De-Identifizierung von Daten nach ISO 20889

Nicht erst seit der DSGVO ist der Schutz personenbezogener Informationen wichtig. An vielen Stellen fallen diese an. Doch irgendwann ist der Zeitpunkt erreicht, an dem diese gelöscht werden müssen. Dies ist aber nicht einfach so möglich. Gerade wenn Datensätze miteinander verknüpft sind. Doch wie kann sichergestellt werden, dass sensible Daten geschützt bleiben – selbst dann, wenn sie geteilt, gespeichert oder analysiert werden?

Die Norm ISO/IEC 20889 liefert darauf eine zentrale Antwort: durch strukturierte Techniken zur De-Identifizierung von Daten. Diese Techniken ermöglichen es, Daten so umzuwandeln, dass sie ihren Nutzen behalten, ohne dabei Rückschlüsse auf einzelne Personen zuzulassen.

Überblick der De-Identifizierungstechniken

Die ISO 20889 unterscheidet acht Hauptkategorien von Techniken, die je nach Datentyp, Risikoszenario und gewünschter Datenverwendbarkeit eingesetzt werden können:

1. Statistische Werkzeuge
2. Kryptografische Werkzeuge
3. Unterdrückungstechniken
4. Pseudonymisierung
5. Anatomie
6. Verallgemeinerung
7. Randomisierung
8. Synthetische Daten

Zusätzlich definiert die Norm formale Modelle wie K-Anonymität, differenzielle Privatsphäre und ein lineares Empfindlichkeitsmodell, um die Wirksamkeit der De-Identifizierung messbar zu machen.

Wie von ISO-Normen gewohnt, werden in Kapitel 3 verwendete Begriffe kurz erläutert. In dieser Norm sind es 39 Begriffe. Zusätzlich werden in Kapitel 4 21 Symbole und Abkürzungen eingeführt. Da die Techniken sehr komplex sein können, geht das Kapitel 6 auf das technische Modell und die Terminologie ein, das Kapitel 7 beschreibt, wie das Risiko einer Re-Identifizierung reduziert wird sowie erläutert das Kapitel 8 die Nützlichkeit von de-identifizierter Daten. Das Kapitel 9 zeigt in der Folge die verschiedenen Techniken, auf welche wir in der Folge kurz eingehen.

1. Statistische Werkzeuge

Diese Methoden verändern die Gesamtstruktur von Datensätzen, um Gruppen statt Einzelpersonen abzubilden.

- Probenahme: Es wird nur eine zufällige Teilmenge der Daten verwendet. Dies reduziert das Risiko, da Angreifer nicht sicher sein können, ob ein bestimmter Datensatz in der Stichprobe enthalten ist.
- Aggregation: Daten werden zu statistischen Kennzahlen zusammengefasst (zum Beispiel Mittelwerte oder Summen). Aggregierte Daten sind weniger granular, behalten aber ihren Nutzen für Analysen.

Beispiel: Anstelle individueller Gehälter wird das Durchschnittsgehalt einer Abteilung veröffentlicht.

2. Kryptografische Werkzeuge

Hier werden Verschlüsselungsmethoden genutzt, um Daten sicher zu transformieren.

- Deterministische Verschlüsselung: Ersetzt Klartext durch verschlüsselte Werte, die bei gleichem Schlüssel identisch sind. Ermöglicht exakte Abgleiche, bietet aber begrenzten Schutz gegen Verknüpfungsangriffe.
- Homomorphe Verschlüsselung: Ermöglicht Berechnungen auf verschlüsselten Daten ohne Entschlüsselung. Ideal für Cloud-Umgebungen, aber rechenintensiv.
- Formaterhaltende Verschlüsselung: Behält das Format der Originaldaten bei (zum Beispiel eine verschlüsselte Sozialversicherungsnummer bleibt eine 9-stellige Zahl).

Anwendung: Medizinische Daten können verschlüsselt an Dritte übermittelt werden, die statistische Analysen durchführen, ohne sensible Details einzusehen.

3. Unterdrückungstechniken

Auch als suppressive Methoden genannt. Die Daten werden teilweise oder vollständig entfernt oder unkenntlich gemacht.

