

Anonymisierung, Schwärzen und Pseudonymisierung von Daten

1. Was bedeutet Anonymisierung?

Anonymisierung bedeutet, dass personenbezogene Daten so verändert werden, dass eine natürliche Person nicht mehr identifiziert werden kann. Dies gilt nicht nur für direkte Angaben wie Name, Adresse, Telefonnummer oder E-Mail-Adresse, sondern auch für indirekte Merkmale, die in Kombination Rückschlüsse auf eine Person zulassen können.

Eine echte Anonymisierung liegt nur dann vor, wenn die Zuordnung zu einer Person nicht mehr oder nur mit unverhältnismäßig hohem Aufwand möglich ist. Entscheidend ist daher nicht nur, ob offensichtliche Namen entfernt wurden, sondern ob eine Re-Identifikation praktisch ausgeschlossen oder zumindest nach heutigem Stand nicht realistisch ist.

2. Begründung mit Bezug zur DSGVO

Die Datenschutz-Grundverordnung gilt für personenbezogene Daten. Nach Erwägungsgrund 26 DSGVO fallen anonymisierte Informationen nicht unter die Datenschutzgrundsätze, wenn sich die Informationen nicht mehr auf eine identifizierte oder identifizierbare natürliche Person beziehen.

Wichtig ist jedoch: Sobald Daten mit zusätzlichen Informationen wieder einer Person zugeordnet werden können, handelt es sich nicht um echte Anonymisierung. Pseudonymisierte Daten bleiben daher in der Regel personenbezogene Daten, wenn eine Rückführung über Schlüssel, Listen oder andere Zusatzinformationen möglich ist.

3. Methoden zur Anonymisierung

- Schwärzen
- Pseudonymisierung
- Generalisierung
- Aggregation
- Synthetische Daten

4. Schwärzen ist nicht automatisch Anonymisieren

Beim Schwärzen werden sensible Informationen in einem Dokument unkenntlich gemacht, zum Beispiel durch schwarze Balken über Namen, Adressen oder Aktenzeichen. Für eine echte, datenschutzkonforme Schwärzung reicht es jedoch nicht aus, Text nur optisch zu überdecken.

Ein häufiger Fehler besteht darin, Text in Word, PDF-Dateien oder Standard-Viewern lediglich mit einem schwarzen Balken zu übermalen. Der Text kann dann im Hintergrund weiterhin vorhanden sein, kopiert, durchsucht oder technisch wiederhergestellt werden. Auch Metadaten wie Autor, Änderungsverlauf, Kommentare, Dateieigenschaften oder eingebettete Informationen können personenbezogene Daten enthalten.

Datenschutzkonformes Schwärzen bedeutet daher: Die betroffenen Inhalte müssen tatsächlich aus der Datei entfernt werden. Zusätzlich sollten Metadaten geprüft und bereinigt werden. Das Ergebnis muss kontrolliert werden, indem versucht wird, den geschwärzten Text zu markieren, zu kopieren oder über die Suche wiederzufinden.

5. Generalisierung

Bei der Generalisierung werden genaue Angaben durch allgemeinere Kategorien ersetzt. Dadurch bleiben Informationen für Auswertungen nutzbar, während das Risiko einer Identifizierung sinkt.

Beispiel:

- Geburtsdatum: 14.03.1982 → Altersgruppe: 40–49 Jahre
- Wohnort: Eschwege → Region: Nordhessen

Je allgemeiner die Kategorien gewählt werden, desto geringer ist in der Regel das Risiko einer Re-Identifikation.

6. Aggregation

Aggregation bedeutet, dass Einzelangaben zu Gruppen- oder Summenwerten zusammengefasst werden. Die Betrachtung einzelner Personen tritt dabei in den Hintergrund.

Beispiel:

Statt die Weiterbildungsstunden einzelner Beschäftigter aufzulisten, wird nur angegeben:

- Abteilung A: 320 Weiterbildungsstunden
- Abteilung B: 280 Weiterbildungsstunden

Die Auswertung bezieht sich damit auf Gruppen und nicht mehr auf einzelne Personen.

7. Synthetische Daten

Synthetische Daten sind künstlich erzeugte Daten, die echte Daten ersetzen oder nachbilden können. Sie können vollständig fiktiv erstellt werden, zum Beispiel für Schulungen, Softwaretests oder Demonstrationen. Sie können aber auch aus Mustern echter Datensätze abgeleitet werden, etwa wenn statistische Zusammenhänge einer Studie erhalten bleiben sollen, ohne die Originalpersonen zu übernehmen.

