Ein Glossar zur Datenqualität

Sedir Mohammed[1], Lou Brandner[2], Sebastian Hallensleben[3], Hazar Harmouch[1], Andreas Hauschke[3], Jessica Heesen[2], Stefanie Hildebrandt[3], Simon David Hirsbrunner[2], Julia Keselj[4], Philipp Mahlow[4], Felix Naumann[1], Frauke Rostalski[4], Anna Wilken[4], Annika Wölke[4]

[1] Hasso-Plattner-Institut, Universität Potsdam
[2] Internationales Zentrum für Ethik in den Wissenschaften, Universität Tübingen
[3] VDE
[4] Universität zu Köln

DOI: 10.5281/zenodo.7702426
Version 1.2 (6.3.2023)

Gemeinhin wird **Datenqualität** definiert als die Eignung von Daten für einen bestimmten Anwendungszweck (fitness for use). Diese allgemeine Definition wird üblicherweise in eine Menge an **Datenqualitätsdimensionen** oder -kriterien aufgetrennt. Diese Dimensionen sind teils subjektiv, teils objektiv. Teils können sie automatisch gemessen werden, teils sind sie überprüfbar und teils können sie nur durch Expert*innen bewertet werden.

Das Glossar zur Datenqualität listet alle relevanten Datenqualitätsdimensionen mit einer knappen allgemeinen Definition. Der Fokus liegt auf der Nutzung von Daten für Anwendungen der Künstliche Intelligenz. Die Dimensionen sind nicht orthogonal – ihre Bewertungen können voneinander abhängen; dennoch wirft jede Dimension einen etwas anderen Blick auf die Qualität von Daten. Sofern die Dimensionen oder wortgleiche Begriffe bereits in Gesetzen aufgegriffen wurden, enthält das Glossar einen entsprechenden Hinweis.* Einige klassische Dimensionen aus der Datenqualitätsliteratur (bspw. Objektivität, Compliance) werden hier näher ausdifferenziert oder in mehrere Elemente aufgespalten. In künftigen Ausgaben des Glossars planen wir, systematisch Beispiele hinzuzufügen sowie passende Literatur zu verknüpfen. Hinweise und Fragen zu den Begriffen und deren Definitionen nehmen wir gerne entgegen: felix.naumann@hpi.de

** Es ist zu beachten, dass der hier angeführte KI-Verordnungsentwurf der Kommission sich noch in einem laufenden Gesetzgebungsverfahren befindet. Ob das Gesetz tatsächlich in Kraft tritt, ist dementsprechend nicht gesichert. In jedem Fall wird in der finalen Version mit Änderungen gegenüber dem Entwurf zu rechnen sein.*

Aktualität (timeliness)

Die Lebenswirklichkeit verändert sich fortlaufend: Aktienkurse schwanken, Fußballspiele werden ausgetragen, Produkte werden verkauft und Menschen passen ihre Gewohnheiten ihrer Umwelt neu an. Die Aktualität einer Datenmenge beschreibt den zeitlichen Unterschied zwischen dem elektronisch erfassten Ereignis in der realen Welt und dessen digitaler Repräsentation in der Datenmenge. Änderungen können sich daraus ergeben, dass neue Datensätze erfasst werden (z.B. ein Verkauf), bereits bestehende Datensätze von der Realität überholt werden (z.B. Umzug

eines Kunden) bzw. aus der Löschung von Datensätzen (z.B. Insolvenz eines Unternehmens).

Verwandte Begriffe: Timeliness, up-to-date

Rechtlicher Hintergrund:
Aktualität steht in Art. 5 Abs. 1 lit. d) der DS-GVO in Bezug zur sachlichen Richtigkeit (*Korrektheit*) personenbezogener Daten. Es kann eine Pflicht zur Aktualisierung von personenbezogenen Daten bestehen, wenn dies aus der Perspektive des Datensubjekts mit Blick auf den Zweck der Datenverarbeitung „erforderlich“ ist. (*Roßnagel*/NK-Datenschutzrecht, 1. Auflage 2019, Art. 5 Rn. 141.).