- Maskierung: Direkte Identifikatoren (zum Beispiel Namen) werden gelöscht oder abgeschnitten.
- Lokale Unterdrückung: Einzelne Werte in bestimmten Datensätzen werden entfernt, wenn sie in Kombination mit anderen Attributen eine Identifizierung ermöglichen (zum Beispiel seltene Berufsbezeichnungen).
- Unterdrückung von Aufzeichnungen: Ganze Datensätze mit Ausreissern oder seltenen Merkmalen werden ausgeschlossen. Falls es aber kein hohes Risiko für eine Re-Identifikation gibt, bleiben die Informationen erhalten.

Der Nachteil ist dabei, dass der Informationsverlust die Nützlichkeit der Daten einschränken kann.

4. Pseudonymisierung

Die Identifikatoren werden durch künstliche Schlüssel (Pseudonyme) ersetzt. Die Rückverfolgung ist nur mit zusätzlichen Informationen möglich, die separat gesichert werden müssen. Die

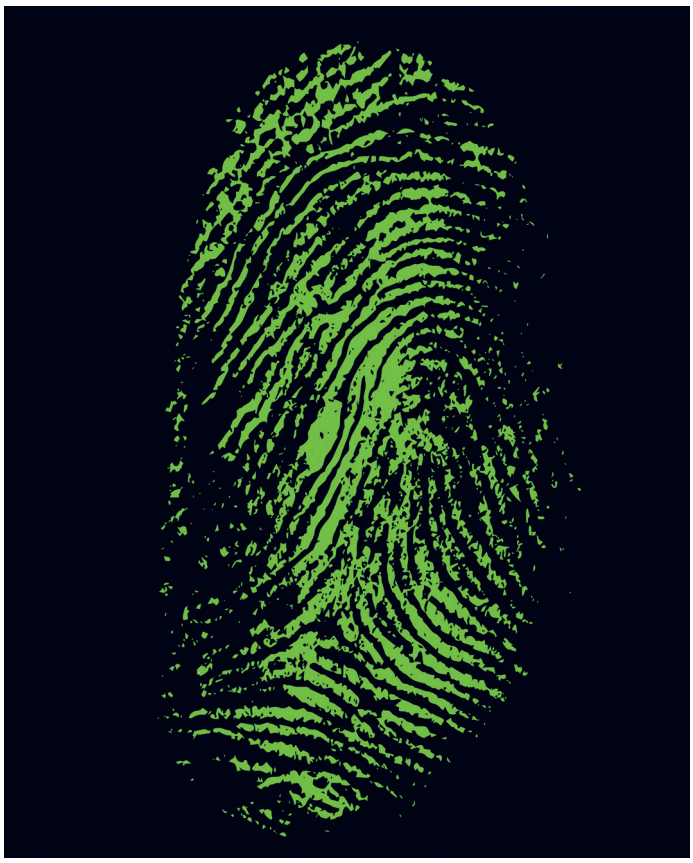

Bild: Pixabay

Die Technik der De-Identifizierung von Daten ermöglicht es, diese so umzuwandeln, dass sie ihren Nutzen behalten, ohne dabei Rückschlüsse auf einzelne Personen zuzulassen.

ZUM AUTOR

Andreas Wisler, Dipl. Ing FH
goSecurity AG
Schulstrasse 11
CH-8542 Wiesendangen

T +41 (0)52 511 37 37
www.goSecurity.ch
wisler@gosecurity.ch

Pseudonymisierung ist somit eine Technik, die es ermöglicht, miteinander verbundene Informationen aus verschiedenen Datensätzen zu verknüpfen, ohne die Identität der Dateninhaber preiszugeben.

- Kryptografische Pseudonyme: Hash-Funktionen oder Verschlüsselung generieren eindeutige Pseudonyme.
- Zufallsgenerierte Pseudonyme: Unabhängig von den Originaldaten erzeugt, oft mit Mapping-Tabellen.

Das Risiko besteht, dass bei unsachgemässer Schlüsselverwaltung eine Re-Identifizierung möglich sein kann.

5. Anatomie

Ein Datensatz wird in zwei Tabellen aufgeteilt: Eine enthält Identifikatoren, die andere sensible Attribute. Ein gemeinsames Schlüsselattribut ermöglicht die Verknüpfung, aber nur berechtigte Stellen haben Zugriff auf beide Tabellen.

Der Vorteil ist, dass die statistische Integrität erhalten bleibt, während direkte Identifikatoren geschützt sind.

6. Verallgemeinerung

Die Granularität der Daten wird reduziert, um Gruppenbildung zu fördern.

