h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbi
FACHBEREICH INFORMATIK

Recherche
Anonymisierung

Inhaltsverzeichnung

1. Hintergrund – Wozu Anonymisieren?

Personenbezogene Daten sind eine schützenswerte Ressource – mehr denn je. Die Digitalisierung erhöht die Zahl der Dienste, für die Nutzer Daten bereitstellen müssen. Datenanbieter möchten ihre sensiblen Daten vertraulich behandeln. Insbesondere beim digitalen Austausch müssen Daten vor Manipulation und Diebstahl geschützt und die Identität des Datenanbieters jederzeit geschützt werden. Gleichzeitig muss sie die hohen Anforderungen der Datenschutzgrundverordnung (DSGVO) erfüllen.

2. Begriffsbestimmungen.

Anonymisierung:

„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 natürlichen Person zugeordnet werden können.“ §3 Abs.6 Bundesdatenschutzgesetz (BDSG)

Personenbezogene Daten:

„Personenbezogene Daten sind Einzelangaben über persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren natürlichen Person (Betroffener).“ §3 Abs.1 BDSG

Personenbeziehbar:

Einzelangaben, die eine bestimmte Person zwar nicht eindeutig oder unmittelbar identifizieren, die es aber erlauben, die Identität der Person mit Hilfe anderer Informationen festzustellen.“ (Metschke & Wellbrock 2002, S. 19)

3. Art der Daten.

Wenn Sie eine Anonymisierung vornehmen möchten, müssen Sie zunächst die direkten und indirekten Identifizierungsmerkmale im Datensatz identifizieren und auswerten, ob dieser im Zuge der Anonymisierung geändert/gelöscht werden muss.

Die zur Verarbeitung vorgesehenen Daten sind also in drei Kategorien einzuteilen:

A- Direkte Identifikationsmerkmale:

Alle Daten, welche eine direkte Identifizierung zulassen. Beispiele für direkte Identifikationsmerkmale sind insbesondere Namen , unter denen die Person bekannt ist. und Orten, Bilder, Stimmen.

B- Indirekte Identifikationsmerkmale:

Alle Daten, welche in Verbindung mit anderen indirekten oder externem Wissen eine Identifikation potentiell ermöglichen. Beispiele für indirekte Identifikationsmerkmale sind:

- Personenbezeichner (z. B. Patienten-ID, Sozialversicherungsnummer, Steuernummer, Autokennzeichen, Kontonummer, Versicherungsnummer, Geburtsdatum)
- Erscheinungsmerkmale (z. B. Körpergröße, Haarfarbe, Kleidung, Tätowierungen)
- Biometrische Kennzeichen (z. B. Gesicht, Stimmprofile, Fingerabdrücke)
- Genetische Daten
- Digitale Zertifikate, welche eine Identifikationsmöglichkeit beinhalten (z. B. Zertifikate zur elektronischen Unterschrift)
- Identifikationsmerkmale basierend auf elektronischer Kommunikation (z B. Telefonnummer, Faxnummer, E-Mailadresse, IP-Adresse)
- Demographische Daten (z. B. Religion, Geburtsland, Muttersprache, Vorstrafen)
- Zuordnungsmerkale (z. B. Beruf, Funktion, Anschriften, Vorstrafen, Name der Mutter/des Vaters)
- Ausreißervariablen (z. B. seltene Diagnosen, Behandlungsbesonderheiten, körperliche Fehlbildungen, für die untersuchte Population untypische Merkmale).

Je nachdem, welche weiteren Informationen dem oder den Verantwortlichen zur Verfügung stehen, sind indirekte Identifikationsmerkmale ggf. als direkte Identifikationsmerkmale anzusehen

C- Nicht identifizierende Daten:

Alle anderen Daten, die weder direkte oder indirekte Identifikationsmerkmale darstellen.

4. Verfahren der Anonymisierung

Es gibt verschiedene Methoden der Anonymisierung, und diese Methoden sollten danach bewertet werden, wie der jeweilige Fall den von der Person verfolgten Zweck am besten erreicht.

4.1 Nichtangabe

Die zu schützenden Daten werden bei dieser Methode nicht verwendet, sondern weggelassen, z.B. indem Sie die Spalte oder den Wert Teilbereich der Tabelle in der Datenbank löschen oder nicht exportieren.

4.2 Maskierung/Ersetzung

Zu schützende Daten werden mit einem konstanten oder sich ändernden Wert, Zeichen oder Zeichenkette ersetzt.

4.3 Mischung/Shuffling

Bei dieser Methode werden die im Datensatz enthaltenen Werte ausgetauscht („scrambled“). Es ist wichtig zu beachten, dass alle Informationen, die eine Person eindeutig identifizieren, wie bspw. eine Telefonnummer oder eine Kreditkartennummer zur Auflösung des Personenbezugs noch zusätzlich mit einer weiteren Methode verfremdet werden müssen, um einen Personenbezug ausschließen zu können.

4.5 Varianz Methode

Verwenden Sie diese Methode, um zahlenbasierte Daten zu verfremden, indem Sie den Wert in einem festen, zufällig ansteigenden oder abnehmenden Varianz Intervall ändern.

