Wie viel Open Data kann es geben? Data Linkage und Re-Identifizierungsrisiken

Elena März, Johann Guggumos, Sebastian Wilhelm

Angaben zur Veröffentlichung / Publication details:

März, Elena, Johann Guggumos, and Sebastian Wilhelm. 2024. “Wie viel Open Data kann es geben? Data Linkage und Re-Identifizierungsrisiken.” Datenschutz und Datensicherheit - DuD 48 (6): 378–82. https://doi.org/10.1007/s11623-024-1925-y.

Elena März, Johann Guggumos, Sebastian Wilhelm

Wie viel Open Data kann es geben?

Data Linkage und Re-Identifizierungsrisiken

In Zeiten wachsender Datenmengen und eines zunehmenden Interesses an Open Data steht die Frage im Raum: „Wie viel Offenheit ist möglich?“ Zentrale Zielsetzung jeder Veröffentlichung von Daten als Open Data ist deren Anonymität. Nur dann liegen die Daten außerhalb des Geltungsbereichs der Datenschutz-Grundverordnung. Die steigende Verfügbarkeit von personenbezogenen Zusatzinformationen aus öffentlichen Quellen wie sozialen Medien erschweren allerdings die Anonymisierung, weil mit zunehmender Datenmenge auch die Möglichkeiten einer Verknüpfung von Daten (Data Linkage) zunehmen und damit auch das Risiko einer möglichen Re-Identifizierung steigt. Die Veröffentlichung von anonymisierten Daten erfordert daher aufgrund des unumkehrbaren Charakters eine umfassende Analyse der Re-Identifizierungsrisiken.

Elena März

ist wissenschaftliche Mitarbeiterin am Lehrstuhl für Bürgerliches Recht, Haftungsrecht und Recht der Digitalisierung der Universität Augsburg und im Drittmittelprojekt EAsyAnon tätig.
E-Mail: elena.maerz@jura.uni-augsburg.de

Johann Guggumos

ist wissenschaftlicher Mitarbeiter am Lehrstuhl für Bürgerliches Recht, Haftungsrecht und Recht der Digitalisierung der Universität Augsburg und im Drittmittelprojekt EAsyAnon tätig.
E-Mail: johann.guggumos@jura.uni-augsburg.de

Sebastian Wilhelm

ist wissenschaftlicher Mitarbeiter im Fachbereich Informatik am Technologie Campus Grafenau der Technischen Hochschule Deggendorf. Sein Forschungsschwerpunkt liegt in der Anwendung von Datenanalyse und maschinellem Lernen zur automatisierten Erkennung von Aktivitäten und Notfällen sowie in der Entwicklung und Umsetzung von Maßnahmen im Bereich des technischen Datenschutzes.
E-Mail: sebastian.wilhelm@th-deg.de

1 Problemaufriss

Die europäische Datenschutz-Grundverordnung bezieht sich in ihrem Anwendungsbereich ausschließlich auf personenbezogene Daten. Damit stehen alle Daten, die nicht unter diesen Schutzbereich fallen, für die Nutzung als Open Data zur Verfügung. Trotz der signifikanten Unterscheidung von Personenbezogenheit und Anonymität findet sich in der Verordnung nur die Definition von personenbezogenen Daten. Gemäß Art. 4 Abs. 1 Nr. 1 DS-GVO sind personenbezogene Daten alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person beziehen. Als identifizierbar wird eine natürliche Person angesehen, die direkt oder indirekt, insbesondere mittels Zuordnung zu einer Kennung wie einem Namen, Kennnummer, Standortdaten, Online-Kennung oder zu einem oder mehreren besonderen Merkmalen, die Ausdruck der physischen, physiologischen, genetischen, psychischen, wirtschaftlichen, kulturellen oder sozialen Identität dieser natürlichen Person sind.