Ansehen (reputation)

Das Ansehen von Daten beschreibt die Vertrauenswürdigkeit der Datenquelle sowie des verarbeitenden Systems. Daten bzw. eine Datenquelle genießen hohes Ansehen, wenn sich diese bereits in der Vergangenheit längerfristig als qualitativ hochwertig erwiesen haben.
Das Ansehen ist niedrig, wenn zu ihnen bisher keine oder nur schlechte Erfahrungen existieren. Insbesondere wenn Datenqualitätsdimensionen wie Fehlerfreiheit nicht adäquat gemessen werden können, stellt das Ansehen eine relevante Dimension dar, welche zugleich als erwartete Qualität aufgefasst werden kann.

Verwandte Begriffe: Reputation, Ruf, Vertrauenswürdigkeit

Ausgewogenheit (balance)

Die Ausgewogenheit von Daten betrachtet die Verteilung der Datenpunkte innerhalb eines Datensatzes. Ein Datensatz ist ausgewogen, wenn die Datenpunkte innerhalb des repräsentierten Wertebereichs im Verhältnis zueinander gleich verteilt sind. In einem ausgewogenen Datensatz, der Kund*innen in Altersgruppen unterteilt, sollten Kund*innen aller Altersgruppen gleich häufig vertreten sein. Dies bedeutet nicht, dass alle Altersgruppen der Gesamtbevölkerung enthalten sein müssen (siehe auch Diversität).

Verwandte Begriffe: Balance, Repräsentativität

Bearbeitbarkeit (ease of manipulation)

Daten sind gut bearbeitbar, wenn Änderungen oder Ergänzungen einfach durchgeführt werden können (Daten in einer Excel-Tabelle vs. Daten auf einer Website). Die Bearbeitbarkeit lässt sich sowohl aus Sicht eines Positivfalls als auch aus Sicht eines Negativfalls betrachten. Zum einen besteht bei einer guten Bearbeitbarkeit die Gefahr, dass die Daten gewollt oder ungewollt verfälscht werden (Negativfall). Andererseits können bei einer guten Bearbeitbarkeit Daten leicht für legitime individuelle Zwecke angepasst werden (Positivfall).

Verwandte Begriffe: Manipulationsfähigkeit, Interoperabilität

Diversität (diversity)

Eine Datenmenge ist divers, wenn jeder Entitätstyp der Gesamtmenge mindestens einmal vorkommt. Das Ziel hierbei ist es, dass die Datenmenge die Vielfalt der Entitätstypen aus der Gesamtmenge widerspiegelt, also alle relevanten Varianten enthält.
Angenommen, eine Mitarbeiterdatenbank (Gesamtmenge) besteht aus männlichen und weiblichen Angestellten, bei der neben dem Geschlecht, ebenso die Abteilung erfasst wurde. Die Datenmenge ist divers, wenn aus allen Abteilungen mindestens eine Mitarbeiterin sowie ein Mitarbeiter enthalten ist.

Verwandte Begriffe: Repräsentativität, Vielfältigkeit

Dokumentation (documentation)

Eine Datenmenge ist gut dokumentiert, wenn sowohl relevante strukturierte Metadaten als auch eine textuelle Beschreibung vollständig und richtig zur Verfügung stehen. Typische Metadaten umfassen den Umfang der Daten, das technische (Datentypen) und semantische (Tabellen- und Spaltennamen) Schema der Daten, Statistiken, Informationen zur Herkunft und zur bisher erfolgten Transformation der Daten. Die textuellen Beschreibungen umfassen unter anderem den Zweck der Daten und die bisherige Nutzung der Daten und können z.B. in sogenannten Data Sheets formalisiert werden.

Verwandte Begriffe: Metadaten, Beschreibung, Transparenz

Eindeutigkeit (uniqueness)

Eine Datenmenge ist eindeutig, wenn jede Entität der realen Welt als höchstens ein Datensatz dargestellt wird: Es existieren keine Duplikate/Dubletten. Beispielsweise kommt ein und derselbe Kunde nur ein einziges Mal in der Kundendatenbank vor - die Kundendatenbank enthält daher keine Dubletten.

Verwandte Begriffe: Duplikatfreiheit, Redundanzfreiheit

Effizienz (efficiency)

Die Effizienz der Daten beschreibt, wie ressourcenschonend verschiedene Prozesse respektive Algorithmen (z.B. Sortieralgorithmen oder Datenanalyseverfahren) auf die Daten angewendet werden können. Ressourcen können etwa Programmieraufwand, Rechenleistung oder Stromverbrauch umfassen. Daten sind demnach effizient, wenn sie leicht nutzbar sind und eine Vielzahl an Prozessen wirksam unterstützen.

