene Daten darstellen und der DSGVO unterliegen, fallen ordnungsgemäß anonymisierte Daten nicht mehr unter den Anwendungsbereich der DSGVO.

Bei Unsicherheit, ob Daten personenbezogen sind, sollte stets die oder der Datenschutzbeauftragte konsultiert werden.

/open

Pseudonymisierung

Personenbezogene Daten werden durch den Austausch personenbezogener Informationen mit künstlichen Bezeichnern (Pseudonymen) so verändert, dass sie ohne die Verwendung zusätzlicher Informationen (eines sogenannten **Schlüssels**) keiner spezifischen Person mehr zugeordnet werden können. Diese zusätzlichen Informationen werden getrennt aufbewahrt und durch technische und organisatorische Maßnahmen geschützt.

Anonymisierung

Personenbezogene Daten werden durch die vollständige Entfernung von personenbezogenen Informationen aus dem Datensatz anonymisiert oder so verändert, dass ein Rückschluss auf eine Einzelperson nicht mehr möglich ist. Das bedeutet, selbst mit zusätzlichen Informationen kann der Personenbezug nicht wiederhergestellt werden.

Für die Veröffentlichung als Open Data ist in der Regel eine vollständige Anonymisierung erforderlich, um die Einhaltung datenschutzrechtlicher Vorschriften zu gewährleisten.