Grundsätzlich kann die Identifizierung durch „direkte“ und „indirekte“ Identifikatoren erfolgen kann.[1] So kann eine Person direkt durch ihren Namen identifiziert werden, während sie indirekt durch eine Telefonnummer, ein Autokennzeichen oder durch eine Kombination signifikanter Kriterien identifiziert werden kann, die es ermöglichen, sie durch Eingrenzung der Gruppe, zu der sie gehört (Alter, Beruf, Wohnort), zu erkennen.[2] Insbesondere bei den indirekten Identifikatoren stellt sich die Frage, wann ein Bezug zu einer Person noch besteht oder hergestellt werden kann. Manche Merkmale sind so eindeutig, dass eine Person

1 In diesem Sinne schon Article 29 Data Protection Working Party, Opinion 4/2007 on the concept of personal data, WP 136, S. 12.

2 Article 29 Data Protection Working Party, Opinion 4/2007 on the concept of personal data, WP 136, S. 13.

ohne Aufwand identifizierbar ist („derzeitiger Ministerpräsident Spaniens"). Aber auch eine Kombination mehrerer Angaben kann spezifisch genug sein, um sie auf eine Person einzugrenzen.[3] Besonders die Kombination solcher verschiedenen Identifikatoren steht nachfolgend im Fokus. Dies gilt gerade im Hinblick auf getrennte Datensätze mit demselben Ursprung, die durch Zusammenführung eine Gesamtheit bilden können.[4] Damit ist Anonymität bislang eine Kontextfrage.[5] Maßgeblich ist der Kontext, in welchem Raum und für welchen Adressatenkreis Daten anonym bereitgestellt werden. Denn aus dieser Empfängersicht beurteilt sich, wie Einträge des Datensatzes mit zusätzlichen Informationen verbunden werden können, um einen hinreichenden Bezug zu einer Person herzustellen. Open Data löst jedoch den beschränkten Adressatenkreis auf und erzeugt einen dynamischen Kontext, der mit jedem neuen Datensatz die Anonymität auflösen kann.

Nach EG 26 DS-GVO sind bei der Feststellung, ob Mittel zur Identifizierung der natürlichen Person nach vernünftigem Ermessen eingesetzt werden können, alle objektiven Faktoren zu berücksichtigen, wie etwa die Kosten und der Zeitaufwand für die Identifizierung, wobei die zum Zeitpunkt der Verarbeitung verfügbare Technologie und die technologische Entwicklung zu beachten sind.

Die Wahrscheinlichkeit des Risikos einer erneuten Identifizierung zu bewerten, bestimmt sich gemäß EG 26 DS-GVO durch eine Analyse der einschlägigen technischen Kenntnisse und Mittel. Diese hat einerseits juristisch, andererseits aufgrund der Referenz der technischen Möglichkeiten aus technischer Sicht zu erfolgen.[6]

2 Zur Verfügung stehende Mittel und Zusatzwissen

Ausgangspunkt für die Risikoanalyse von Open Data ist die Betrachtung des öffentlichen Personenkreises, als Adressaten der Durchführung einer Re-Identifizierung.[7]

Es müssen Informationen aus öffentlichen Suchmaschinen wie Google, Bing oder ähnlichem in die Analyse einfließen, weil diese ein Zusatzwissen darstellen. Zudem werden durch die einzelnen Personen immer mehr persönliche Daten freiwillig in sozialen Netzwerken wie Instagram, Facebook oder X (ehemals Twitter) geteilt und diese Informationen werden häufig auf öffentlichen Profilen zugänglich gemacht. Die frei zugängliche Datenmenge steigt somit kontinuierlich an. Der EuGH hat alles, was frei zugänglich oder durch rechtliche Mittel und Ansprüche zugänglich gemacht werden kann und darf, als zu berücksichtigendes Zusatzwissen eingestuft.[8] Bislang war der Adressatenkreis der Datensätze beschränkt, sodass auch die zur Verfügung stehenden Mittel beschränkt waren. Der Übergang von beschränktem zu unbeschränktem Adressatenkreis führt zu einem größeren Umfang an Mitteln und Wissen, die einem potenziellen Empfänger zur Verfügung stehen könnten.

