DATENQUALITÄT UND QUALITÄTSMETRIKEN IN DER DATENWIRTSCHAFT

GRUNDLAGEN, PRAXIS, HANDLUNGSEMPFEHLUNGEN

Eine Studie im Auftrag des Bundesministeriums für Wirtschaft und Klimaschutz von der Begleitforschung zum Technologieprogramm „Smarte Datenwirtschaft"

Smarte Datenwirtschaft

IMPRESSUM

Die Studie wurde im Auftrag des Bundesministeriums für Wirtschaft und Klimaschutz im Rahmen der Begleitforschung zum Technologieprogramm „Smarte Datenwirtschaft“ erstellt.

AUTOR:INNEN

Marieke Rohde,
Marlene Eisenträger,
Nicole Wittenbrink,
Sebastian Straub,
Peter Gabriel

HERAUSGEBER

Peter Gabriel
Begleitforschung Smarte Datenwirtschaft
Institut für Innovation und Technik (iit)
in der VDI / VDE Innovation + Technik GmbH
Steinplatz 1
10623 Berlin
gabriel@iit-berlin.de

VERÖFFENTLICHUNG

Oktober 2022

GESTALTUNG

LHLK Agentur für Kommunikation GmbH
Hauptstraße 28
10827 Berlin

BILDER

Ilya, Vdant85 – stock.adobe.com (Titel)

INHALT

WENN DATEN ZUM WIRTSCHAFTSGUT WERDEN, IST DEREN QUALITÄT VON WESENTLICHER BEDEUTUNG. DOCH WIE LÄSST SICH DIESE BEURTEILEN UND GEWÄHRLEISTEN?

DIE KURZSTUDIE BELEUCHTET DEN HEUTIGEN STAND VON THEORIE UND PRAXIS ZUM UMGANG MIT DATENQUALITÄT IN DER DATENWIRTSCHAFT.

EXECUTIVE SUMMARY

Die Optimierung interner Betriebsprozesse oder das Angebot datenbasierter Dienste – an solchen Wertschöpfungsprozessen mit digitalen Daten sind oftmals mehrere Unternehmen beteiligt, die Daten tauschen oder miteinander handeln. Dazu zählen unter anderem Datenlieferanten, technische Dienstleister, Plattformbetreiber und Käufer von Informationsdiensten. In der so entstehenden eigenständigen Datenwirtschaft spielt die Datenqualität eine zentrale Rolle für das Wirtschaftsgut Daten.

Diese Kurzstudie beleuchtet den heutigen Stand von Theorie und Praxis zum Umgang mit Datenqualität in der Datenwirtschaft. Sie wurde im Rahmen der Begleitforschung zum Technologieprogramm „Smarte Datenwirtschaft" des Bundesministeriums für Wirtschaft und Klimaschutz (BMWK) erstellt. Die Studie fasst zunächst die Fachdiskussion zu Datenqualität und Qualitätsmetriken in der Datenwirtschaft zusammen, inklusive der rechtlichen Rahmenbedingungen. Anschließend betrachtet sie anhand von Interviews mit Vertretern deutscher Unternehmen und Forschungseinrichtungen den Umgang mit Datenqualität in deutschen Unternehmen und beschreibt zu erwartende Trends. Auf dieser Grundlage werden Handlungsempfehlungen für Wirtschaft und Wissenschaft in drei Handlungsfeldern abgeleitet:

ANWENDUNG ETABLIERTER KONZEPTE

Die Experten und die Fachliteratur beschreiben ein starkes Gefälle zwischen Anwendungsbereichen mit bereits etablierten Datenmärkten, wie etwa die für Adressdaten, Wetterdaten und Bilddaten, und Anwendungsbereichen, in denen die Datenwirtschaft noch am Anfang steht, wie etwa die Wertschöpfung mit Sensordaten in der industriellen Produktion oder mit Lieferkettendaten in der Logistik. Um auch solche Anwendungsbereiche für die Datenwirtschaft zu erschließen, ist es nach Aussagen der Interviewpartner notwendig, allgemeine Konzepte zum Umgang mit Datenqualität in konkreten Datenwertschöpfungsketten anzuwenden und die Anwendung dann branchenweit zu vereinheitlichen.

Da die Qualität von Daten deren wirtschaftlichen Wert maßgeblich beeinflusst, kommt Metriken für Datenqualität die wichtige Aufgabe zu, Vergleichbarkeit und Markttransparenz in der Datenwirtschaft herzustellen. Solche Metriken, etwa zur Messung von Qualitätsdimensionen wie Korrektheit, Aktualität, Konsistenz oder Vollständigkeit, sind in der theoretischen Fachliteratur etabliert. In der Praxis leiten die Anforderungen eines konkreten Anwendungsfalls die Auswahl solcher Metriken. Ihre Messung kann in einigen Fällen nur bei oder unmittelbar nach der Erhebung der Daten erfolgen.

- Für eine effiziente Anwendung von Datenqualitätsmetriken müssen alle Akteure in der jeweiligen Datenwertschöpfungskette zusammenarbeiten. Datengebende Unternehmen sowie auch öffentliche Anbieter von Open Data wie Behörden und Forschungseinrichtungen sollten nutzbar und nachvollziehbar aufbereitete Daten sowie Metriken der anwendungsfallunabhängigen („inhärenten") Datenqualität bereitstellen. Datennutzende Unternehmen sollten anwendungsbezogene Anforderungen sowie Metriken des wirtschaftlichen Mehrwerts beisteuern.
- Um Rechtssicherheit im Haftungsfall zu erhalten, sollten datengebende und datennehmende Unternehmen vertragliche Vereinbarungen zur Beschaffenheit von Daten treffen. Der diesbezüglich bestehende vertragsrechtliche Gestaltungsspielraum bei Datenüberlassungen wird den Aussagen der Experten nach derzeit nicht ausgeschöpft.
- Um branchenweite Standards für Datenqualität zu etablieren, sollten Branchenverbände für in ihren Bereichen relevante Anwendungen Metriken für Datenqualität zusammenstellen und die dazugehörigen Messverfahren spezifizieren sowie auch Standardverträge für die Datenüberlassung entwickeln und bereitstellen.
- Um Synergien zwischen verschiedenen Anwendungsbereichen des Datenqualitätsmanagements zu heben und Metriken mit übergreifender Relevanz zu identifizieren, sollten Wirtschaft und Wissenschaft eine branchenübergreifende Dialogplattform zum Umgang mit Datenqualität aufbauen.

KONZEPTIONELLE WEITERENTWICKLUNGEN

Existierende Rahmenkonzepte für Datenqualität sind in der Regel auf eine interne Verwendung und fortlaufende Pflege von Datenbeständen ausgelegt. In der Datenwirtschaft hingegen sind besonders die Nutzbarkeit eines Datensatzes als alleinstehendes Wirtschafts- oder Handelsgut sowie eine transparente Darstellung der Datenqualität wichtig, um den Tausch mit Dritten anbahnen und durchführen zu können. Der Literatur und den Experten zufolge rückt dabei die sogenannte pragmatische, bereits an einer konkreten Datennutzung ausgerichtete Datenqualität vermehrt in den Fokus, die in der Fachliteratur bisher verhältnismäßig wenig Aufmerksamkeit erhalten hat.