Beispiel: Eine echte Studie enthält Angaben zu Alter, Beruf, Wohnregion und Weiterbildungsinteresse. Aus diesen Mustern kann ein künstlicher Datensatz erzeugt werden, der ähnliche statistische Zusammenhänge enthält, aber keine realen Personen abbildet.

Synthetische Daten sind besonders nützlich, wenn mit realistischen Beispielen gearbeitet werden soll, ohne echte personenbezogene Daten zu verwenden. Sie sind jedoch nicht automatisch anonym. Wenn synthetische Daten zu nah an den Originaldaten erzeugt werden, können unter Umständen Rückschlüsse auf einzelne Personen möglich sein. Deshalb muss auch bei synthetischen Daten geprüft werden, ob ein Re-Identifikationsrisiko besteht.

Für Trainings, Demonstrationen und KI-Übungen sind vollständig fiktive synthetische Daten oft die sicherste Lösung.

8. Pseudonymisierung

Pseudonymisierung bedeutet, dass direkte Identifikatoren durch künstliche Kennungen ersetzt werden. Namen, Personalnummern oder Kundennummern werden zum Beispiel durch IDs, Codes oder Referenznummern ersetzt. Die eigentlichen Personen sind dann in der Arbeitsdatei nicht mehr direkt sichtbar.

Beispiel:

Originaldaten:

Name	Alter	Ort	Krankheit	Kommentar
Maria Müller	42	Eschwege	Migräne	arbeitet bei Firma X

Pseudonymisierte Arbeitskopie:

ID	Alter	Ort	Krankheit	Kommentar
P-83921	42	Eschwege	Migräne	arbeitet bei Firma X

Der Name wurde ersetzt, aber die Daten sind dadurch nicht automatisch anonym. Die Person könnte weiterhin erkennbar sein, wenn mehrere Merkmale zusammengeführt werden. Alter, Wohnort, Krankheit, Arbeitgeber oder Freitext-Kommentare können in Kombination ausreichen, um eine Person wiederzuerkennen.

Pseudonymisierung ist daher eine wichtige Schutzmaßnahme, aber keine vollständige Anonymisierung. Solange eine Zuordnungstabelle existiert, zum Beispiel „P-83921 = Maria Müller“, bleiben die Daten grundsätzlich personenbezogen. Die DSGVO gilt weiterhin.

In der Praxis bedeutet das:

1. Direkte Identifikatoren sollten entfernt oder ersetzt werden. Dazu gehören Namen, Adressen, Telefonnummern, E-Mail-Adressen, Kundennummern, Personalnummern und Aktenzeichen.
2. Indirekte Identifikatoren müssen ebenfalls geprüft werden. Dazu gehören Alter, Wohnort, Beruf, Abteilung, Diagnose, seltene Ereignisse, kleine Gruppen und auffällige Kombinationen von Merkmalen.
3. Genaue Angaben sollten bei Bedarf in allgemeinere Kategorien überführt werden. Aus einem exakten Geburtsdatum kann zum Beispiel eine Altersgruppe werden. Aus einem konkreten Wohnort kann eine Region werden.
4. Freitextfelder sind besonders kritisch. Dort stehen häufig unbeabsichtigt Namen, Orte, Arbeitgeber, familiäre Informationen oder konkrete Einzelfälle.

5. Die Zuordnungstabelle zwischen Klarnamen und IDs muss getrennt gespeichert werden. Sie darf nicht gemeinsam mit der pseudonymisierten Arbeitskopie weitergegeben oder in ein KI-Tool hochgeladen werden.
6. Die Zuordnungstabelle sollte stark zugriffsbeschränkt und verschlüsselt gespeichert werden. Nur Personen mit einem berechtigten Zweck sollten Zugriff darauf haben.
7. Auch die Originaldaten dürfen bestehen bleiben, wenn dafür ein legitimer Zweck und eine Rechtsgrundlage vorliegen. Sie müssen jedoch geschützt, zweckgebunden verarbeitet und nach Ablauf der erforderlichen Aufbewahrungsfrist gelöscht werden.
8. Für KI-Analysen ist besondere Vorsicht erforderlich. Eine Arbeitskopie mit ersetzten Namen reicht in der Regel nicht aus, wenn weitere Merkmale eine Re-Identifikation ermöglichen. Für externe KI-Tools sollten nach Möglichkeit nur wirklich anonymisierte oder vollständig synthetische Daten verwendet werden.

9. Merksatz

Anonymisierung bedeutet nicht nur, Namen zu entfernen. Entscheidend ist, ob eine Person mit realistischem Aufwand wiedererkannt werden kann. Pseudonymisierung reduziert Risiken, ersetzt aber keine vollständige Anonymisierung.