Diese steigenden Datenmengen und Zusatzinformationen könnten in der Zukunft zu einer Re-Identifizierung führen.[9] Bei der Frage der Re-Identifizierung müssen im Rahmen der Risikoanalyse insbesondere drei Faktoren bedacht werden, wie sie schon von der Artikel 29 Datenschutzgruppe für die Anonymisierungsbewertung entwickelt worden sind.[10] Erster Risikoaspekt ist die Gefahr des Aussonderns einer einzelnen Person[11] aufgrund von Alleinstellungsmerkmalen innerhalb eines Datensatzes.[12] Ein weiterer Risikofaktor ist die Inferenz[13], das heißt eine Ableitung von personenbezogenen Informationen aufgrund logischer Schlussfolgerung. Der hier behandelte dritte und zentrale Risikoaspekt ist die Verknüpfbarkeit von Daten, weil mit steigenden verfügbaren Daten auch das Risiko wächst.[14] Werden mehrere Datensätze durch denselben Datenverantwortlichen als Open Data veröffentlicht, müssen diese als besonders sensibel betrachtet werden. Es ist sicherzustellen, dass die Gesamtheit dieser Datensätze die Anonymität bewahrt. Wenn durch die Kombination dieser Datensätze die Identifikation einzelner Personen möglich wird, da nun einzelne Attribute miteinander verknüpft werden können, müssen die Daten nach wie vor als personenbezogen betrachtet werden.[15] Führt man diesen Gedanken fort, gilt dies nicht nur für andere veröffentlichte Datensätze, sondern für jegliche Informationen, die dazu geeignet sind, eine ausreichende Verknüpfung herzustellen. Werden viele Datensätze als Open Data veröffentlicht und/oder sind zudem viele personenbezogene Daten verfügbar, sollten laut Art. 29 Datenschutzgruppe all diese Zusatzinformationen in die Risikobewertung der Verknüpfbarkeit bei Open Data einbezogen werden.

Nach EG 26 S. 3 und 4 DS-GVO müssen diese Informationen jedoch auch „nach allgemeinem Ermessen wahrscheinlich zur Identifizierung natürlicher Personen genutzt werden". Daher ist entscheidend, wie viel Zeitaufwand, Know-how und damit verbundene Kosten notwendig sind, um mit derart verfügbaren Informationen einen Datensatz zu De-Anonymisieren.

Ausgangspunkt für die Wahrscheinlichkeitsprognose einer De-Anonymisierung von Open Data sind Datenschutzmodelle wie k-Anonymität, l-Diversität und t-Closeness, weil sie die Datensätze in ein strukturiertes Format übersetzen, das als Bewertungsgrundlage für die Prognose dient.[16] Diese können bewirken, dass einzelne Daten innerhalb eines Datensatzes nicht ohne weiteres einer bestimmten Person zugeordnet werden können.[17] Anhand von Generalisierungsverfahren werden durch die k-Anonymität Gruppen mit selben Werten pro Einträgen gebildet, um vor dem Risiko der Aussonderung zu schützen.[18] Weiterhin werden durch l-Diversität und t-Closeness Risiken durch Inferenz aus-