Allerdings werden in der Datenwirtschaft Daten häufig auch für andere Zwecke genutzt als ursprünglich geplant („Repurposed Data"). Die nutzungsbezogenen Qualitätsanforderungen sind dann nicht im Vorfeld bekannt. Nachträgliche Prüfungen der Datenqualität werden in diesen Fällen häufig dadurch erschwert, dass Unternehmen aus Sicherheitsbedenken nur Metadaten

herausgeben oder einen eingeschränkten Zugriff auf ihre Daten erlauben, zulasten der Nachvollziehbarkeit von Daten gehen kann.

- Um auf solche Besonderheiten einzugehen, sollten Branchen- und Fachverbände ein eigenes Rahmenkonzept für die Datenqualität in der Datenwirtschaft entwickeln. Dieses sollte sich am Dictionary of Data Quality Dimensions (3DQ, DAMA NL) orientieren, einem erfolgversprechenden Ansatz zur Harmonisierung bestehender Rahmenkonzepte.
- Wirtschaft und Wissenschaft sollten die Nutzung von Metriken für subjektiv wahrgenommene, pragmatische Qualitätsdimensionen wie die Relevanz und Glaubwürdigkeit von Daten in der Praxis etablieren. Für eine Messung kommen unter anderem Reputationssysteme, wie sie im E-Commerce üblich sind, in Betracht.
- Zur Einschätzung der rechtssicheren Verwendbarkeit von Daten – ebenfalls eine pragmatische Qualitätsdimension, die aber nicht durch Metriken gemessen werden kann – sollten Auditierungsverfahren entwickelt werden, die die gesetzlichen Vorschriften zum Datenschutz, zum Urheberrecht und zum Schutz von Geschäftsgeheimnissen berücksichtigen.
- Um die Nutzung von Daten für nicht-vorhergesehene Zwecke zu erleichtern, sollten Datengebende und Betreibende von Datenplattformen ihre Datenangebote so präsentieren, dass auf variable Qualitätsanforderungen eingegangen werden kann, z. B. durch parametrisierbare Angebote und Preismodelle im Datenhandel.
- Anbietende von Diensten zur Datenbewertung sollten die technischen Möglichkeiten dafür schaffen, dass Datenprüfungen on-premise auf den Systemen der datengebenden Unternehmen durchgeführt werden können. Sie sollten dabei auf den Ansätzen der International Data Spaces Association und von GAIA-X aufbauen.

UMFELD

Sowohl die Experten als auch die Fachliteratur sehen Hemmnisse, die im Umfeld der Datenwirtschaft einem effizienten Umgang mit Datenqualität entgegenstehen. In der Breite der Unternehmen wird häufig noch Silo-Datenhaltung betrieben, Standards für Datenformate und Schnittstellen werden nicht beachtet und die Datenqualität als Belang wird vernachlässigt. Außerdem mangelt es potenziell datengebenden Unternehmen an Vertrauen, an Kenntnis über den wirtschaftlichen Wert von Daten und Data Sharing (dem Teilen von Daten) und an Datenkompetenz (Data Literacy) bei Domänenexperten und Management gleichermaßen.

- Für besonders relevante oder sicherheitskritische Datenverwendungen sollte ein Zertifizierungssystem für Datenqualität aufgebaut werden.
- Um die Möglichkeit des Datenteilens und die Vergleichbarkeit von Datensätzen in digitalen Ökosystemen zu fördern, sollten Wissenschaft, Fachverbände und Normungsgremien weiter an der Etablierung von für die Datenwirtschaft geeigneten Standards für Datenformate und Nutzungsschnittstellen arbeiten.
- Um intern eine Datenkultur zu schaffen, sollten Unternehmen bei ihren Führungskräften und Mitarbeitenden die Bedeutung der Datenwirtschaft, Grundkenntnisse zu datenwirtschaftlichen Anwendungen und die Wichtigkeit von Datenqualität vermitteln und verankern.

01

1 EINLEITUNG

Digitale Daten nehmen in der Wertschöpfung von Unternehmen eine zentrale Rolle ein. Sie sind eine wichtige Grundlage für die Entscheidungsfindung in allen betrieblichen Prozessen: von der Produktentwicklung über die Produktion bzw. das Erbringen einer Dienstleistung bis hin zu Marketing und Vertrieb. Das gilt insbesondere für selbstlernende KI-Systeme, die zunehmend in Produkten und Diensten integriert werden oder mit denen Unternehmensprozesse gesteuert werden. In Assistenzsystemen sowie in der Optimierung und Automatisierung von Geschäfts-, Produktions- oder Logistikprozessen fließen vermehrt Verfahren des Maschinenlernens ein. Jedoch stehen auch für herkömmliche Entscheidungsprozesse, z. B. in der Markt- oder Wettbewerbsbeobachtung, durch die fortschreitende Digitalisierung neue Datenbestände als wertvolle Ressource für den Wertschöpfungsprozess zur Verfügung. Damit werden Daten zum Wirtschaftsgut mit einem ökonomischen Wert, das in einer eigenen Datenwertschöpfungskette gesammelt, aufbereitet, konsolidiert und bereitgestellt wird.

In vielen Anwendungsfällen sind die erforderlichen Daten aber gar nicht im Unternehmen selbst vorhanden. Für das teilautonome Fahren müssen etwa hochpräzise Straßenkarten von spezialisierten Dienstleistern bezogen werden. Der Hersteller einer Maschine kann nur eine vorausschauende Wartung seiner Produkte mittels KI anbieten, wenn er die Betriebsdaten seiner Kundinnen und Kunden zum Anlernen der Algorithmen nutzt. Ebenso werden technische Dienstleistende für die Aufbereitung und Analyse von Daten sowie Betreibende von Datenplattformen in die Wertschöpfung eingebunden. Data Sharing über Unternehmensgrenzen hinweg ist damit oft Voraussetzung für die datengetriebene Wertschöpfung. Vielfach werden Daten auch zum Handelsgut, das entweder in bilateralen Geschäftstransaktionen oder auf Datenhandelsplattformen ausgetauscht wird.

DER TAUSCH DES WIRTSCHAFTSGUTS DATEN MIT DRITTEN SETZT VORAUS, DASS SEINE BESCHAFFENHEIT UND GÜTE ALLEN SEITEN IN AUSREICHENDEM MASSE BEKANNT SIND.

Der Tausch des Wirtschaftsguts Daten mit Dritten setzt voraus, dass seine Beschaffenheit und Güte allen Seiten in ausreichendem Maße bekannt sind. Nur dann kann das datengebende Unternehmen den Wert seines Angebots für andere begründen und nur dann kann das datennehmende Unternehmen eine informierte Entscheidung zum Nutzen des Angebots für seine eigene Wertschöpfung treffen. Das gilt noch einmal stärker, wenn das Handelsgut Daten gegen ein Entgelt ausgetauscht wird. Wenn beide Seiten dieselben Informationen zum gehandelten Gut haben, kann dies einen effizienten Ausgleich von Angebot und Nachfrage über den Kaufpreis begünstigen.

Damit kommt der Bestimmung der Datenqualität eine zentrale Rolle in der Datenwirtschaft zu. Für den Datenhandel wurde in den letzten Jahren sogar mehrfach die Zielstellung formuliert, Abschätzungen des wirtschaftlichen Datenwerts direkt an Messungen der Datenqualität zu binden (BVDW 2018; Stein et al. 2022; Rea und Sutton 2019; Band et al. 2022).