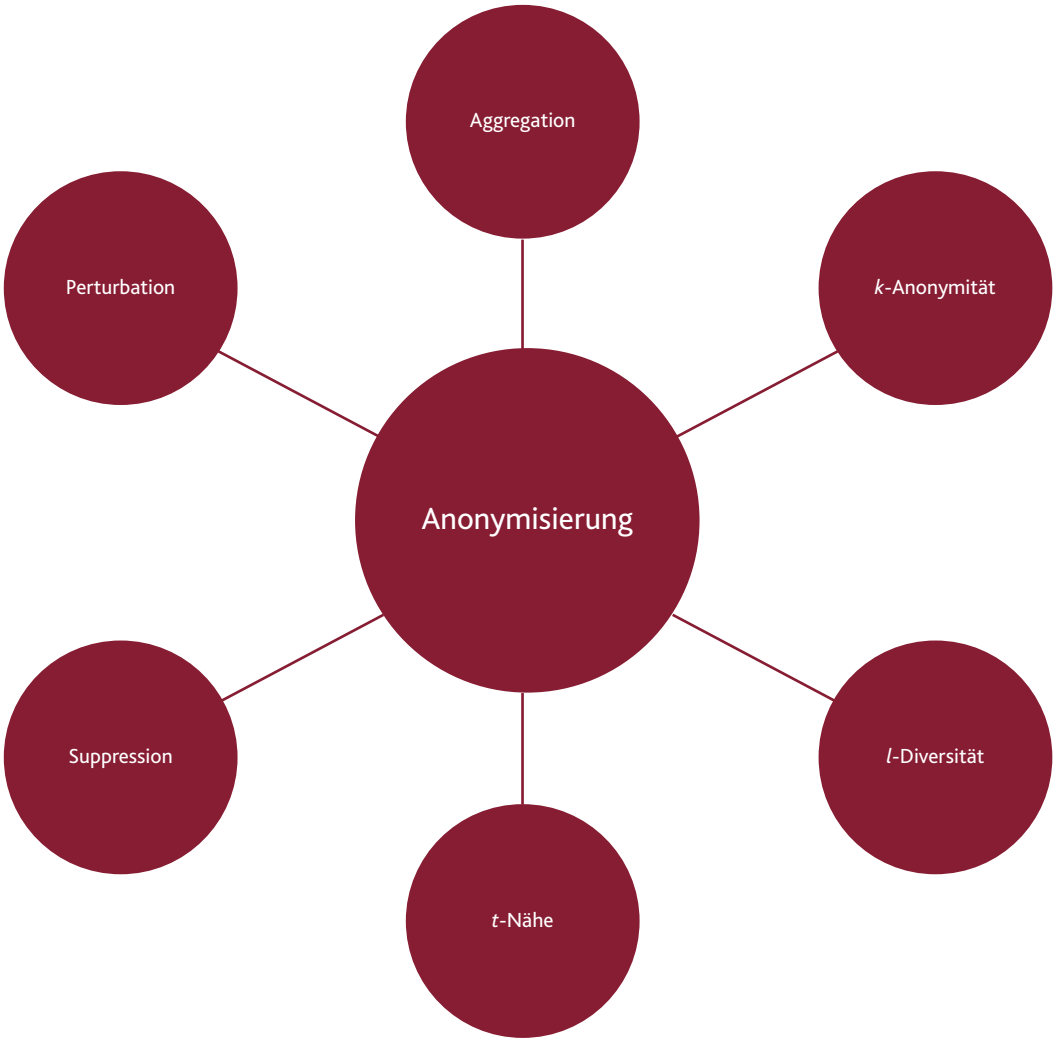
Methoden zur Anonymisierung von Daten

Die Anonymisierung von Daten kann durch verschiedene Verfahren erfolgen, die oft in Kombination angewendet werden. Die Wahl der Methode hängt stark von der Art der Daten und dem gewünschten Nutzbarkeitsgrad des anonymisierten Datensatzes ab.

Bevor Anonymisierungstechniken angewandt werden können, müssen die Daten und die möglichen Identifikatoren genau analysiert und verstanden werden. Hierbei wird insbesondere den **Quasi-Identifikatoren** eine hohe Bedeutung beigemessen. Darunter versteht man Datenpunkte, die allein nicht identifizierend sind, in Kombination mit weiteren Informationen jedoch eine Re-Identifizierung ermöglichen können.

Quasi-Identifikatoren: Ein Beispiel

Der Datensatz enthält Vor- und Nachname, Postleitzahl, Geburtsdatum sowie Vermögensinformationen einer Frau. Die Entfernung der Namen scheint auszureichen, um ihre Re-Identifikation zu verhindern. Jedoch wohnt die Frau auf einem abseits gelegenen Grundstück, in dessen Postleitzahlenbereich insgesamt nur drei Personen, zwei Männer und die Frau, wohnen. Damit lässt sich die Frau aus dem Geschlecht und der Postleitzahl re-identifizieren, auch ohne die offensichtlich personenbezogenen Datenpunkte, wie den Vor- und Nachnamen.



/open

Aggregation

Diese Methode zielt darauf ab, die Granularität von Daten zu reduzieren, um eine Identifizierung zu verhindern, während die Nützlichkeit für die Analyse erhalten bleibt.

Vorgehen: Daten von der Individualebene werden zu Gruppen zusammengefasst oder bestimmte Attribute werden verallgemeinert.

Limitation: Sind die gebildeten Gruppen zu klein und bestehen im ungünstigen Fall aus einer Person, ist weiterhin das Risiko der Re-Identifikation gegeben.

Anwendungsbeispiele:

- Geburtsdatums- oder Altersangaben werden durch Altersspannen (z.B. 20-29 Jahre, 30-39 Jahre) ersetzt.
- Präzise GPS-Koordinaten werden zu größeren Gebieten (z.B. Stadtteilen, Kreisen) aggregiert.
- Straßennamen und Hausnummern werden durch Postleitzahlen ersetzt. Für größere Städte kann auch eine weitere Generalisierung auf Postleitzahlgebiete sinnvoll sein.
- Genaue Zeitstempel werden auf Tage, Wochen oder Monate gerundet.

k-Anonymität

Die *k*-Anonymität erweitert die Methode der Aggregation. Diese Methode stellt sicher, dass jede Kombination von Quasi-Identifikatoren in einem Datensatz für eine Mindestanzahl *k* an Personen vorkommt.

Vorgehen: Im ersten Schritt wird der Wert *k* definiert, d.h. die Mindestgröße einer Gruppe. Im zweiten Schritt werden Quasi-Identifikatoren identifiziert. Im dritten Schritt werden die Datenpunkte aggregiert, sodass die Gruppen mindestens *k* Personen mit den gleichen Werten der Quasi-Identifikatoren umfassen.

Limitation: *k*-Anonymität erschwert die Re-Identifikation, schützt jedoch nicht immer vor ihr. Beispielsweise können durch Social-Engineering-Angriffe Informationen in einer Gruppe gesammelt werden und per Ausschluss-Prinzip auf eine Person in der Gruppe geschlossen werden. Außerdem kann es zum sogenannten Homogenitätsproblem kommen, das auftritt, wenn alle Datensätze innerhalb einer Gruppe ähnlich sensible Attribute aufweisen, beispielsweise alle Personen die gleiche Krankheit haben.