- Rundung: Zahlen werden auf glatte Werte gerundet (zum Beispiel Alter 23 > 20 bis 25).
- Kategorisierung: Kontinuierliche Werte werden in Intervalle gruppiert (zum Beispiel Einkommen: «niedrig», «mittel», «hoch»).
- Kombination von Attributen: Mehrere Merkmale werden zu einem übergeordneten Attribut zusammengefasst (zum Beispiel «IT-Beruf» statt «Softwareentwickler»).

Das Ziel ist stets, die Wiedererkennung einer Person zu erschweren, ohne die Aussagekraft der Daten vollständig zu verlieren. Jedoch ist die Herausforderung, dass zu starke Verallgemeinerung die Aussagekraft der Daten mindert.

7. Randomisierung

Die Daten werden gezielt verrauscht oder permutiert (verändert), um Rückschlüsse zu erschweren.

- Lärmzusatz: ein Wert wird künstlich leicht verändert (zum Beispiel die Note 3.33 wird zu 3.53).
- Permutation: Attributwerte werden innerhalb des Datensatzes vertauscht (zum Beispiel Geschlechter werden zufällig zugeordnet, aber die Gesamtverteilung bleibt erhalten).
- Mikroaggregation: Ähnliche Datensätze werden gruppiert und durch Durchschnittswerte ersetzt.

Der Nutzen ist, dass die statistischen Eigenschaften erhalten bleiben, während Einzeldaten unkenntlich gemacht werden.

8. Synthetische Daten

Dabei handelt es sich um künstlich generierte Datensätze, die reale Muster nachbilden, aber keine echten Personen darstellen. Sie werden mithilfe von Algorithmen erzeugt, die auf Originaldaten trainiert wurden.

Als Anwendung sind Softwaretests oder Schulungszwecke, wo realistische, aber anonyme Daten benötigt werden.

Formale Modelle zur Messung des Datenschutzes

Die Wirksamkeit von De-Identifizierungsverfahren lässt sich nicht allein durch technische Intuition bewerten. Formale Modelle liefern messbare Kriterien, wie gut die Privatsphäre geschützt wird. Um diese Wirksamkeit der De-Identifizierung zu quantifizieren, definiert die ISO 20889 mehrere Modelle:

- K-Anonymität: Jeder Datensatz in einer Gruppe ist mindestens K-mal vertreten. Ein Angreifer kann Einzelpersonen nicht von anderen unterscheiden.
- L-Diversität: In jeder Gruppe gibt es mindestens L unterschiedliche Werte für sensible Attribute. Dies verhindert Rückschlüsse auf spezifische Merkmale.
- T-Knappheit: Die Verteilung sensibler Attribute in einer Gruppe darf nicht mehr als T von der Gesamtverteilung abweichen.
- Differenzielle Privatsphäre: Fügt kontrolliertes Rauschen hinzu, um zu garantieren, dass die An- oder Abwesenheit eines Einzeldatensatzes die Ergebnisse nicht oder nur kaum beeinflusst.

Fazit: Die richtige Technik wählen

Die ISO 20889 bietet keinen «Einheitsansatz», sondern betont die kontextspezifische Auswahl:

1. Datenart: Strukturierte Daten (Tabellen) erlauben andere Methoden als Freitext oder Multimedia-Dateien.
2. Risikobewertung: Bei hochsensiblen Daten (zum Beispiel Gesundheitsinformationen) sind starke kryptografische oder randomisierte Verfahren ratsam.
3. Nützlichkeit: Aggregation und Verallgemeinerung erhalten die Analysefähigkeit, während Unterdrückung oder Anatomie gezielt Risiken minimieren.
4. Kombination: Oft sind mehrere Techniken nötig (zum Beispiel Pseudonymisierung und Randomisierung).

Die Norm ISO/IEC 20889 gibt Unternehmen ein umfassendes Instrument an die Hand, um personenbezogene Daten effektiv zu de-identifizieren und dennoch ihren Nutzen zu bewahren. Die Vielfalt der beschriebenen Techniken zeigt, dass Datenschutz nicht durch ein einzelnes Werkzeug erreicht werden kann. Vielmehr bedarf es einer bewussten Kombination von Methoden, die auf den jeweiligen Kontext und das Schutzniveau abgestimmt sind. Die Norm kann mittels QR-Code gekauft werden.