Ziel dieser Studie ist es daher, für Fachleute aus Wirtschaft und Wissenschaft, die an Projekten und Themen der Datenwirtschaft arbeiten, zunächst die aktuelle Fachliteratur zu Datenqualität und Qualitätsmetriken zusammenzufassen. Dies schließt die Betrachtung der rechtlichen Rahmenbedingungen mit ein, soweit sie Relevanz für die Bewertung der Datenqualität haben. Das trifft unter anderem auf den Datenschutz, das Urheberrecht und den Schutz von Betriebs-

geheimnissen zu, aber auch auf das Haftungsrecht. Die Studie wirft außerdem einen Blick auf den heutigen Umgang mit Datenqualität in der Datenwirtschaft bei deutschen Unternehmen und leitet daraus Handlungsempfehlungen für Unternehmen, Branchen- bzw. Fachverbände und für die Wissenschaft ab.

Da Datenwirtschaft und Datenqualitätsmanagement noch nicht in der Breite der deutschen Unternehmen angekommen sind und um Wirkzusammenhänge besser erfassen zu können, wurde die Studie als qualitative Untersuchung mit Experteninterviews angelegt. Befragt wurden Vertreter von Dienstleistungsunternehmen und Forschungseinrichtungen, die als sogenannte verarbeitende Intermediäre an der Umsetzung von Datenwertschöpfung beteiligt sind sowie Vertreter von Unternehmen, die Plattformen für den Datenhandel oder das Data Sharing betreiben. Interviews mit Vertreterinnen oder Vertretern von Unternehmen, die in der Datenwertschöpfungskette ausschließlich als datengebende oder datennutzende Unternehmen agieren, konnten trotz mehrerer Anfragen nicht geführt werden. In Gesprächen mit den Experten wurde die Vermutung bestätigt, dass solche Unternehmen derzeit nicht offen auftreten wollen und sehr großen Wert auf die Geheimhaltung ihrer konkreten Anwendungsfälle und Wertschöpfungsprozesse legen. Ihre Positionen und Bedarfe konnten aber anhand der Erfahrungen der Befragten mit ihren Partnern und Kunden ausreichend abgebildet werden.

Zunächst wurden in einer Literaturrecherche die fachwissenschaftlichen Grundlagen zu Datenqualität und Qualitätsmetriken in der Datenwirtschaft herausgearbeitet. Die Ergebnisse wurden im Juli 2021 in einem Workshop mit Experten und Expertinnen der Arbeitsgruppe Data Economy des KI Bundesverbands validiert. Anschließend wurden leitfadengestützte Interviews mit elf Experten aus der Datenwirtschaft zu ihrem Umgang mit Datenqualität, zu ihrer allgemeinen Einschätzung des Feldes und zu Handlungsempfehlungen geführt. Die konsolidierten Resultate wurden in einem zweiten Workshop im Januar 2022 mit mehreren Interviewpartnern und dem Leiter der Arbeitsgruppe Data Economy des KI Bundesverbands validiert. Dabei wurden auch die aus der Literaturrecherche und den Interviews abgeleiteten Handlungsempfehlungen diskutiert und ergänzt.

Abbildung 1: Ablauf der Erstellung der Studie

Kapitel 2 der Studie fasst die aktuelle Fachliteratur zu Datenqualität und Qualitätsmetriken in der Datenwirtschaft zusammen, einschließlich der rechtlichen Rahmenbedingungen. Kapitel 3 reflektiert anhand der Experteninterviews die heutige Praxis in deutschen Unternehmen und beschreibt zu erwartende Entwicklungen. Kapitel 4 gibt aus der Literatur und den Expertengesprächen abgeleitete Handlungsempfehlungen für Wirtschaft, Verbände und Wissenschaft.

Die Autorinnen und Autoren bedanken sich herzlich bei den Befragten für die Teilnahme an den Interviews und den Workshops:

- Daniel Abbou, KI Bundesverband
- Stephan Boch, KEB Automation KG
- Dan Follwarczny, zum Zeitpunkt des Interviews Uniserv GmbH, jetzt ecovium GmbH
- Dr. Andreas Herzog, Fraunhofer-Institut für Fabrikbetrieb und -automatisierung IFF
- Tobias Manthey, Evotegra GmbH und Leiter der Arbeitsgruppe Data Economy im KI Bundesverband
- Kai Meinke, deltaDAO AG
- Fabian Müller, STATWORX GmbH
- Will Perkins, Augmented-Reality-Entwickler
- Calvin Rix, FIR e. V. an der RWTH Aachen
- Dennis Weber, Advaneo GmbH
- Sebastian Wiemann, T-Systems International GmbH
- Dr. Georg Wittenburg, Inspirient GmbH

Die Verantwortung für den Inhalt dieser Studie liegt ausschließlich bei den Autorinnen und Autoren.

Die Studie wurde im Rahmen der Begleitforschung zum Technologieprogramm „Smarte Datenwirtschaft“ (SDW) des Bundesministeriums für Wirtschaft und Klimaschutz erstellt. Im Programm arbeiten 21 Projekte an der Erprobung innovativer Digitaltechnologien für die Datenwirtschaft (www.smarte-datenwirtschaft.de). Mehrere Interviewpartner sind bzw. waren an den SDW-Projekten Future Data Assets und EVAREST beteiligt oder sind im KI Bundesverband, einem Partner der Begleitforschung, aktiv.

Die Autorinnen und Autoren bedanken sich sehr herzlich bei Christoph Pflock (BMWK), Dr. Regine Gernert (DLR Projektträger), Dr. Christiane Graß (DLR Projektträger) und Dr. Barbara Schmitz (DLR Projektträger) für die Unterstützung und wertvollen Hinweise im Verlauf der Kurzstudie.

Den Autorinnen und Autoren war die Einhaltung einer geschlechtergerechten Sprache ein Anliegen. Deshalb wurden in Bezug auf natürliche Personen geschlechtergerechte Beschreibungen gewählt. Bei Organisationen wurden sowohl geschlechterneutrale Formen als auch das generische Maskulinum verwendet, wobei jeweils eine Abwägung zwischen Lesegewohnheiten und Geschlechtsneutralität getroffen wurde.

02

2 GRUNDLAGEN

Datenqualität ist kein Selbstzweck, sondern dient dem Betrieb einer effizienten Datenwirtschaft. Daher werden zunächst die grundlegenden Begriffe zur Datenwirtschaft vorgestellt, bevor es anschließend um die Konzepte von Datenqualität und Qualitätsmetriken und die besonderen Anforderungen des Datenhandels geht. Ebenfalls ein Qualitätsmerkmal ist die Möglichkeit zum rechtssicheren Einsatz von Daten. Zudem können die Qualitätseigenschaften von Daten selbst auch ein Rechtsgegenstand sein. Dieses Grundlagenkapitel schließt daher mit einer Betrachtung des juristischen Rahmens für Datenwirtschaft und Datenqualität.

2.1 Datenwirtschaft, Formen des Datentausches und Anforderungen an die Datenqualität

In der Datenwirtschaft gibt es verschiedene Formen des Datentausches: den direkten Austausch von Daten zwischen Unternehmen, das Data Sharing in geschlossenen Netzwerken, die kostenfreie Bereitstellung meist öffentlicher Daten als Open Data und den Datenhandel. Je nach Datenstrategie, die ein Unternehmen verfolgt, können Daten als IT-Ressource, als Wirtschaftsgut oder als Handelsgut gesehen werden, woraus sich jeweils andere Betrachtungen des Datenwerts und andere Anforderungen an die Datenqualität ergeben. Die folgenden Abschnitte führen daher nicht nur den Begriff der Datenwirtschaft ein, sondern erläutern auch die wichtigsten Formen des Datentausches und ihre besonderen Anforderungen an die Datenqualität.