Verwandte Begriffe: Wirtschaftlichkeit, Nachhaltigkeit

Fairness

Fairness im KI-Kontext beschäftigt sich mit der Identifizierung, Analyse und Quantifizierung von Verzerrungen (Bias), die Individuen und Gruppen nach bestimmten Merkmalen wie Geschlecht, Ethnie, Behinderung etc. unzulässig diskriminieren. Ein KI-System ist demnach fair, wenn es anhand bestimmter Metriken Standards zur Freiheit von diskriminierenden Verzerrungen erfüllt. Daten im KI-Kontext können nicht grundsätzlich fair sein, aber die Struktur und Qualität von Datenmengen können eine starke Auswirkung auf Verzerrungen und die potenzielle Fairness von KI-Systemen haben. Daten-spezifische Verzerrungen treten beispielsweise aufgrund mangelnder Vielfalt in der Repräsentation sozialer Gruppen und deren Perspektiven zum Beispiel infolge ungeeigneter Auswahl-Methoden auf.

Verwandte Begriffe: Gerechtigkeit, Bias

Rechtlicher Hintergrund:
Laut der englischen Sprachfassung der DS-GVO müssen personenbezogene Daten im Grundsatz „fair“ verarbeitet werden. Jedoch zeigt unter anderem die deutsche Sprachfassung, die „fairness“ mit dem deutschen Rechtsbegriff „Treu und Glauben“ übersetzt, dass der Grundsatz nicht auf Fairness im Sinne von Diskriminierungsfreiheit abzielt. Er dient lediglich als „Auffangbecken“ für Fälle, die von den übrigen Schutzvorschriften der DS-GVO nicht erfasst werden, aber dennoch das von ihr angestrebte Gleichgewicht zwischen den Interessen des Datensubjekts und des Datenverantwortlichen erheblich stören könnten. [Art. 5 Abs. 1 lit. a) Alt. 2 DS-GVO] (Roßnagel/NK-Datenschutzrecht, 1. Auflage 2019, Art. 5 Rn. 47).
Dass der Begriff „Fairness“ im deutschsprachigen juristischen Kontext nicht unbedingt gleich verstanden wird, zeigt auch das Allgemeine Gleichbehandlungsgesetz (kurz: AGG). Dieses beschäftigt sich zwar ebenfalls mit dem Schutz vor Diskriminierungen (bei bestimmten Merkmalen, vgl. § 1 AGG). Jedoch wird hier nicht der Begriff „Fairness“ verwendet, sondern es wird von dem Verhindern oder Beseitigen von „Benachteiligungen“ gesprochen. Auch in den englischen Versionen der europäischen Richtlinien zu Diskriminierung (z.B. COUNCIL DIRECTIVE 2000/43/EC) wird nicht der Begriff „Fairness“ benutzt, sondern der Begriff „discrimination“. Wichtig ist an dieser Stelle, dass es zulässige und unzulässige Formen der Diskriminierung gibt. Eine Benachteiligung iSd §3 Abs.1, 2 AGG liegt vor, wenn eine Person wegen eines bestimmten Merkmals (vgl. §1 AGG) eine weniger günstige Behandlung erfährt, als eine andere Person in einer vergleichbaren Situation erfährt, oder wenn dem Anschein nach neutrale Vorschriften Personen wegen bestimmter Merkmale benachteiligen können. Unzulässig werden diese Benachteiligungen erst dann, wenn sie nicht gerechtfertigt sind. Eine Rechtfertigung einer Diskriminierung kann sich zum Beispiel ergeben, wenn durch die Benachteiligung ein sachliches Ziel verfolgt wird, und die gewählten Mittel zur Erreichung dieses Ziels erforderlich und angemessen sind.