4.6 Kryptografische Methoden

Hierbei kommen Verschlüsselungs- und /oder Hash-Algorithmen zum Einsatz. Dabei ist zu beachten, dass kryptografische Eigenschaften wie Blocklänge, Ausgabealphabet und Kollisionen der jeweils verwendeten Methoden Auswirkungen auf das Ergebnis der Anonymisierung haben. Weiterhin ist zu beachten, dass hier kryptografische Methoden im speziellen Kontext der Anonymisierung bzw. Pseudonymisierung betrachtet werden, d. h. einige Betrachtungen im anderen Kontext ggf. zu anderen Ergebnissen führen können

4.7 Verschlüsselungsverfahren

Moderne kryptographische Methoden sind nahezu ausschließlich Binär Chiffren, die sich in Block- und Stromchiffren sowie in symmetrische und asymmetrische Verfahren unterteilen lassen. Hierbei ist Folgendes zu beachten:

1) Stromchiffren müssen zum Erhalt von Eigenschaften mehrfach denselben Schlüsselstrom verwenden, was die kryptografische Stärke der Verfahren abgeschwächt. Daher sind Stromchiffren für die Pseudonymisierung/Anonymisierung i.d.R. eher ungeeignet

2) Den Vorteilen im Umgang mit dem Schlüsselmaterial stehen bei asymmetrischen Verfahren sehr hohe Performance-Einbußen und relativ große Chiffrat-Blöcke entgegen.

Dabei erhalten Binär Chiffren weder den Zeichensatz noch die Zeichenart oder die Zeichenlänge. Jedoch sind Binärchiffren sowohl kollisionsfrei als auch eindeutig.

Andere Verschlüsselungsverfahren können Anforderungen bzgl. Zeichenart, Zeichensatz und Zeichenlänge ggf. erhalten. Dieses ist z. B. bei entsprechender Implementierung beim symmetrischen Verfahren „One-Time-Pad“ der Fall. Hier wiederum kann ggf. die Anforderung der Eindeutigkeit nicht mehr gegeben sein.

4.8 Hash-Funktionen.

Während sich anonymisierte Daten einfach durch das Weglassen der personenbeziehbaren Daten erzeugen lassen gibt es, wie oben erwähnt, für die Erzeugung von Pseudonymen verschiedene Methoden. Für Datenverarbeitungssysteme eignet sich am besten die Hashfunktion.

Hashfunktionen werden in Sicherheitsverfahren zur Unterstützung der Authentikation (Identifikation), der Erkennung der Datenunversehrtheit (Signaturen) oder dem Urheber- und Empfängernachweis verwendet. Eine Hashfunktion ist ein Algorithmus, der eine Nachricht (Bitfolge) beliebiger Länge auf einen Hashwert, eine Nachricht (Bitfolge) fester und kurzer Länge abbildet.

Eine Hashfunktion muss über folgende Eigenschaften verfügen:

• Einwegfunktions-Eigenschaft, d. h. zu einem vorgegebenen Wert soll es mit vertretbarem Aufwand unmöglich sein, eine Nachricht zu finden, die eben diesen Wert als Hashwert hat. Dieser „vertretbare Aufwand” hängt immer vom Entwicklungsstand der einsetzbaren Technik und den jeweiligen Sicherheitsanforderungen ab.

• Kollisionsfreiheit, d. h. es soll mit vertretbarem Aufwand unmöglich sein, zwei Nachrichten mit demselben Hashwert zu finden.

Zu den bekanntesten Hashfunktionen gehören MD 4, MD 5 SHA 1, RIPEMD und RIPEMD 160. Einige davon sind zur Erzeugung von Pseudonymen unbrauchbar, da sie nicht kollisionsfrei sind.

4.9 Salt

„Salt“ (= „Salz“) bezeichnet in der Kryptografie eine zufällig gewählte Zeichenfolge, die an einen gegebenen Klartext vor der Verwendung als Eingabe einer Hash-Funktion angehängt wird, um die Entropie der Eingabe zu erhöhen, was letztlich zu einer höheren Streuung des Ergebnisses führt. Hierdurch kann z. B. verhindert werden, dass Originaldaten bspw. mit Hilfe von Rainbow-Tabellen identifiziert werden können.

5. Beispiele für die Vorgehensweise

Datentyp	Methode
Zahl	– Neuvergabe der letzten x Stellen (x = abhängig von den Zahlenwerten) – Ersetzen durch Zufallszahlen – Nutzung einer Varianz (z. B. ± x%) – Löschung
String	– Neuvergabe über Tabelle – Ersetzung durch feste Zeichenkette – Ersetzung durch feste Zeichenkette mit laufender Nummer zwecks Beibehaltung der Unterscheidbarkeit
Datum	– Setzen von Tag und Monat auf festen Wert – Setzen des Datums auf einen festen Wert
Postleitzahl	– Neuvergabe von mindestens den letzten 2 Stellen über Umsetzungstabelle – Ersetzen von mindestens den letzten beiden Stellen durch festen Wert – Ersetzen von mindestens den letzten beiden Stellen durch festen Zufallswert
E-Mail-Adresse	– Löschen – Ersetzen durch festen Dummy-Wert
Religion	– Löschen – Ersetzen durch festen Dummy-Wert
Medizinische Code-Systeme wie ICD, OPS, usw.	– Verkürzen der Kodierung – Löschung

6. Checkliste

Daten	Methode
Patientenname	• verschlüsselt/gehasht
Geburtsdatum	• Setzen von Tag und Monat auf festen Wert
Alte	• Nutzung einer Varianz (z. B. ± x%) •
Rasse	• Ersetzung durch feste Zeichenkette •
Arztbesuch	• Setzen von Tag und Monat auf festen Wert
Andere Daten	• Unverändert

- https://www.srd-rechtsanwaelte.de/blog/anonymisierung-pseudonymisierung/
- https://www.bmwi.de/Redaktion/DE/Publikationen/Industrie/anonymisierung-im-datenschutz.pdf?__blob=publicationFile&v=4
- https://dsgvo-gesetz.de/
- https://issuu.com/bdi-berlin/docs/20201103_leitfaden_bdi_anonymisierung_personenbezo
- https://www.gesundheitsdatenschutz.org/download/Pseudonymisierung-Anonymisierung.pdf