2.1.1 DATENWIRTSCHAFT: DEFINITION UND ROLLEN

Mit Datenwirtschaft (Data Economy) ist nach allgemeiner Definition eine datengetriebene Wertschöpfung in einem oder über mehrere Unternehmen hinweg gemeint:

„Im Kern beschäftigt sich die Data Economy mit der Monetarisierung von Informationen auf Basis gewonnener Daten, welche mit einem Algorithmus zu werthaltigen Informationen transformiert und anschließend auf Basis der betriebswirtschaftlichen Funktionen zugänglich gemacht werden. Data Economy kann als eigenes Business-Modell betrieben werden oder unterstützt, verändert oder ersetzt bestehende Wertschöpfungsmodelle durch eine zunehmende Digitalisierung“ (BVDW 2018).

Dies beinhaltet die „Erzeugung, Erhebung, Speicherung, Verarbeitung, Verteilung, Analyse, Aufbereitung, Lieferung und Nutzung von Daten mit Hilfe der Digitaltechnik" (Europäische Kommission 2017). Datenwertschöpfung erfolgt als schrittweise Transformation, die dazu dient, Daten wertschöpfend nutzbar zu machen (siehe Abbildung 2). In dieser Datenwertschöpfungskette werden Rohdaten zunächst erhoben (Schritt 1), dann durch Strukturierung, Bereinigung und Vorverarbeitung zu nutzbaren Daten aufbereitet (Schritt 2), anschließend durch Analyse, Integration mit anderen Datensätzen und inhaltliche Aufbereitung zu anwendungsrelevanten Informationen veredelt (Schritt 3), welche dann über Schnittstellen bereitgestellt (Schritt 4) und letztendlich wirtschaftlich genutzt werden (Schritt 5).

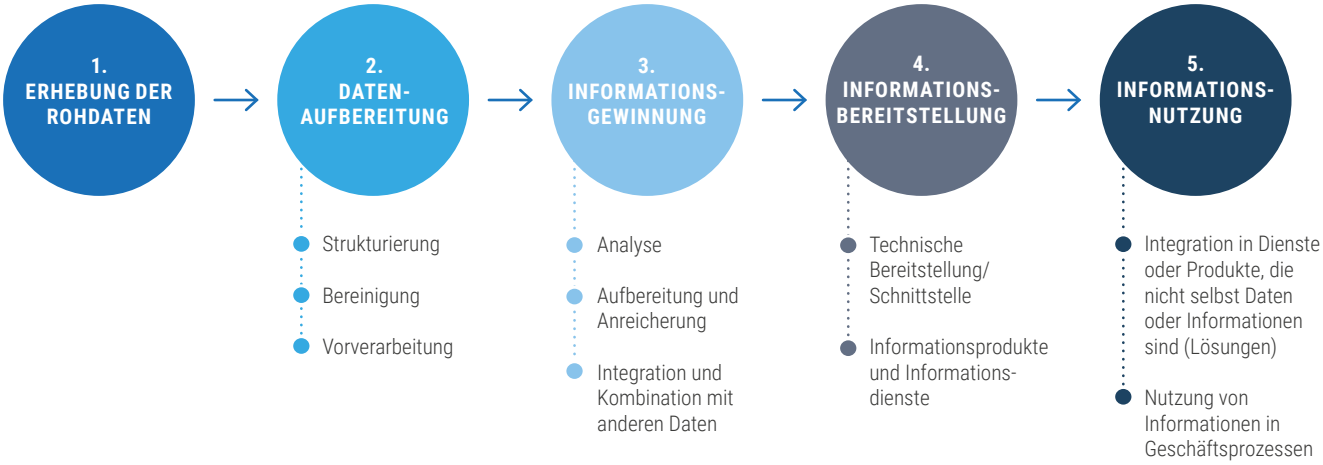

Abbildung 2: Die Datenwertschöpfungskette von den Rohdaten bis zur Informationsnutzung (BVDW 2018)

Im Einzelfall kann es schwierig sein, Daten und die daraus entstehenden, wirtschaftlich nutzbaren Informationen scharf voneinander zu trennen, da diese Unterscheidung anwendungsfallspezifisch ist (⟶ siehe „Daten, Informationen und Wissen"). Wenn neue Datenwertschöpfungsketten entstehen, können Informationsangebote, die für einen Anwendungsfall entwickelt wurden, als „Rohstoff Daten" für einen neuen Anwendungsfall dienen, was auch als Repurposed Data bezeichnet wird (Zhang et al. 2019). Hierbei müssen Informationsangebote meist mit anderen Datenbeständen harmonisiert werden und die Informationsgewinnung erneut im Hinblick auf das neue Wertschöpfungsziel durchlaufen. Beim „Repurposing" von personenbezogenen Daten bestehen neben technischen auch regulatorische Herausforderungen, da die europäische Datenschutz-Grundverordnung (DSGVO) die Nachnutzung von personenbezogenen Daten für nicht vorhergesehene Zwecke deutlich einschränkt (Schweitzer und Peitz 2017).

DATEN, INFORMATIONEN UND WISSEN

Die Unterscheidung von Daten und Informationen geht auf die Wissensrepräsentation (Aamodt und Nygård 1995) sowie die Semiotik (Shanks und Darke 1998) zurück, die drei Betrachtungsebenen von Daten berücksichtigen. Die Datenebene (Syntax) setzt sich mit der Form der Daten auseinander. Syntaktisch betrachtet sind Daten nach bestimmten Regeln erstellte symbolische Repräsentationen. So ist „grün" syntaktisch eine Zeichenfolge aus vier Buchstaben. Die Informationsebene (Semantik) setzt sich mit dem Inhalt der Daten auseinander. Semantisch haben Daten eine Bedeutung, d. h. sie bildet etwas ab und dieses Abbildungsverhältnis ist objektiv verifizierbar. Eine Datenzeile zu einer Person weist etwa als Attribut Augenfarbe „grün" auf und die Person hat tatsächlich grüne Augen. Die Wissensebene (Pragmatik) bezeichnet den Gebrauch von Daten. Durch Auswertung im Kontext des gesamten Wissens zum Anwendungsfall werden Entscheidungen oder Handlungen abgeleitet, die eine pragmatische Relevanz haben. Der Empfehlungsdienst eines Onlineshops schlägt z. B. einer Person vermehrt Kleidung in Herbstfarben vor, da dies laut Wissensbasis gut zu grünen Augen passt.