Genauigkeit (precision)

Zum einen ist die Genauigkeit ein Maß dafür, wie hoch der Grad der Varianz wiederholt aufgenommener Daten ist. Dies ist streng von der Korrektheit der Daten zu unterscheiden (siehe dort). Ein Beispiel für eine hohe Genauigkeit ist die Messung eines Vitalparameters einer Patientin in einem Krankenhaus exakt alle 120 Sekunden. Von diesen zeitlich genau aufgenommenen Daten lässt sich jedoch nicht auf die Korrektheit der Daten schließen, etwa

Messfehler. Zum anderen wird die Genauigkeit als Maß für den Detailgrad von Informationen genutzt. So kann in einem Formular nach dem Geburtsjahr oder nach dem genauen Geburtsdatum gefragt werden. Ein Indikator für Genauigkeit ist die Zahl der Nachkommastellen bei gemessenen Werten. Ein dritter Aspekt der Genauigkeit ist die Präzision der vordefinierten Werteklassen, z.B. bei der Angabe von Hautfarbe, bei der zwischen braun und schwarz unterschieden wird.

Verwandte Begriffe: Präzision, Exaktheit

Glaubwürdigkeit (credibility)

Die Glaubwürdigkeit von Daten ist geprägt durch vielfältige Charakteristiken. Dies beinhaltet psychologische Marker (unbewusste Annahmen, Stereotypen), soziale Marker (anerkannte Titel und Institutionen, Quellenautorität), ästhetische Marker (vertraute Darstellungskonventionen im Design von Diagrammen oder einer Webseite) und kognitive Marker (Lesbarkeit, Expertise des Betrachters). Glaubwürdigkeit ist in der Regel eine kontextbezogene und teils subjektive Qualität von Daten, Informationen und deren Ursprung. Sie kann entsprechend nicht ausschließlich als statistische Größe dargestellt werden.

Verwandte Begriffe: Vertrauenswürdigkeit, Reputation, Autorität

Konsistente Darstellung (consistent representation)

Eine Datenmenge (z.B. Tabelle) ist in ihrer Darstellung konsistent, wenn kein Attribut (Spalte) zwei oder mehr eindeutige Werte hat, die semantisch gleichwertig sind (z.B. New York vs. NYC oder 1/12/2022 vs. 12.1.2022). Dies schließt auch exakte Duplikate/Dubletten ein.

Verwandte Begriffe: Einheitliche Darstellung, Categorical duplicates

Konsistenz (consistency)

Eine Datenmenge (z.B. Tabelle) ist konsistent, wenn alle Bedingungen, die an den Zustand der Datenmenge gestellt werden, erfüllt sind. Konsistenz umfasst Integritätsbedingungen an die Daten wie Datentypen oder Wertebereiche, Abhängigkeiten oder auch Beziehungen über verschiedene Datenmengen hinweg. Beispiele für mangelnde Konsistenz sind unterschiedliche Datumsformate in einer Spalte, verschiedene Städte bei gleicher Postleitzahl, oder Bestellungen mit ungültigen Kundennummern.

Verwandte Begriffe: Integrität, Nebenbedingungen, Widerspruchsfreiheit, Validität

Korrektheit (free of error, accuracy)

Korrektheit beschreibt die Übereinstimmung zwischen einem Phänomen in der Welt und dessen Beschreibung als Daten. Beim Vergleich des Datenwertes mit dem empirisch feststellbaren Wert kann der Unterschied entweder binär bestimmt werden (gleich oder ungleich) oder man kann

mittels eines Ähnlichkeitsmaßes den Grad des Unterschieds bestimmen (z.B. als Ähnlichkeit zwischen 0 und 1). Die Qualität der Korrektheit spielt insbesondere bei Daten eine Rolle, deren sachliche Richtigkeit abschließend geklärt werden kann und deren Bedeutung nicht ambivalent ist.

Verwandte Begriffe: Fehlerfreiheit, Richtigkeit

Rechtlicher Hintergrund:
Art. 10 Abs. 3 S. 1 KI-VO-E greift die Korrektheit von Daten auf. Gesprochen wird von „fehlerfreien" Daten, im englischen Originalwortlaut verwendet die Kommission den Ausdruck: „free of errors". Die Europäische Kommission legt den juristischen Begriffen des KI-VO-E dabei gängige informatische Definitionen zugrunde.

Im Umgang mit personenbezogenen Daten ist die Gewährleistung von „Korrektheit" durch die DS-GVO in Art. 5 Abs. 1 lit. d vorgeschrieben. In der englischen Sprachfassung wird der Begriff „accurate" verwendet, in der deutschen „sachliche Richtigkeit".

