How To

Anonymisierung

Einführung und Praxistipps zur Anonymisierung personenbezogener Daten

Es gibt Datensätze, die so persönlich sind, dass sie aus gutem Grund hinter verschlossenen Türen bleiben. Doch einige enthalten auch unabhängig vom Personenbezug wertvolle Informationen. Es lohnt sich deshalb auch bei sensiblen Daten eine Veröffentlichung von Open Data in Betracht zu ziehen. Voraussetzung dafür ist eine Verfremdung der Daten, die Rückschlüsse auf Personen verhindert. In diesem How To erklären wir, welche Methoden es dafür gibt – ein kleiner Einstieg zur Anonymisierung von Datensätzen mit persönlichem Touch!

Was sind personenbezogene Daten?

Als personenbezogene Daten sind alle Informationen zu verstehen, die sich auf eine identifizierte oder identifizierbare natürliche Person beziehen. Dazu zählen neben Namen oder Kontakdaten z. B. auch das Alter, Geburtsjahr oder der Wohnort einer Person. Werden diese Informationen durch eine sogenannte Anonymisierung von der Zuordnung zu einzelnen Personen entkoppelt, gelten die Daten nicht mehr als personenbezogen. Sie können dann rein rechtlich der Öffentlichkeit frei zur Verfügung gestellt werden. Es gibt keinen absoluten Maßstab darüber, wann Informationen sich unter keinerlei Umständen wieder Personen zuordnen lassen, also wann Daten vollumfänglich anonymisiert sind. Daher geht es bei der Anonymisierung vielmehr darum, dass der Personenbezug nur mit einem unverhältnismäßig hohen Aufwand an Zeit und Kosten wieder hergestellt werden kann. Jeder zu veröffentlichende Datensatz muss als Einzelfall betrachtet werden, mit dem Ziel eine gute **Balance zwischen** einem hohen **Datenschutzniveau** und einer hohen **Datenqualität** zu finden.

Anonymisieren bedeutet, personenbezogenen Daten so zu verändern, dass betroffene Personen nicht mehr identifizierbar sind.

Das Fallbeispiel

In diesem How To verweisen wir für anschauliche Beispiele auf die Anonymisierung eines realen Datensatzes zu getätigten Ausleihen einer Berliner Bibliothek. Die folgende Tabelle zeigt einen vereinfachten Ausschnitt aus dem originalen Datensatz. Jede Zeile beschreibt einen Ausleihvorgang und enthält personenbezogene Informationen zu den Ausleihenden.

Ausleihvorgang			*Ausleihende Person*			
Medium	Datum	Uhrzeit	Geschlecht	Geburtsjahr	Verkehrszelle	...
Exit Racism	02.01.22	14:01	weiblich	1991	16021	...
Star Wars - Die hohe Republik	02.01.22	14:01	weiblich	1991	16021	...
Öko-Test. - (2022), Heft 10	02.01.22	17:34	weiblich	1956	16021	...
Reggae nights [CD]	03.01.22	10:10	männlich	1999	16021	...

Anonymisierung in der Praxis

Müssen meine Daten anonymisiert werden?

Bei jedem Datensatz muss grundlegend geprüft werden, ob er (in dieser Form) veröffentlicht werden kann oder Formen der Anonymisierung notwendig sind. Diese Risikoeinschätzung sollte von Personen, die den Datensatz gut kennen und/oder Datenschutzexpertise haben und möglichst nach dem Vier-Augen-Prinzip, erfolgen. Dabei sind folgende Fragen hilfreich:

- **Sind personenbezogene Informationen enthalten und ist es wahrscheinlich, dass einzelne Personen identifiziert werden können?** Falls ja, identifizieren Sie, welche Variablen personenbezogen sind, welche sensibel sind und welche besonders wertvoll sind.
- **Sind Variablen in den Daten vorhanden die in Kombination mit zusätzlichem Wissen aus z.B. anderen Datensätzen Rückschlüsse ermöglichen?** Falls ja, sammeln Sie gemeinsam mit Kolleg:innen Szenarien, wie aus dem Datensatz Personen re-identifiziert werden könnten. Überlegen Sie, wie und warum Ihre Daten mit diesen anderen Datensätzen verknüpft werden könnten.

Wenn einer der beiden Punkte zutrifft, muss eine Anonymisisrung vorgenommen werden.

Das Anonymisierungsvorgehen ist immer im konkreten Einzelfall zu entwickeln. Es kommt dabei auch auf das Thema und das Anwendungsfeld an, zum Beispiel gelten Geo- und Gesundheitsdaten als ganz besonders sensibel.