Anwendungsbeispiel: Der Datensatz enthält Informationen über Postleitzahl, Alter und Geschlecht sowie weitere, nicht-personenbezogene, Datenpunkte. Für *k* wird ein Wert von fünf definiert, d.h. die Kombination aus Postleitzahl, Altersgruppe und Geschlecht umfasst mindestens fünf Personen. Im letzten Schritt werden Altersgruppen und Postleitzahl-Gruppen sowie Summen der Geschlechter gebildet, sodass die Daten von mindestens 5 Personen je Gruppe im finalen Datensatz aggregiert sind.

l-Diversität

l-Diversität ist eine Verfeinerung des *k*-Anonymitätsmodells, um die Limitationen bei homogenen sensiblen Attributen innerhalb einer *k*-anonymen Gruppe zu adressieren.

Vorgehen: Zusätzlich zur *k*-Anonymität wird festgelegt, dass für jede *k*-anonyme Gruppe mindestens *l* verschiedene Werte für ein bestimmtes Attribut vorliegen müssen. Das heißt, innerhalb einer *k*-anonymen Gruppen darf ein weiteres Attribut nicht weniger als *l* Ausprägungen haben.

Limitation: Wenn die Anzahl der Werte innerhalb einer Gruppe unterschiedlich ist, jedoch ungleich verteilt, lässt sich in Einzelfällen nicht ausschließen, dass einzelne Personen re-identifiziert werden.

Anwendungsbeispiel: Wenn eine *k*-anonyme Gruppe aus 10 weiblichen Personen besteht, die aus dem gleichen Postleitzahlengebiet (56867) stammen und in einer identischen Altersspanne liegen (30 bis 39 Jahre) und die gleiche Sportverletzung haben (Syndesmosebandriss), dann kann für jede dieser Personen auf die Verletzung geschlossen werden. *l*-Diversität kann hergestellt werden, wenn die Gruppe weiter aggregiert wird, sodass mindestens 2 Sportverletzungen in der *k*-anonymen Gruppe vorkommen.

t-Nähe

T-Nähe geht über *l*-Diversität hinaus und stellt nicht nur Anforderungen an die Anzahl unterschiedlicher sensibler Attribute innerhalb einer *k*-anonymen Gruppe, sondern auch an die Verteilung dieser Attribute innerhalb der Gruppe. Sie stellt sicher, dass die Verteilung sensibler Attribute innerhalb der Gruppen ähnlich verteilt ist wie im Gesamtdatensatz.

Vorgehen: Zusätzlich zur *k*-Anonymität und der *l*-Diversität wird eine Anforderung an die Verteilung der sensiblen Attribute innerhalb der Gruppe definiert und die Aggregation so justiert, dass diese Anforderung erfüllt ist. Für jede der *k*- und *l*-anonymisierten Gruppen wird zusätzlich die darin enthaltene Verteilung sensibler Attribute mit ihrer Verteilung im Gesamtdatensatz verglichen. Der Schwellenwert *t* gibt dabei an, wie groß der Unterschied zwischen den beiden Verteilungen sein darf. Ist *t* in einer Gruppe im Vergleich zum Gesamtdatensatz kleiner oder gleich t, wird die Gruppe als *t*-nah bezeichnet. Je kleiner *t* ist, desto höher ist der Schutz der Privatsphäre.

Limitation: Mit steigenden Anforderungen an die Anonymisierung, beginnend mit *k*-Anonymität bis zur *t*-Nähe, kann der Nutzen der veröffentlichen Daten nicht mehr im Verhältnis zum Aufwand stehen.

Anwendungsbeispiel: Wenn eine *k*-anonyme Gruppe mit *l*-Diversität insgesamt 10 Personen umfasst, wovon 9 Spezialisten sind und eine Person Manager, dann kann über zusätzliches Hintergrundwissen auf den Manager zurückgeschlossen werden. Wenn zusätzlich definiert wird, dass die *k*-anonyme Gruppe im Hinblick auf die Berufsgruppe nicht weniger als 30 % einer Ausprägung (bspw. Manager) umfassen darf, dann wäre *t*-Diversität implementiert.

Methoden der Suppression

Suppression bedeutet, dass bestimmte Informationen aus dem Datensatz entfernt werden.

Vorgehen: Personenbezogene Daten oder Quasi-Identifikatoren werden aus dem Datensatz gänzlich entfernt und der Datensatz wird, wenn möglich, in der Reihenfolge randomisiert, sodass das auch die Beschaffung der gelöschten Datenpunkte nicht mittels Merging-Methoden zur Re-Identifikation führen kann. Alternativ können einzelne Datensätze entfernt werden.