Kosten (cost)

Die Kosten für einen Datensatz umfassen sowohl die monetären Kosten, die bei der Erzeugung oder Beschaffung und der dauerhaften Speicherung der Daten entstehen, als auch die Personalkosten zur Beschaffung und Aufbereitung der Daten. Kosten können für eine gesamte Datenmenge berechnet werden, oder sie fallen pro Anfrage an die Daten an.

Verwandte Begriffe: Preis

Beispiele: Annotieren von Daten durch Expertinnen oder Crowd-Worker; Einkauf der Daten bei Data Broker; Speicherung in der Cloud; Personalkosten z.B. zur Bereinigung, Umformatierung usw.

Portabilität (portability)

Portabilität beschreibt die Fähigkeit, strukturierte Daten zuverlässig und sicher von einem Computer zu einem anderen zu übertragen. Das einfache Übertragen von persönlichen Daten von einem sozialen Netzwerk auf einen externen Datenspeicher wäre ein Beispiel für eine gute Portabilität. Portable Daten sind nach gängigen Standards formatiert.

Verwandte Begriffe: Übertragbarkeit, Transferierbarkeit

Privatheit (privacy)

Daten sind privat, wenn die in den Daten beschriebenen Personen Kontrolle über und Zugriff auf diese Daten haben. Private (auch vertrauliche) Daten wahren das Recht auf informationelle Selbstbestimmung der Nutzer*innen. Die rechtliche Absicherung von Privatheit kann organisatorisch und technisch sichergestellt werden. Zur organisatorischen Herstellung der Privatheit dienen Einwilligungserklärungen durch Nutzer*innen, die die gesamte Nutzung der

Daten untersagen oder Anweisungen zu deren Gebrauch (teilweise Löschung, Verarbeitung etc.) enthalten können. Zur technischen Herstellung der Privatheit können die Daten beispielsweise verschlüsselt oder der physische Zugriff auf das Speichermedium begrenzt werden.

Verwandte Begriffe: Vertraulichkeit, Privatsphäre, Datenschutz

Rechtlicher Hintergrund:
Privatheit als solche ist keine konkrete gesetzliche Anforderung, sondern als Ausdruck des Grundrechts auf informationelle Selbstbestimmung bzw. Schutz der Privatsphäre das Schutzziel des Datenschutzrechts. Insbesondere die DS-GVO trägt mit ihrem Rechte- und Pflichtenkatalog zur Wahrung der Privatheit personenbezogener Daten bei.

Relevanz (relevancy)

Daten sind relevant, wenn sie aus Sicht der Anwenderin bzw. für den Zweck einer Anwendung notwendige Informationen enthalten und somit zur konkreten Realisierung eines Ziels beitragen. So sind beispielsweise in einem Online-Shop der Artikelname und der Preis für die Vergleichbarkeit von Produkten relevant. Die Anzahl der beteiligten Personen zur Herstellung einzelner Produkte hat dagegen, abhängig vom Anwendungsfall, mitunter eine geringe Relevanz.
Im Kontext des Datenschutzes dürfen Daten nur aus Einträgen bestehen, erhoben und verarbeitet werden, wie sie für den jeweiligen Zweck erforderlich sind. Hierbei gilt, dass diese in einem den geltenden Rechtsvorschriften konformen Zustand bzw. Ausmaß vorliegen. So darf mit dem Ziel der Datensparsamkeit bei der Anmeldung für einen Newsletter nicht die Adresse eines Kunden eine verpflichtende Angabe sein.

Verwandte Begriffe: Adequacy, Erheblichkeit

Rechtlicher Hintergrund:
Art. 10 Abs. 3 S. 1 KI-VO-E setzt voraus, dass verwendete Trainingsdaten für Hochrisikosysteme mit relevanten Trainings-, Test- und Validierungsdaten konfrontiert werden. Der englischsprachige Originaltext verwendet „relevant“. Die Europäische Kommission legt den juristischen Begriffen des KI-VO-E dabei gängige informatische Definitionen zugrunde.

Der datenschutzrechtliche Grundsatz der „Datenminimierung“ verlangt gem. Art. 5 Abs. 1 lit. c DS-GVO, dass personenbezogene Daten „dem Zweck angemessen“ und „erheblich“ sein müssen, was in der englischen Fassung mit „adequate“ und „relevant“ bezeichnet wird.