3 Article 29 Data Protection Working Party, Opinion 4/2007 on the concept of personal data, WP 136, S. 13.

4 Specht/Matz, Handbuch Europäisches und deutsches Datenschutzrecht, § 15 Rn. 70.

5 Vgl. Gierschmann, ZD 2021, 482, 483.

6 Gierschmann, ZD 2021, 482, 485.

7 Vgl. Stummer, ZfDR, 2023, 263, 269 f.

8 EuGH, Urteil vom 19.10.2016, C-582/14, Rn. 49.

9 Vgl. Schlussantrag GA Sanchez, C- 582/14 Breyer, Rn. 68; zum Zeitpunkt Kühling/Buchner, DS-GVO, Art. 4 Nr. 1, Rn. 24.

10 Article 29 Data Protection Working Party, Opinion 5/2014 on Anonymisation Techniques, WP 216, S. 10 ff.

11 Article 29 Data Protection Working Party, Opinion 5/2014 on Anonymisation Techniques, WP 216, S. 10 ff.

12 So auch Anderl/Kruesz, MMR 2023, 255, 257.

13 Article 29 Data Protection Working Party, Opinion 5/2014 on Anonymisation Techniques, WP 216, S. 10 ff.

14 Article 29 Data Protection Working Party, Opinion 5/2014 on Anonymisation Techniques, WP 216, S. 10 ff.

15 Bischoff, PharmaR 2020, 309, 313.

16 Article 29 Data Protection Working Party, Opinion 5/2014 on Anonymisation Techniques, WP 216, S. 19 f.

17 Mühlenbeck, Anonyme und pseudonyme Daten, S. 313 f.

18 Article 29 Data Protection Working Party, Opinion 5/2014 on Anonymisation Techniques, WP 216, S. 19 f.

geschlossen, weil hierbei die Verteilung der Werte von Einträgen beschränkt wird, um keine logischen Schlussfolgerungen ziehen zu können.[19] Diese Datenschutzmodelle haben jedoch gemein, dass sie keine Lösung gegen die Verknüpfbarkeit bieten. Es besteht weiterhin die Möglichkeit, dass Datensätze mit ähnlichen Einträgen kombiniert werden können. Sobald ein Eintrag aufgrund von Zusatzwissen nach EG 26 S. 3 DS-GVO ausgesondert werden kann und damit identifizierbar ist, liegt folglich wieder ein personenbezogenes Datum vor.

3 Technische Verknüpfbarkeit von anonymisierten Datensätzen

Damit zeigt sich, dass Datenschutzmodelle hinsichtlich des Risikofaktors Verknüpfbarkeit an ihre Grenzen stoßen. Zur Evaluierung des Re-Identifizierungsrisikos sind die technischen und realistischen Möglichkeiten zur Herstellung von Verknüpfungen zwischen Datensätzen entscheidend. Obwohl diese Datensätze an sich keine direkten personenbezogenen Informationen preisgeben, ermöglicht die Kombination verschiedener Datenquellen eine potenzielle Re-Identifizierung von Individuen. In diesem Kapitel stellen wir ausschließlich fiktive Beispiele vor, die jedoch realistische Methodiken der Datenverknüpfung aufzeigen.

3.1 Verknüpfbarkeit durch Überschneidungen

Eine grundlegende Technik zur Verknüpfung verschiedener Datenquellen besteht in der Erstellung einer Zuordnungstabelle. Diese Methode nutzt überlappende Informationen, die in zwei unterschiedlichen Datensätzen vorhanden sind, um Korrelationen zwischen einzelnen Datensatzzeilen herzustellen. Existiert innerhalb dieser Zeilen eine pseudonymisierte ID, ermöglicht dies den Aufbau einer Zuordnungstabelle. Diese Tabelle verknüpft die pseudonymisierten IDs aus Datensatz A mit jenen aus Datensatz B, wodurch eine direkte Verbindung zwischen den Datensätzen etabliert wird.[20]

Fallbeispiel 1:
Zur Veranschaulichung dient das folgende hypothetische Szenario mit zwei Datensätzen: ein von einem Kreditkartenunternehmen anonym bereitgestellter Datensatz von Finanztransaktionen (siehe Tabelle 1) und ein Datensatz von Flugbuchungsinformationen, bereitgestellt von einer Fluggesellschaft (siehe Tabelle 2). Beide Datensätze enthalten pseudonymisierte Kunden-IDs, sind jedoch unabhängig voneinander.

Durch die Analyse der Überschneidungen zwischen beiden Datensätzen lassen sich Zuordnungen zwischen Transaktionen und Flugbuchungen herstellen, etwa auf Basis des *Buchungsdatums* und der *Transaktionskategorie*. Beispielsweise:

- Transaktion #2 --> Flugbuchung #1 oder #2
- Transaktion #5 --> Flugbuchung #4

Durch logische Verknüpfungen der Zuordnungen lassen sich damit die unabhängigen pseudonymen Kunden-IDs miteinander verbinden. Ein illustratives Beispiel hierfür ist die Zuordnung des Kunden *CUST-171* aus dem Finanztransaktionsdatensatz zum Passagier *PAS-535* aus dem Flugbuchungsdatensatz, was eine direkte Verbindung zwischen diesen Datensätzen etabliert.