Die Transformationsschritte der Datenwertschöpfungskette (⟶ siehe Abbildung 2) stellen diese Betrachtungsebenen nacheinander in den Fokus: Am Anfang steht bei der Erhebung und Aufbereitung der Daten (Schritte 1 und 2) die technische Nutzbarkeit im Vordergrund, die eher auf formalen, syntaktischen und statistischen Eigenschaften beruht. Beispielsweise könnten die Werte „grün", „gruen" und „green" vereinheitlicht werden. Bei der Informationsgewinnung (Schritt 3) werden auf der Ebene der Inhalte (Semantik) die relevanten Aspekte der Daten herauskristallisiert, beispielsweise durch Kombination der Augenfarbendaten mit Daten zu vorherigem Kaufverhalten und Integration dieser Daten zu einem persönlichen Farbprofil, welches entscheidungsrelevante Informationen enthält. Irrelevante Datenattribute eines Datensatzes (z. B. welche Haustiere die betroffene Person besitzt) werden zur Verschlankung entfernt. Die Interpretation im Anwendungskontext (Pragmatik) generiert dann entscheidungs- bzw. handlungsrelevantes Wissen (Schritt 5) wie etwa eine Kaufempfehlung.

Unternehmen können entweder eine oder auch mehrere Rollen in der Datenwertschöpfungskette einnehmen (Europäische Kommission 2017):

Datenerhebende Unternehmen verfügen über selbsterhobene Daten. Hier gibt es sowohl Unternehmen, bei denen die Datenerhebung das Kerngeschäft ist, wie Dienstleister für Kundenbefragungen, als auch Unternehmen, die Datenbestände aus ihrem Kerngeschäft in einem Nebengeschäft zweitverwerten, wie etwa Zustellunternehmen, die Adressdaten anbieten.

Datennutzende Unternehmen nutzen Daten, Datenderivate oder Informationen zur Wertschöpfung. Auch hier kann unterschieden werden zwischen Datennutzenden, deren Kerngeschäft datenbasierte Produkte oder Dienste sind (wie Googles Werbeangebote oder der automatisierte Übersetzungsdienst Deepl), sowie Datennutzenden, bei denen die Nutzung von Daten der Verbesserung oder Erweiterung des Kerngeschäfts dient (beispielsweise ein Unternehmen, das einen Adressdatendienst bezieht, um sein CRM-System zu aktualisieren).

Intermediäre sind an der Datenwertschöpfung beteiligt, sind aber weder endnutzende noch datenerhebende Akteure.

- **Verarbeitende Intermediäre** sind direkt an der Datenwertschöpfung beteiligt. Sie befassen sich mit Aspekten der Aufbereitung und Informationsgewinnung von Daten. Zu den verarbeitenden Intermediären gehören Analysedienstleistende, Anbietende von Diensten zur Datenveredlung und Datenbroker, die Datensätze verschiedener Datengebender kompilieren (Dewenter und Lüth 2019).
- **Ermöglichende Intermediäre** bieten Dienste und Produkte an, die Datenwertschöpfung ermöglichen oder unterstützen. Zu ihnen zählen etwa Infrastruktur-, Plattform- und Softwareanbieter, Rechtsexpert:innen oder Zertifizierungsdienstleister.

2.1.2 FORMEN DES DATENTAUSCHES: DATA SHARING, OPEN DATA UND DATENHANDEL

Die Zusammenarbeit verschiedener Unternehmen in der Datenwirtschaft erfordert einen Datenaustausch zwischen Unternehmen. Oft geschieht das noch direkt, sei es per E-Mail oder über einen Cloud-Speicherdienst. Mittlerweile werden dafür aber auch spezialisierte Datenplattformen mit erweiterten Funktionen für das Daten- und Nutzermanagement eingesetzt. Im einfachsten Fall tauscht ein Unternehmen über eine eigene Plattform oder Schnittstelle bilateral Daten mit seinen Kundinnen und Kunden, Partnern und Dienstleistern. Auch der Verkauf von Daten wird zum Teil über eigene Plattformen abgewickelt.

→ MEHR INFORMATIONEN zum Thema Data Sharing in der Studie „How to share data?“

Zunehmend gibt es jedoch auch Plattformen, die von Intermediären für einen multilateralen Datenaustausch mit jeweils mehreren Datengebenden und Datennehmenden betrieben werden. Für solche Datenplattformen haben sich mehrere Formen des Austausches entwickelt: Ein gängiges Modell ist das unentgeltliche Data Sharing zwischen Unternehmen über entsprechende Plattformen (Schweitzer und Peitz 2017). Häufig fokussieren sich diese Plattformen auf eine bestimmte Branche, für die der Tausch von Stammdaten oder Prozessdaten angeboten wird (Lindner et al. 2021). Eng verwandt damit ist die kostenfreie, offene Bereitstellung von Daten (Open Data) durch öffentliche oder private Anbietende. Die Motivation hierfür ist entweder das Allgemeinwohl oder die Hoffnung auf implizite Gegenleistungen wie z. B. ein Imagegewinn, die Möglichkeit, Standards zu setzen, oder eine Verbesserung der eigenen Angebote durch komplementäre Dienste, die

mithilfe der bereitgestellten Daten entwickelt werden. Häufig stellen datengebende Unternehmen Daten auch gegen nicht-monetäre Gegenleistungen zur Verfügung. Nutzende von Datenplattformen beispielsweise räumen den Plattformbetreibern meist die Rechte zur Erhebung und wirtschaftlichen Verwertung der Nutzungsdaten ein, um die Plattformdienste nutzen zu können (Schweitzer und Peitz 2017).

Als Datenhandel (Schweitzer und Peitz 2017) wird die Bereitstellung von Daten gegen ein Entgelt bezeichnet. Eine Transaktion findet zwischen mindestens zwei Handelspartnern statt, dem datengebenden und dem datennehmenden Unternehmen. Der Handel geschieht entweder direkt zwischen Partnerunternehmen oder in etablierten, offenen Marktstrukturen, wie es sie z. B. für Adress-, Markt-, Konsumenten- und Geodaten gibt (Dewenter und Lüth 2019). Vereinzelt findet er auch schon auf breiter angelegten Datenmarktplätzen statt, die diverse Datenangebote für eine breite Käuferschaft zur Verfügung stellen (Lindner et al. 2021).

Erfolgt der Datenhandel bilateral direkt vom Datengebenden an den Datennehmenden, so spricht man von Primärmärkten für Daten (Schweitzer und Peitz 2017). Dem primären Datenhandel gegenüber steht der sekundäre Datenhandel über Datenmarktplätze von Intermediären. Es gibt allerdings keine einheitliche Definition von Datenmarktplätzen (Meisel und Spiekermann 2019). In Anlehnung an die Arbeiten von Meisel und Spiekermann (2019) bzw. Koutroumpis et al. (2017) definiert diese Studie Datenmarktplätze als Plattformen, deren Wertversprechen es ist, Datengebende und Datennehmende zu vermitteln und die Leistungen zur technischen, vertraglichen und monetären Abwicklung der Transaktionen bereitzustellen (siehe Abbildung 3). Andere Autorinnen und Autoren gehen weiter und beziehen in ihre Definition auch den Handel mit Datenderivaten und datennahen Diensten über die Plattform mit ein (Trauth und Mayer 2022). Auch der Übergang von Data-Sharing-Plattformen zu Datenmarktplätzen ist noch fließend (Lindner et al. 2021).