Repräsentativität (representativity)

Eine Datenmenge ist statistisch repräsentativ, wenn jede Entität der zu repräsentierenden Gesamtmenge die gleiche Chance hat, in der Datenmenge repräsentiert zu sein. So sind die relativen Verteilungen der charakteristischen Eigenschaften der Gesamtmenge in der Datenmenge wiederzufinden. Eng verwandt mit der statistischen Repräsentativität, aber dennoch separat zu betrachten, sind die Dimensionen Ausgewogenheit (siehe dort) und Diversität (siehe dort).

Beispielsweise bestehe ein Datensatz (Gesamtmenge) aus 70 männlichen und 30 weiblichen Studierenden. Von den männlichen Studierenden sind 40 im Fach Kunst und 30 im Fach Geschichte eingeschrieben. Im Vergleich studieren 15 der weiblichen Studierenden Kunst und die übrigen 15 Geschichte. Auf Basis der zuvor genannten Gesamtmenge wäre eine Datenmenge repräsentativ, wenn diese aus jeweils 9 weiblichen Kunst- und Geschichtsstudierenden besteht, sowie aus 24 männlichen Kunst- und 18 männlichen Geschichtsstudierenden. Diese Datenmenge ist statistisch repräsentativ, da die relativen Verhältnisse zwischen den Studierenden des gleichen Geschlechts, als auch die Verhältnisse zwischen den Studierenden des gleichen Studienfachs im Vergleich zur Gesamtmenge identisch sind.

Rechtlicher Hintergrund:
In Art. 10 Abs. 3 S. 1 des KI-Verordnungsentwurfs der Europäischen Kommission wird Repräsentativität als Anforderung an Datensätze in Hochrisikosystemen genannt. Eine Legaldefinition wird nicht gegeben. Ergänzend sieht jedoch Art. 10 Abs. 3 S. 2 vor, dass die verwendeten Datensätze „die geeigneten statistischen Merkmale, gegebenenfalls auch bezüglich der Personen oder Personengruppen, auf die das Hochrisiko-KI-System bestimmungsgemäß angewandt werden soll", aufweisen. Auch müssen die Datensätze „soweit dies für die Zweckbestimmung erforderlich ist, den Merkmalen oder Elementen entsprechen, die für die besonderen geografischen, verhaltensbezogenen oder funktionalen Rahmenbedingungen, unter denen das Hochrisiko-KI-System bestimmungsgemäß verwendet werden soll, typisch sind", Art. 10 Abs. 4 KI-VO-E.

Unklar ist, ob ein Datensatz repräsentativ i.S.d. KI-VO-E ist, wenn er schlicht die Verhältnisse der real zu repräsentierenden Gesamtmenge, wie oben im Beispiel dargestellt, wiedergibt. Denn dabei könnte es dazu kommen, dass die Datenmenge für Randgruppen nicht ausreicht, um das KI-System für die Verwendung durch Zugehörige ebendieser zu trainieren (vgl. **Umfang**). Sinn und Zweck der Repräsentativitätsanforderung ist es jedoch gerade sicherzustellen, dass das trainierte KI-System für all diejenigen Bevölkerungsgruppen funktioniert, auf die es zweckgemäß angewendet werden soll. Auch im Lichte der Konkretisierungen durch Art. 10 Abs. 3 S. 2, Abs. 4 KI-VO-E spricht mithin viel dafür, zumindest für menschenbezogene Anwendungsgebiete ein Kriterium der Erfolgswahrscheinlichkeit zu fordern, welches erfüllt ist, wenn der Datensatz die Verhältnisse der entsprechenden Personengruppen richtig wiedergibt, zumindest aber genügend Daten für eine einzelne Gruppierung enthält, um das System mit hinreichender Wahrscheinlichkeit für die Nutzung durch Angehörige dieser kleinen Gruppen zu trainieren.

Rückverfolgbarkeit (traceability)

Die Rückverfolgbarkeit beschreibt die Fähigkeit, die Herkunft der Daten sowie alle Transformationen, die auf Daten ausgeführt wurden, nachzuvollziehen. Somit ist einsehbar, welcher Datenquelle die Daten entstammen und welche Änderungen auf ihnen durchgeführt wurden. Damit können Daten z.B. in einen vorherigen Zustand versetzt werden, sodass entweder die aktuelle Version ersetzt wird oder verschiedene Versionen der Daten parallel existieren.