Dieses vereinfachte Beispiel verdeutlicht das Potenzial der Datenverknüpfung in einem Big-Data-Kontext, bei dem auch minimale Schnittmengen an übereinstimmenden Informationen zur Integration voneinander unabhängiger Datensätze führen können.

Tabelle 1 | Datensatz von Finanztransaktionen eines Kreditkartenunternehmens

	Kunde	Datum	Betrag	Kategorie
1	CUST-202	2023-07-23	145,39 €	Unterhaltung
2	CUST-171	2023-03-29	1.225,25 €	Flugbuchung
3	CUST-960	2023-11-27	6,55 €	Abonnements
4	CUST-370	2023-11-15	98,46 €	Restaurant
5	CUST-171	2023-09-24	302,25 €	Flugbuchung
6	CUST-714	2023-06-10	51,01 €	Unterhaltung
7	CUST-370	2023-03-21	23,43 €	Unterhaltung

Tabelle 2 | Datensatz von Flugbuchungsinformationen einer Fluggesellschaft

	Passagier	Datum Buchung	Alter	Geschlecht	Abflug	Ziel
1	PAS-535	2023-03-29	57	Männlich	JFK	MUC
2	PAS-370	2023-03-29	52	Weiblich	SIN	JFK
3	PAS-171	2023-10-11	61	Männlich	SIN	DXB
4	PAS-535	2023-09-24	57	Männlich	MUC	AMS
5	PAS-206	2023-04-01	39	Männlich	DXB	AMS
6	PAS-221	2023-05-31	44	Weiblich	FRA	MUC

3.2 Verknüpfbarkeit durch Mustererkennung

Eine weitere Möglichkeit, Datensätze miteinander zu verknüpfen, basiert auf der Identifikation von Mustern. Diese Methode beruht auf der Analyse und dem Abgleich von Verhaltensmustern, Aktivitätsprofilen oder anderen charakteristischen Mustern, die in verschiedenen, unabhängigen Datensätzen erkannt werden können. Im Gegensatz zur Nutzung direkter Überschneidungen in den Datenquellen, nutzt die Mustererkennung indirekte Merkmale und Verhaltensweisen, um Übereinstimmungen zwischen Datensätzen zu finden. Dies ermöglicht die Verknüpfung von Daten, selbst wenn keine offensichtliche Verbindung auf der Ebene individueller Identifikatoren besteht.[21]

Fallbeispiel 2:
Zur Veranschaulichung dient das folgende hypothetische Szenario mit zwei unabhängigen Datensätzen: ein Datensatz eines Mobilfunkanbieters über Funkzelleneinwahlen von Smartphone-Nutzern (siehe Tabelle 3) und ein Datensatz über Beiträge in Sozialen Medien (siehe Tabelle 4).

Die Analyse von Bewegungsmustern aus Mobilfunkdaten in Verbindung mit Standorttags, Aktivitäten und Beiträgen aus sozialen Medien ermöglicht das Erkennen von Übereinstimmun-

19 Article 29 Data Protection Working Party, Opinion 5/2014 on Anonymisation Techniques, WP 216, S. 22f.

20 Vgl. Christen, Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection, 2012; Herzog/Scheuren/Winkler, Record Linkage – Methodology, 2007.

21 Vgl. Zhang/Yang/Zhang/Xu, De-anonymization Attack Method of Mobility Trajectory Data Based on Semantic Trajectory Pattern, 2021.