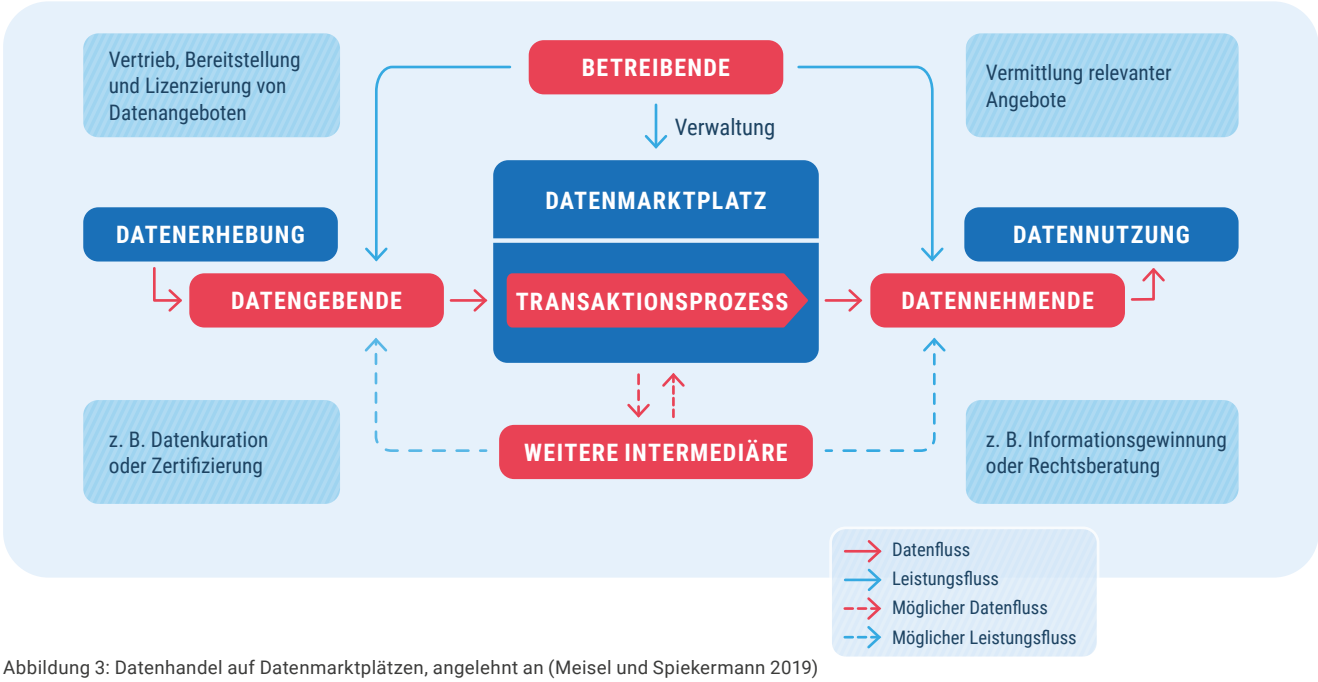

Abbildung 3: Datenhandel auf Datenmarktplätzen, angelehnt an (Meisel und Spiekermann 2019)

Am sekundären Datenhandel können neben dem Datengebenden, dem Datennehmenden und dem Marktplatzbetreiber noch eine Reihe weiterer Intermediäre an einer Datenhandelstransaktion beteiligt sein, so zum Beispiel Bereitsteller von digitalen Diensten oder Dienstleistungen zur Herstellung von Datenqualität, Beratende (z. B. für rechtliche Fragen) oder Anbietende anderer Support-Leistungen (Meisel und Spiekermann 2019). Der Datengebende des Datenhandels ist nicht immer der Datenerhebende der Datenwertschöpfungskette und der Datennehmende des Datenhandels ist nicht immer der Datennutzende der Datenwertschöpfungskette: Beispielsweise kann ein KI-Dienstleister Daten von einem Datenbroker kaufen, der diese aus mehreren Datenangeboten zusammengestellt hat, um diese dann bei einem Unternehmenskunden wertschöpfend einzusetzen: Dies wäre ein Datenhandel zwischen zwei verarbeitenden Intermediären der Datenwertschöpfungskette. Der Abgleich der Datenbeschaffenheit mit den Qualitätsanforderungen der Datennutzungsseite ist ein vorbereitender Schritt einer Datenhandelstransaktion. Bei einem Handel zwischen Intermediären sind diese daher auf aussagekräftige Angaben von datenerhebenden und datennutzenden Unternehmen angewiesen.

2.1.3 DATENWERT UND DATENQUALITÄT IN DER DATENWERTSCHÖPFUNG

Je nachdem, welche betriebswirtschaftliche Strategie ein Unternehmen mit seinen Daten verfolgt, kann der wirtschaftliche Wert von Daten unterschiedlich bestimmt werden. Hierbei spielt es eine Rolle, ob Daten primär intern wertschöpfend genutzt werden oder als datengetriebene Angebote am Markt monetarisiert werden (Rea und Sutton 2019). Wenn Unternehmen bislang nicht in der Datenwirtschaft aktiv waren, nehmen sie meist zunächst, aufbauend auf existierenden Datenbeständen und IT-Ressourcen aus dem Kerngeschäft, verhältnismäßig leicht umzusetzende interne Datenwertschöpfungsprojekte ins Visier. Eine Datenmonetarisierung am offenen Markt erfolgt dann erst in einem zweiten optionalen Schritt (BVDW 2018). Diese unterschiedlichen Datenstrategien bringen jeweils andere Anforderungen an die Daten mit sich, wobei eine zunehmende Öffnung nach außen mit wachsenden Ansprüchen einhergeht (siehe Abbildung 4).

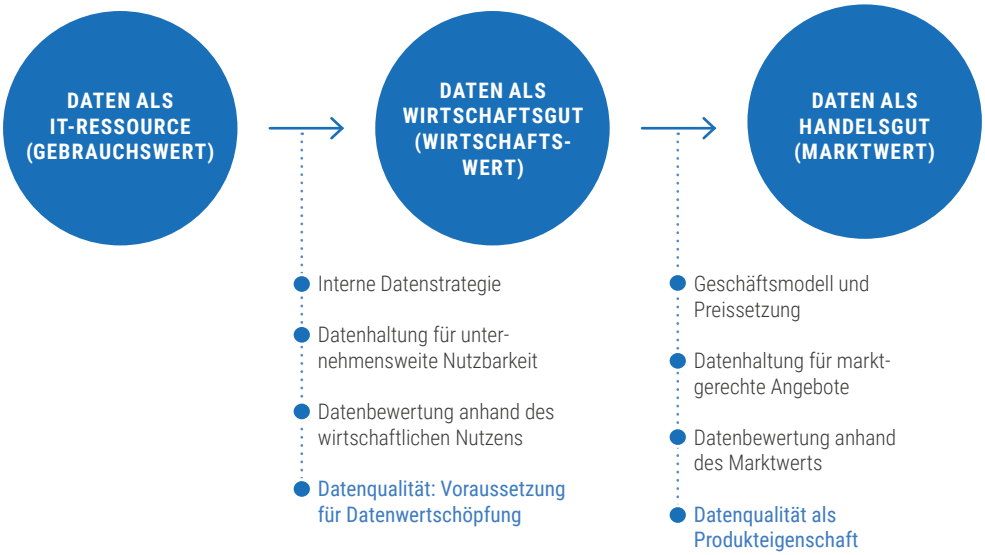

Abbildung 4: Die Abhängigkeit von Datensicht, Datenwert und Datenqualität (BVDW 2018; Rea und Sutton 2019)

Daten als IT-Ressource. Daten haben in Unternehmen zunächst nur einen Gebrauchswert (User Value), aber keinen explizit abgeschätzten wirtschaftlichen Wert im Sinne der Datenwertschöpfung (BVDW 2018). Eine wirtschaftliche Bewertung von Daten kann in solchen Fällen nur über einen kostenbasierten Ansatz erfolgen, der abschätzt, wie viel eine Reproduktion oder ein Ersatz der Daten das Unternehmen kosten würde (Rea und Sutton 2019; Stein et al. 2022). Der Umgang mit Daten und Datenqualität obliegt meist IT-Fachkräften und orientiert sich an der operativen Verwendung in einem dezidierten Geschäftsbereich.