Beispiel: Der Bibliotheksdatensatz enthält keine Namen oder Adressen, durch die sich einzelne Personen identifizieren lassen könnten. Er enthält aber potentiell sensible Informationen, wie die Verkehrszelle (eine Raumeinheit, die meist mehrere Häuserblöcke enthält), also den ungefähren Wohnort der Person. Die zu einer Verkehrszelle gehörenden Häuserblöcke sind im Internet recherchierbar. Durch die Kombination dieser offenen Informationen wird die Menge an Personen, auf den sich eine Ausleihe beziehen kann, deutlich kleiner, es sind aber immer noch keine Einzelpersonen identifizierbar. Wir gehen noch einen Schritt weiter und fragen, ob es möglich ist, dass ich mit zusätzlichem, persönlichem Vorwissen eine Einzelperson einer Ausleihe zuordnen kann. Ein mögliches Szenario: Wenn ich weiß, dass meine Nachbarin am 02.01.22 das Buch “Exit Racism” ausgeliehen hat, und an diesem Tag in dieser Verkehrszelle keine andere Frau etwas ausgeliehen hat, kann ich aus dem Datum und Zeitstempel der Daten auch herausfinden, dass sie ebenfalls das Buch „Star Wars“ ausgeliehen hat. Ich kann aus dem Ausleihvorgang auch sensible Informationen über sie ableiten, wie das Alter. Dieses und ähnliche Szenarien sind eher unwahrscheinlich, aber denkbar. Daher haben wir uns entschieden, Methoden zur Anonymisierung der Daten anzuwenden.

Methoden der Anonymisierung

Für die Anonymisierung bieten sich verschiedene Methoden an, die miteinander kombiniert bzw. iterativ angewendet werden können. Die Wichtigsten stellen wir hier vor.

Aggregieren: Daten zu Kategorien zusammenfassen

Das Aggregieren von Daten bietet sich an, um mehr Überschneidungen zwischen den Einträgen zu erreichen, sie also weniger einmalig zu machen. Welche Variablen wie stark aggregiert werden sollen und welche Granularität

Wie stark zu aggregieren ist, hängt auch von der Menge und Stichprobe der Daten ab und welche anderen, personenbezogenen Variablen der Datensatz enthält. Ist beispielsweise der Geburtsmonat und das Geburtsjahr gemeinsam mit dem Geschlecht und Wohnort zum Beispiel auf Ebene der lebensweltlich orientierten Räume (LOR) gegeben, lässt sich daraus eher eine bestimmte natürliche Person ableiten, als wenn nur das Geburtsjahr bekannt ist.

beibehalten werden kann, ist immer fallspezifisch. Können beispielsweise Geburtsdaten zu Geburtsjahren verallgemeinert werden? Oder ist vielleicht der Geburtsmonat besonders relevant im Kontext des Datensatzes? Sollten besser Altersgruppen gebildet werden? Zu beachten sind dabei auch Informationen, die vielleicht nicht im Datensatz explizit enthalten sind, aber den gesamten Datensatz betreffen, zum Beispiel wenn Daten nur aus einem bestimmten Bezirk stammen, oder nur eine bestimmte Gruppe, z. B. Frauen im öffentlichen Dienst, umfassen. Dadurch wird die Gruppe möglicher Personen wieder kleiner, was die De-Anonymisierung vereinfacht.

- **Räumliche Daten:** Konkrete Adressdaten können auf eine höhere räumliche Ebene aggregiert werden wie zum Beispiel auf Gebäudeebene, Blockebene oder planungsrelevante Einheiten wie LOR-Räume oder politische Einheiten.
- **Altersangaben:** Konkrete Altersangaben wie Alter oder Geburtsdatum können zu Altersgruppen zusammengefasst werden.
- **Zeitangaben:** Konkrete Datumsangaben wie die genaue Uhrzeit können zu übergeordneten Datumsangaben wie (Werk-)Tage, Wochen oder Monate aggregiert werden.

Beispiel: Der Bibliotheksdatensatz enthält sehr genaue Angaben wie Uhrzeit und Datum der Ausleihe und das konkrete Alter einer Person. Dadurch ist jeder Eintrag in der Kombination aller Merkmalen sehr „eindeutig“. Es gibt wenige Ausleihvorgänge, die die gleichen Merkmale besitzen. Je einmaliger jeder Ausleihvorgang, desto wahrscheinlicher wird es, dass ich mit weniger Vorwissen schneller Rückschlüsse auf eine mir bekannte Person ziehen kann. Um die Ausleihvorgänge weniger eindeutig zu machen, wird eine Aggregierung vorgenommen. Es bietet sich z.B. an statt Uhrzeit und Datum nur die Stunde und den Monate zu behalten. Dadurch bleibt die zeitliche Komponente in den Daten enthalten, was weiterhin viele Auswertungen über das Geschehen im Tages- und Jahresverlauf zulässt. Jedoch lassen sich mehrere Einträge zu Ausleihen nicht mehr eindeutig miteinander verknüpfen. Um sicherzugehen, dass die Ausleihen einmalig genug sind, wird als nächstes das Prinzip der k-Anonymität angewandt.