Tabelle 3 | Datensatz über Eingewählte eines Mobilfunkanbieters

	Nutzer ID	Datum	Uhrzeit	Funkzelle
1	48e84d329	2023-04-01	10:00	Zentrum Stadt A
2	49x4917c9	2023-04-01	17:00	Industriegebiet Stadt B
3	49x4917c9	2023-04-01	18:00	Wohngebiet Stadt A

Tabelle 4 | Datensatz aus sozialen Medien

	Nutzername	Datum	Uhrzeit	Veröffentlichter Inhalt
1	Nutzer123	2023-04-01	08:00	„Toller Start in den Tag“
2	Nutzer456	2023-04-02	10:00	„Neuer Job: Bei Firma XYZ“
3	Nutzer456	2023-04-02	18:15	„Endlich zuhause! Feierabendlauf durch das Wohngebiet“

gen, die auf dieselbe Person hinweisen könnten. Die regelmäßige Bewegung zwischen einem Industriegebiet in Stadt B und einem Wohnbereich in Stadt A, ergänzt durch Posts in sozialen Netzwerken über den Wohnort und Arbeitsplatz, kann eine Zuordnung von einem anonymen Mobilfunknutzer zu einem Social-Media-Profil ermöglichen. Der Einsatz von Deep-Learning-Modellen verstärkt diese Möglichkeit, indem Zusammenhänge aufdeckt werden, die für menschliche Analysten verborgen bleiben könnten.

3.3 Verknüpfbarkeit durch trainierte KI-Modelle

Eine weitere Methode zur Verknüpfung von Datensätzen bietet der Einsatz trainierter KI-Modelle. Diese Technik verwendet identische oder ähnliche Merkmale aus zwei unabhängigen Datensätzen, um mittels prädiktiver KI-Analysen Vorhersagen zu generieren und so einen Datensatz zu bereichern. Ein Datensatz dient dabei als Trainingsgrundlage für ein KI-Modell, das anschließend auf einen zweiten Datensatz angewendet wird, um diesen um zusätzliche Informationen zu erweitern. Dadurch werden die Datensätze, zumindest indirekt, miteinander verbunden. Von besonderer Relevanz ist die Fähigkeit, vollständig unabhängige Datensätze zu verknüpfen, selbst wenn die Individuen aus Datensatz A nicht in Datensatz B vertreten sind.[22]

Fallbeispiel 3:
Zur Veranschaulichung dient das folgende hypothetische Szenario mit zwei unabhängigen Datensätzen: ein Gesundheitsdatensatz, der von einer Krankenkasse zu Forschungszwecken veröffentlicht wurde (siehe Tabelle 5), und ein Datensatz aus einer Lifestyle-Befragung (siehe Tabelle 6).

Im Beispiel wird auf dem Datensatz aus Tabelle 5 ein KI-Modell mit dem Gesundheitsdatensatz trainiert, um basierend auf Attributen wie Alter, Geschlecht, körperlichem Aktivitätslevel und Ernährungsgewohnheiten eine potenzielle medizinische Diagnose vorherzusagen. Dieses trainierte Modell kann anschließend auf den Lifestyle Datensatz (Tabelle 6) angewendet werden, um diesen um ein Attribut “potenzielle Diagnose” zu erweitern. Hierzu kann im Lifestyle Datensatz direkt auf die Attribute Alter und

Tabelle 5 | Auszug aus einem Datensatz von Gesundheitsdaten

	Alter	M/W	Arztbesuche im Jahr	Körperliches Aktivitätslevel	Ernährung	Diagnose
1	50-59	W	6	mittel	fast-food	Herzkrankheit
2	60-69	W	3	niedrig	ausgewogen	keine
3	40-49	W	8	mittel	fast-food	keine
4	60-69	M	2	mittel	fast-food	Typ 2 Diabetes

Tabelle 6 | Auszug eines Datensatzes aus einer Lifestyle-Befragung

	Alter	M/W	PLZ	Schritte pro Tag	Ernährungsprofil
1	40-44	M	68923	5731	kohlenhydratarm
2	55-59	W	45614	13965	kohlenhydratarm
3	65-69	W	71845	11154	proteinreich
4	20-24	M	93921	11110	vegan

Geschlecht zurückgegriffen werden. Das Attribut “Aktivitätslevel” kann abgeleitet werden dem Attribut “Schritte pro Tag”, das Attribut “Ernährungsgewohnheiten” kann aus dem Attribut “Ernährungsprofil” abgeleitet werden. Dieses Verfahren ermöglicht nicht nur eine indirekte Verknüpfung beider Datensätze, sondern beeinflusst auch maßgeblich die Risikobewertung des erweiterten Datensatzes.