Daten als Wirtschaftsgut. Als Einstieg in die Datenwirtschaft wird häufig das Ziel gesetzt, das Kerngeschäft durch Einsparungen oder zusätzliche Erlöse zu stärken und unternehmensintern Datenwert zu erschließen (Business Value), z. B. durch Optimierung oder Effizienzsteigerung verschiedener Wertschöpfungsprozesse (Rea und Sutton 2019; BVDW 2018). Daten werden erstmals unternehmensweit als Wirtschaftsgut betrachtet und einer Datenstrategie und Kosten-Nutzen-Abschätzung folgend bewertet (BVDW 2018; Stein et al. 2022), was in Zukunft etwa auch im Rahmen einer Bilanzierung des Vermögenswerts von Daten eines Unternehmens relevant sein könnte.[1] Durch solche Kosten-Nutzen-Betrachtungen kann ein direkter Bezug zwischen Metriken für Datenqualität und wirtschaftlichem Datenwert hergestellt werden (BVDW 2018), sodass Datenqualität zu einem strategischen Belang wird.

DIE INTEROPERABLE NUTZBARKEIT UND ZUGREIFBARKEIT VON DATEN TRETEN IN DER DATENWIRTSCHAFT ALS QUALITÄTSEIGENSCHAFTEN IN DEN VORDERGRUND.

Durch die unternehmensweite Betrachtung von Daten als Wirtschaftsgut rücken die interoperable Nutzbarkeit und Zugreifbarkeit von Daten als Qualitätseigenschaften erstmalig in den Fokus: Wenn Unternehmen datenwirtschaftlich aktiv werden, erfolgt daher häufig eine Modernisierung der Dateninfrastruktur sowie der Prozesse und Regelungen zur Datenhaltung und -nutzung (Datenmanagement, Data Governance) (Demary et al. 2019; BVDW 2018). Die initial hohen Aufwände hierfür rechnen sich meist als Teil einer umfassenden Datenstrategie, da der Wert der Unternehmensdaten durch Zusammenführung mit anderen Daten steigt und die Kosten der Umsetzung zukünftiger Datenwertschöpfungsaktivitäten sinken (Stein et al. 2022). Der Digitalisierungsgrad eines Unternehmens kann daher auch zur wirtschaftlichen Bewertung eines Unternehmens herangezogen werden (Band et al. 2022).

Eine Zwischenform zwischen der internen Datenwertschöpfung und dem offenen Handel von Daten bzw. datenbasierten Diensten und Produkten ist die Datenwertschöpfung in geschlossenen Netzwerken. Werden Daten über Unternehmensgrenzen hinweg geteilt, misst sich die Datenqualität zunehmend auch an der Konformität mit gängigen Standards für Datenformate. Zudem rücken rechtliche Vereinbarungen zur Datenüberlassung sowie die rechtssichere Verwendbarkeit von Datensätzen (⟶ siehe Abschnitt 2.3) in den Fokus.

1 Die Bilanzierung des Datenkapitals eines Unternehmens als Asset und entsprechende Regulierungen werden im Projekt Future Data Assets des Technologieprogramms „Smarte Datenwirtschaft" behandelt und münden in ein „Rahmenwerk[s] zur Identifikation, Analyse, Bewertung, Steuerung sowie der internen und externen Berichtserstattung" (Stein et al. 2022; Band et al. 2022).

Daten als Handelsgut. Unternehmen können neue Geschäftsbereiche erschließen, indem sie Daten oder datenbasierte Angebote für den offenen Markt entwickeln. Sowohl die Anforderungen an die Datenqualität als auch der Datenwert entstehen dann nicht mehr primär intern, sondern ergeben sich aus den Anforderungen und der Zahlungsbereitschaft der Kundschaft. Den Preis legt der Anbietende unter Berücksichtigung des Wettbewerbs und des betriebswirtschaftlichen Nutzens beim Kunden so fest, dass Vorteile für Datengebende und Datennehmende entstehen. So entstehen Anreize für längerfristige Geschäftsbeziehungen (Leiting et al. 2022). Der Angebotspreis bzw. der Preis vergleichbarer Angebote bestimmt dann den Marktwert (Market Value) datenbasierter Angebote, welcher zur Bestimmung des wirtschaftlichen Datenwerts herangezogen wird (Rea und Sutton 2019).

DIE ANFORDERUNGEN AN DIE DATENQUALITÄT UND DER DATENWERT ERGEBEN SICH AUS DEN ANFORDERUNGEN UND DER ZAHLUNGSBEREITSCHAFT DER KUNDSCHAFT.

Finden Datentransaktionen in geschlossenen Netzwerken statt, in denen dauerhaft zusammengearbeitet wird und ein Vertrauensverhältnis besteht, so sind Anpassungen der Daten und Nachfragen zu den Daten mit nur geringen Aufwänden und Kosten verbunden. Werden Daten oder datenbasierte Angebote hingegen offen am Markt angeboten, so sind Qualitätsmerkmale auch Produkteigenschaften. Die bedarfsgerechte Beschaffenheit des datenbasierten Angebots als alleinstehende Einheit gewinnt an Bedeutung, ebenso wie seine Ex-ante-Bewertbarkeit. Beim Handel mit Daten und anderen Immaterialgütern besteht eine Informationsasymmetrie zwischen Datengebenden und Datennehmenden hinsichtlich der Beschaffenheit der Güter, welche es im Sinne der Markttransparenz und -effizienz abzubauen gilt (Dewenter und Lüth 2019). Aussagekräftige und transparente Angaben zur Datenqualität sind ein wichtiger Teil der Beschaffenheitsangaben, anhand derer die Nachfrageseite entscheidet, ob sie die Transaktion zu den vorliegenden Bedingungen durchführen möchte.

Schlüsselherausforderung Offenlegung von Daten. Häufig bestehen in Unternehmen Bedenken hinsichtlich eines möglichen Abflusses von Geschäftsgeheimnissen durch die Offenlegung von Daten, selbst wenn der Empfängerkreis noch limitiert ist, wie beim Data Sharing in geschlossenen Netzwerken. Das Risiko eines potenziell geschäftsschädigenden Datenabflusses sowie auch die Aufwände, um dem entgegenzuwirken, können sich negativ auf die unternehmensinterne Kosten-Nutzen-Abschätzung für einen datenwirtschaftlichen Anwendungsfall auswirken.

Infrastrukturangebote, die eine strikte Nutzungskontrolle und eine dezentrale Datenverarbeitung unterstützen, wie beispielsweise in den Rahmenentwicklungen des Industriekonsortiums International Data Spaces Association (IDSA)[2] und im europäischen Projekt GAIA-X[3] spezifiziert, erleichtern den Umgang mit diesen rechtlichen und technischen Herausforderungen. Auch eine strategische Ausgestaltung des Datenangebots, beispielsweise durch Aggregation, Vorverarbeitung oder Löschung sensibler Datenfelder kann Rückschlüsse auf sensible Inhalte in vielen Fällen schwer oder unmöglich machen. Solche Maßnahmen, die der Informationssicherheit dienen, können sich jedoch negativ auf die Datenqualität auswirken (aufgrund geringerer Transparenz oder Unvollständigkeit der Daten).