K-Anonymität meint das Verstecken in der Menge: Wenn jede Person Teil einer größeren Gruppe ist, dann kann jeder Datensatz in dieser Gruppe einer einzelnen Person entsprechen. Es muss dafür genügend Überschneidungen zwischen den Einträgen geben und die Dimensionalität der Daten muss ausreichend gering sein.

k-Anonymität: Verstecken in der Menge ermöglichen

Damit k-Anonymität erreicht werden kann, müssen mindestens k Personen im Datensatz vorhanden sein, die den Satz von personenbezogenen Attributen teilen. Gleichzeitig müssen sich die Personen einer solchen Gruppe in anderen, sensitiven Merkmalen zumindest teilweise unterscheiden. Dazu müssen Daten teilweise aggregiert oder gelöscht werden. Für die konkrete Prüfung der k-Anonymität ist die Wahl eines Schwellenwerts, also des k, nötig: Wie viele Einträge muss es mindestens innerhalb einer Gruppe von gleichen personenbezogenen Attributen geben, damit Einträge innerhalb dieser Gruppe nicht mehr einer bestimmten, natürlichen Person zugeordnet werden können? Hierfür gibt es keine konkrete Vorgabe, sondern auch hier gilt es fallspezifisch abzuwägen.

Die ODIS steht bei Fragen unterstützend zur Seite.

Ziehen Sie bei ihrem Anonymisierungsvorhaben am besten die mit dem Datenschutz beauftragte Person ihrer Behörde mit ein.

Beispiel: Wir haben uns nach der Analyse des Bibliotheksdaten für einen Schwellenwert von k=20 entschieden. Es wird für die Kombination aus allen personenbezogenen Spalten (Verkehrszelle, Altersgruppe, Geschlecht) geprüft, wie viele Einträge mit gleichem, kombiniertem Wert es gibt. Gibt es weniger als 20 Einträge, wurde für eine der Spalten der Wert gelöscht. Anschließend findet erneut eine k-Prüfung statt und ggf. wird für eine weitere Spalte ein Wert gelöscht. Dies wird durchgeführt, bis es keine Gruppe von eindeutigen Kombinationen dieser Spalten mit weniger als 20 Einträgen gab, die k-Anonymität also erreicht wurde. Anstatt bei einer Unterschreitung des Schwellenwerts direkt alle personenbezogenen Daten dieser Einträge zu löschen, haben wir uns für dieses schrittweise Vorgehen entschieden, um so viel wie nötig und so wenig wie möglich an Datenqualität einzubüßen.

Löschen: In Einzelfällen notwendig

Sehr fallspezifisch kann zum Schutz der Privatsphäre das weitere Löschen von Werten nötig sein. Dabei ist wieder die Gesamtmenge an Personen, auf die sich ein Datensatz beziehen kann, relevant. Wenn diese Zahl zu klein ist und eine Aggregation sich nicht anbietet, müssen u. U. Werte gelöscht werden.

Beispiel: Im Fall der Bibliotheksdaten, ist der Rückschluss auf Personen über die Verkehrszelle ein kitischer Faktor. Es gibt Verkehrszellen, in denen weniger als 100 Pesonen gemeldet sind. Bei einer so kleinen Gesamtmenge könnte in Kombination mit den anderen, personenbezogenen Angaben im Datensatz hinreichend leicht eine natürliche Person identifiziert werden. Diese Einträge werden mit ‚NA' überschreiben.

Zur weiterführenden Recherche empfehlen wir u. a. folgende Ressourcen:

Stiftung Datenschutz (Hrsg.) (2023): Anonymisierung und Pseudonymisierung von Daten. Gesetzestexte, Leitfaden und Grundregeln für die Praxis. C.E. Müller Verlag.

Anonymisierung prüfen und dokumentieren

Nach der durchgeführten Anonymisierung gilt es, sich nochmal die Fragen der anfänglichen Prüfung zu stellen und die entwickelten Szenarien zur möglichen Re-Identifikation zu prüfen. Werden die Daten anschließend veröffentlicht, sollte die Anonymisierung der Daten und das angewendete Verfahren grob als Teil der Metadaten bzw. der Datenbeschreibung erklärt werden. Es gilt auch nach der Veröffentlichung, den Datenschutz nicht aus den Augen zu verlieren, da der jeweilige Kontext sich stets ändern kann. So bietet es sich zum Beispiel an, bei einer erneuten Veröffentlichung oder Aktualisierung des Datensatzes, das Anonymisierungsverfahren erneut zu reflektieren.

Grenzen von Anonymisierungsverfahren

Bei allen Chancen und Potenzialen einer Anonymisierung, sollte bedacht werden, dass eine absolut unumkehrbare Anonymisierung praktisch unmöglich ist. Die Grenzen der Anonymisierung und Beispiele für Re-Identifikation wurden bereits in einigen wissenschaftlichen und praktischen Arbeiten untersucht. Mit einem qualitativ guten Anonymisierungsverfahren lässt sich die Gefahr der De-Anonymisierung sehr gut minimieren und die Qualität der Daten bewahren.

OPEN DATA INFORMATIONSSTELLE BERLIN

Die Open Data Informationsstelle wird gefördert von der Senatskanzlei und der Investitionsbank Berlin aus den Mitteln des Landes Berlin.

BERLIN
Der Regierende Bürgermeister von Berlin
Senatskanzlei

TECHNOLOGIE STIFTUNG BERLIN