3.4 Zwischenfazit

Die vorgestellten Methoden zur Datenverknüpfung repräsentieren lediglich eine Auswahl aus einem breiten Spektrum an verfügbaren Methoden. Unter dem Oberbegriff *Data Linkage* findet sich eine Vielzahl unterschiedlicher Ansätze zur Datenverknüpfung, die teilweise speziell für bestimmte Datentypen oder Datenquellen entwickelt wurden. In der Literatur finden sich zudem zahlreiche Beispiele, die belegen, dass solche Verknüpfungstechniken bereits erfolgreich auf reale Datensätze angewendet wurden.

Der Aufwand bzw. die notwendigen Mittel für die Verknüpfung zweier Datenquellen sind dabei pauschal nicht festzulegen, sondern der Aufwand für die De-Identifikation hängt maßgeblich von den vorhandenen Merkmalen im Datensatz ab. Komplexe Verfahren wie das Training von KI-Modellen oder fortgeschrittene Mustererkennung erfordern oft tiefgehende Kenntnisse in Datenverarbeitung und Künstlicher Intelligenz. Einfachere Ansätze wie das Identifizieren klarer Überschneidungen setzen hingegen nur grundlegende Datenverarbeitungsfähigkeiten voraus. Es sollte jedoch betont werden, dass durch moderne Analysewerkzeuge und Entwicklungen im Bereich der Künstlichen Intelligenz der Aufwand, der betrieben werden muss, um Datensätze zu verknüpfen, tendenziell sinkt. Die Verknüpfbarkeit wird zudem durch die Verfügbarkeit relevanter Daten beeinflusst, wobei mit steigender Datenmenge auch die Verknüpfungsmöglichkeiten zunehmen.

22 Siehe dazu auch Mühlhoff, Prädiktive Privatheit: Kollektiver Datenschutz im Kontext von Big Data und KI, In: Künstliche Intelligenz, Demokratie und Privatheit, 2022, S. 31 ff.

4 Ausblick

Die gezeigten Beispiele erfordern aufgrund des technischen Fortschritts nur einen geringen Aufwand für einen Datenanalysten, um geeignete Datensätze miteinander zu verknüpfen. Der Blick in die Zukunft lässt eine einfachere Bedienbarkeit dieser technischen Möglichkeiten erwarten und vergrößert folglich den Adressatenkreis an potenziellen Anwendern von Software zur De-Anonymisierung. Der Umfang der öffentlichen zugänglichen Datenmengen wie anonymisierte Datensätze oder in sozialen Medien wird steigen und begünstigt damit die Mustererkennung und die Erstellung von Zuordnungstabellen. Die Anonymisierung wird hinsichtlich des Risikofaktors *Data Linkage* zunehmend anspruchsvoller, im Zweifel wird sie kaum noch zu bewerkstelligen sein. Hat die Menge an Open Data einen Schwellwert erreicht, bei der Verknüpfungen über mehrere Datensätze möglich sind, so reicht eine auflösende Verknüpfung als Windstoß, um das Kartenhaus der Anonymisierung einstürzen zu lassen. Daher sind die Entwicklungen der Analysefähigkeiten von Big-Data-Unternehmen und benutzerfreundlichen KIs, die geringe IT-Expertise erfordern, im Auge zu behalten.

Der Beitrag ist im Rahmen des Projekts EAsyAnon – Empfehlungs- und Auditsystem zur Anonymisierung von Daten – entstanden. Gefördert durch das Bundesministerium für Bildung und Forschung (BMBF) – Finanziert durch die Europäische Union (NextGenerationEU).

Funding

Open Access funding enabled and organized by Projekt DEAL.