Limitationen: Die Entfernung von ganzen Attributen kann den Nutzen des Datensatzes minimieren oder die Verwendung der Daten gänzlich nutzlos machen. Wenn Randomisierung nicht möglich ist (bspw. bei Zeitreihendaten), dann ist Re-Identifikation durch Datenlecks nach wie vor möglich. Die Entfernung von einzelnen Datensätzen kann den Nutzen durch eine Verzerrung der ursprünglichen Verteilung und eine fehlende Repräsentativität ebenfalls schmälern.

Anwendungsbeispiel: In einem Datensatz mit Geburtsdatum, Name, Vorname, Geschlecht, Postleitzahl und Einkommen werden einzelne Zeilen mit bestimmten Einkommen entfernt. Zusätzlich werden Geburtsdatum, Name, Vorname und Geschlecht entfernt.

Methoden der Perturbation

Während der Datenperturbation werden die Originaldaten mit zufälligen Elementen angereichert. Darunter fallen Rauschaddition (engl. adding noise), Datenrotation (engl. circular permutation) oder die Generierung von synthetischen Daten (engl. synthetic data generation).

Vorgehen: Zu den (numerischen) Daten werden Werte aus einer Zufallsverteilung addiert (Rauschaddition). Werte werden nicht verändert, aber die Reihenfolge einzelner Werte wird innerhalb einer Spalte zufällig verändert (Datenrotation). Die Erstellung synthetischer Daten impliziert die Ableitung der Verteilung der Originaldaten und computergestützte Generierung von neuen Daten mit gleichen Verteilungs- und Zusammenhangsmustern.

Limitation: Bei allen Verfahren werden die Daten zunehmend verzerrt und der erhoffte Nutzen durch die Verwendung der Daten sinkt. Schlussendlich lassen sich die Daten nicht mehr für reale Anwendungen verwenden, sondern lediglich für Simulationszwecke im Forschungskontext. Zusätzlich kann beispielsweise durch Rückschlüsse auf die verwendete Verteilung bei Rauschaddition eine Re-Identifikation nicht ausgeschlossen werden.

Anwendungsbeispiel: Bei Einkommensdaten werden zufällige Werte aus einer Normalverteilung mit Excel generiert und zu den Originaldaten addiert (Rauschaddition). Alternativ wird die Reihenfolge der Einkommen innerhalb einer Spalte händisch oder mithilfe eines Computerprogramms so verändert, dass die Zuordnung zu den anderen Attributen (Geschlecht, Wohnort etc.) nicht mehr der Realität entspricht.

Empfehlungen für die Praxis

Die Anonymisierung von Daten ist ein komplexer Prozess, der Sorgfalt und Fachwissen erfordert. Es gibt keinen Idealprozess, der sich generisch anwenden lässt. Folgende Überlegungen helfen dabei, das richtige Vorgehen zu entwickeln:

- Bauen Sie Datenverständnis auf, um personenbezogene Attribute, insbesondere Quasi-Identifikatoren, zu erkennen.
- Führen Sie eine Risikoabschätzung für eine Re-Identifikation nach jedem Anonymisierungsschritt durch und bauen Sie den Prozess iterativ auf. Fragen Sie sich, ob personenbezogene Informationen weiterhin in den Daten enthalten sind und ob, z.B. mit zusätzlichem Wissen, ein Rückschluss auf Einzelpersonen möglich ist.
- Bewerten Sie im Prozess, ob eine weitere Anonymisierung im Verhältnis zum Nutzenverlust bei der Verwendung der Daten steht.
- Ziehen Sie bei Unsicherheiten Experten wie Datenschutzbeauftragte und Datenfachleute hinzu.
- Dokumentieren Sie den Anonymisierungsprozess, um seine Nachvollziehbarkeit sicherzustellen.
- Wenn Sie Daten aktualisieren, überprüfen Sie das Ergebnis des Anonymisierungsprozesses. Veränderungen in der Verteilung der Daten können die Anonymisierung unwirksam machen.

Die Einhaltung dieser praktischen Schritte wird Ihnen helfen, wertvolle Daten sicher als Open Data bereitzustellen und gleichzeitig die Privatsphäre der Betroffenen zu schützen.

/open

RheinlandPfalz
MINISTERIUM FÜR ARBEIT,
SOZIALES, TRANSFORMATION
UND DIGITALISIERUNG

Herausgeber

Open-Data-Kompetenzzentrum Rheinland-Pfalz im
Ministerium für Arbeit, Soziales, Transformation
und Digitalisierung des Landes Rheinland-Pfalz

Bauhofstr. 9
55116 Mainz

E-Mail: cc-od@open.rlp.de