Verwandte Begriffe: Provenance, Lineage

Sicherheit (security)

Die Sicherheit von Daten beschreibt den zu jedem Zeitpunkt existierenden Schutz vor unautorisierten Zugriff auf die Daten oder deren Diebstahl oder Beschädigung. Systeme müssen eine korrekte Zugriffsverwaltung garantieren. Zur Aufrechterhaltung dieser Garantie ist daher auch die Funktionssicherheit eines Systems relevant, sodass bei einem Funktionsausfall das System trotzdem in einen definierten Zustand übergeht, bei dem die Sicherheit der Daten gewährleistet wird. Beispielsweise sollte ein*e Kund*in eines Online-Shops nur Zugriff auf ihre zuvor getätigten Bestellungen erhalten und nicht auf die Verkaufszahlen aller Produkte. Daten sollen stets vor Hackerangriffen geschützt werden, bei dem die Angreifer*innen die Daten stehlen oder verschlüsseln könnten.

Verwandte Begriffe: Datenschutz

Transparenz (transparency)

Fragen der Transparenz betreffen Offenlegungspflichten in Bezug auf die Herkunft der Daten, mit denen Modelle trainiert werden, Informationen zu den Qualitätsprüfungen, denen Datensätze unterzogen wurden, wer die Datensätze gelabelt hat, welche Lernziele verfolgt werden, ob und inwiefern Quellcode eingesehen werden kann und vieles mehr. Transparenz ermöglicht, dass von technischen Systemen betroffene Personen zu informierten Entscheidungen kommen können und dass Rechtsverletzungen identifiziert und korrigiert werden. Transparenz ermöglicht zudem gesellschaftliche Debatten und den Aufbau von Vertrauensverhältnissen.

Verwandte Begriffe: Interpretierbarkeit, Zugänglichkeit, Dokumentation

Umfang (appropriate amount of data)

Der Umfang von Daten beschreibt die Größe des Datensatzes und kann z.B. als die Anzahl der Byte oder die Anzahl der Zeilen gemessen werden. Der Umfang von Daten kann zu gering oder zu groß sein. Zum Beispiel ist ein gewisser Umfang an Trainingsdaten nötig, um ein KI-Modell angemessen zu trainieren. Umgekehrt kann ein zu hoher Umfang an Daten, z.B. eine unnötig hohe Auflösung von Bilddateien, zu Problemen bei der Datenverwaltung führen.

Verwandte Begriffe: Conciseness

Übersichtlichkeit (concise representation)

Die Übersichtlichkeit betrachtet die Darstellungsform von Daten. Übersichtliche Daten sind abhängig vom Verwendungszweck geeignet und deutlich erkennbar dargestellt.
Ein Beispiel hierfür sind Aktienkurse, welche in einem Liniendiagramm dargestellt werden, sodass die Kursentwicklungen über einen bestimmten Zeitraum übersichtlich einsehbar sind. Durch die Einbeziehung zu vieler Aktienverläufe sowie durch die Betrachtung zu großer Retrospektiven kann jedoch die Lesbarkeit eingeschränkt werden – die Darstellung wirkt unübersichtlich.

Verwandte Begriffe: Übersichtliche Darstellung, Kompaktheit

Verlässlichkeit (reliability)

Verlässlichkeit bezeichnet den Zustand, bei dem Daten, ohne von Vorurteilen bestimmt zu sein, erfasst wurden. Unabhängig vom Datenerfasser müssen die erfassten Daten bei gegebener Verlässlichkeit immer identisch sein.

Verwandte Begriffe: Reliability, Objektivität

Verständlichkeit (understandability)

Daten sind in einem verständlichen Zustand, wenn sie durch die Anwender*innen unmittelbar verstanden werden können. Verständliche Daten haben ein aussagekräftiges Schema und verwenden typische Datenformate. Ein Beispiel hierfür ist es, wenn in einem Online-Shop zu einem Artikel der vollständige Artikelname hinterlegt ist, sodass der Kunde unmittelbar erkennen kann, um was für einen Artikel es sich handelt. Dagegen ist die Verständlichkeit beeinträchtigt, wenn statt des Artikelnamens ausschließlich die Artikelnummer hinterlegt ist.

Verwandte Begriffe: Explainability, Lesbarkeit, Interpretierbarkeit

Vollständigkeit (completeness)

Vollständigkeit ist das Verhältnis zwischen der Menge der repräsentierten Daten und der Menge der zu repräsentierenden Daten. Während ersteres gezählt werden kann (Anzahl Zeilen, Anzahl nicht-null Werte), kann letzteres oft nur geschätzt werden.
Eine Datenmenge (z.B. Tabelle) ist in Bezug auf eine Domäne vollständig, wenn jede Entität der Domäne in der Datenmenge repräsentiert ist. Ein Datensatz (z.B. Zeile) ist vollständig, wenn für jedes Attribut (Spalte) ein Wert vorhanden ist.

Verwandte Begriffe: Fehlende Werte

Rechtlicher Hintergrund:
Art. 10 Abs. 3 S. 1 KI-VO-E sieht vor, dass Trainings-, Test- und Validierungsdaten vollständig sein müssen. Der englische Originalwortlaut spricht von „complete“. Die Europäische Kommission legt den juristischen Begriffen des KI-VO-E dabei gängige informatische Definitionen zugrunde.

Wertschöpfung (value-added)

Die Wertschöpfung von Daten beschreibt, dass Daten in einer Anwendung nützlich verwertet werden können. Die Verwertung von Daten innerhalb der Anwendung ist dann nützlich, wenn dadurch ein Gewinn (monetär, Erkenntnisse) für die Dateneigentümer*innen oder die Nutzer*innen der Anwendung entsteht.

Wiederherstellbarkeit (recoverability)

Daten sind wiederherstellbar, wenn sie redundant und sicher z.B. auf einem separaten Datenträger gespeichert werden. Die Wiederherstellbarkeit von verlorenen Daten zum Beispiel bei Systemfehlern oder Datenträgerverlust ist auf diese Weise gewährleistet. Neben der Wiederherstellbarkeit selbst kann diese Dimension auf Basis der Zeit gemessen werden, die benötigt wird, um die Daten vollständig zurückzugewinnen.

Verwandte Begriffe: Fehlererholung, Wiederbeschaffbarkeit

Zugänglichkeit (accessibility)

Die Zugänglichkeit hat technische, organisatorische, finanzielle und rechtliche Komponenten. Technische Zugänglichkeit stellt zu jedem Punkt der Verarbeitung ausreichende Ressourcen (Computer, Netzwerk) sicher, um einen reibungslosen und schnellen Zugriff zu ermöglichen. Organisatorische Zugänglichkeit erlaubt Nutzerinnen und Nutzern ohne technische Kenntnisse oder Technologien einen leichten Zugriff auf die Daten. Die rechtliche Zugänglichkeit ergibt sich durch eine Lizenzierung rechtlich geschützter Datensätze, welche eine Weiternutzung der Daten ermöglicht, während eine finanzielle Zugänglichkeit durch angemessene oder entfallende Nutzungsgebühren erreicht werden kann.

Verwandte Begriffe: Antwortzeit, Latenz, Barrierefreiheit, Offenheit, Verfügbarkeit, Auffindbarkeit

---

Vorgeschlagene Zitationsweise:
Sedir, Mohammed, Lou Brandner, Sebastian Hallensleben, Hazar Harmouch, Andreas Hauschke, Jessica Heesen, Stefanie Hildebrandt, Simon David Hirsbrunner, Julia Keselj, Philipp Mahlow, Felix Naumann, Frauke Rostalski, Anna Wilken und Annika Wölke (2023): Glossar Datenqualität. Version 1.2 (6.3.2023), doi: 10.5281/zenodo.7702426.

Dieses Glossar entstand im Rahmen des Forschungsprojekts KITQAR (KI-Test- und Trainingsdatenqualität in der digitalen Arbeitsgesellschaft).

Projektpartner:

Verband der Elektrotechnik, Elektronik und Informationstechnik (VDE); Lehrstuhl für Informationssysteme am Hasso-Plattner-Institut (HPI)/Universität Potsdam; Lehrstuhl für Strafrecht, Strafprozessrecht, Rechtsphilosophie und Rechtsvergleichung/Universität zu Köln; Internationales Zentrum für Ethik in den Wissenschaften (IZEW)/Universität Tübingen

Gefördert durch:

Bundesministerium
für Arbeit und Soziales

DENK FABRIK
DIGITALE ARBEITS-GESELLSCHAFT
im Bundesministerium für Arbeit und Soziales

aufgrund eines Beschlusses
des Deutschen Bundestages