2 https://internationaldataspaces.org/ (Abruf am 08.03.2022).
3 https://www.gaia-x.eu/ (Abruf am 08.03.2022).

2.2 Datenqualität

Die folgenden Abschnitte geben einen Überblick über die Definition von Datenqualität, die für die konkrete Arbeit wichtigen Rahmenkonzepte für Datenqualität sowie das Vorgehen bei der Erhebung von Datenqualität anhand von Metriken. Abschließend werden Anforderungen für ein auf den Datenhandel ausgerichtetes Rahmenkonzept diskutiert.

2.2.1 DEFINITION VON DATENQUALITÄT

Unter Qualität versteht man nach gängiger Definition einen objektivierten Maßstab, der abbildet, inwieweit die realisierte Beschaffenheit einer betrachteten Einheit, z. B. eines gefertigten Produkts, der geforderten Beschaffenheit entspricht (Geiger und Kotte 2008). Datenqualität ist seit Beginn der 1990er-Jahre ein aktiver Forschungsgegenstand und kann aus verschiedenen Perspektiven definiert werden (⟶ siehe Tabelle 1):

Eine der ersten Definitionen von Datenqualität nimmt die Perspektive der Datennutzenden („Konsumenten“) ein: „Daten sind qualitativ hochwertig, wenn sie sich für die Nutzung durch Datenkonsumenten eignen“ (Strong et al. 1997), was zwei Aspekte beinhaltet: ihre Gebrauchsfähigkeit (Usability) und ihre Dienlichkeit (Usefulness). Diese Definition ist bis heute anerkannt und weit verbreitet, wenn auch manchmal in geänderter oder erweiterter Form. Neuere Ansätze fügen der Definition eine technische Sicht oder eine wirtschaftliche Sicht hinzu. Die Definition wird dann erweitert um Aspekte wie „Mängelfreiheit“ oder „Einhaltung der Spezifikationen“ (Kahn et al. 2002; Redman 2001) bzw. „Eignung für den vorgesehenen Verwendungszweck in Praxis, Entscheidungsfindung oder Planung“ oder „Fähigkeit, die festgelegten Geschäfts- und Systemanforderungen sowie die technischen Anforderungen eines Unternehmens zu erfüllen“ (Mahanti 2019; Redman 2001).

DATEN SIND QUALITATIV HOCHWERTIG, WENN SIE SICH FÜR DIE NUTZUNG DURCH DATENKONSUMENTEN EIGNEN.

Neben den wissenschaftlichen Definitionen von Datenqualität existiert auch eine Definition aus Sicht der Normung. Laut ISO 9000:2015 (Qualitätsmanagementsysteme – Grundlagen und Vokabular) und ISO 8000-2:2020 (Datenqualität – Teil 2: Vokabular) ist die Datenqualität ein Maß für den Grad, mit dem ein Satz inhärenter Datenmerkmale die Anforderungen erfüllt. Unter inhärenten Merkmalen versteht man dabei den Daten innewohnende Eigenschaften, die objektiv messbar sind. Eine Anforderung ist als ein Bedarf oder eine Erwartung definiert, der/die festgelegt, vorausgesetzt oder verpflichtend ist. Unabhängig von der zugrunde liegenden Perspektive ist allen Definitionen von Datenqualität gemein, dass sich der Qualitätsgrad über den Vergleich des vorliegenden Zustands der Daten mit einem erwünschten Zustand ermitteln lässt.

QUELLE	DEFINITIONEN
Strong et al. (1997)	„We define high-quality data as data that is fit for use by data consumers—a widely adopted criteria. This means that usefulness and usability are important aspects of quality."
Redman (2001)	„Data are of high quality if they are fit for their intended uses in operations, decision making, and planning. Data are fit for use if they are free of defects and possess desired features."
Kahn et al. (2002)	„We assigned the latter two views of quality, conforming to specifications, and meeting or exceeding consumer expectations."
Mahanti (2019)	„the capability of data to satisfy the stated business, system, and technical requirements of an enterprise"
ISO 8000-2:2020 ISO 9000:2015	Data Quality: „degree to which a set of inherent characteristics of data fulfils requirements"

Tabelle 1 Die Funktionen von Data-Sharing-Plattformen in Anlehnung an das Referenzmodell von Meisel und Spiekermann (2019)

2.2.2 RAHMENKONZEPTE FÜR DATENQUALITÄT

Ein theoretisches Rahmenkonzept für Datenqualität ist eine Grundvoraussetzung, um Datenqualität messbar zu machen. Es legt dar, was genau gemessen werden soll. Die Vielschichtigkeit der Datenqualität ist für die Entwicklung von Rahmenmodellen für Datenqualität eine besondere Herausforderung: Zum einen muss ein vielfältiges Spektrum möglicher Datenqualitätsprobleme berücksichtigt werden; zum anderen hängt deren Relevanz von den jeweiligen Perspektiven der beteiligten Parteien und vom Anwendungszweck der Daten ab. Für ein Unternehmen, das Adressdaten für das Dialogmarketing im Handel anbietet, sind auch kleine Anschriftsfehler in den internen Datenbeständen problematisch. Für eine Forschungseinrichtung, die Adressdaten zur demografischen Forschung verwendet, sind solche Mängel, die die Zustellbarkeit von Postsendungen betreffen, hingegen irrelevant.

— ES GIBT EIN VIELFÄLTIGES SPEKTRUM VON MÄNGELN DER DATENQUALITÄT, VON FORMFEHLERN ÜBER MANGELHAFTE ZUGÄNGLICHKEIT BIS HIN ZU FEHLERHAFTEN WERTEN. —

Die Bandbreite der im Rahmen der Forschung identifizierten typischen Datenqualitätsprobleme reicht von fehlerhaften oder gänzlich fehlenden Werten über Anomalien, logische Inkonsistenzen und Syntaxverstöße bis hin zu mangelnder Zugänglichkeit (für eine umfassende Übersicht siehe Oliveira et al. 2005; Fürber 2016). Die Forschung zu Rahmenkonzeptentwicklung hat sich daher auf die Sammlung und Konzeptualisierung der Dateneigenschaften konzentriert, die sich in Form von Datenqualitätsproblemen auswirken können, z. B. die Korrektheit oder die Konsistenz.

Methodisch können drei Herangehensweisen für die Entwicklung von Rahmenkonzepten für Datenqualität unterschieden werden: intuitiv, empirisch und theoretisch (Wang und Strong 1996). Intuitiv heißt, dass für die Konzeptualisierung ausschließlich auf persönliche Erfahrungswerte und subjektive Einschätzungen zurückgegriffen wird. Diese Herangehensweise ist mittlerweile in den Hintergrund getreten.

Empirisch basierte Rahmenkonzepte. Eine der ersten und bis heute bekanntesten empirischen Studien wurde 1996 von Wang und Strong durchgeführt (Wang und Strong 1996). Ausgehend von 118 Dateneigenschaften, die einzelne Aspekte oder Konstrukte der Datenqualität ausmachen (sogenannte Datenqualitätsdimensionen) wurden durch Befragung von 355 Datenkonsumenten die 15 wichtigsten identifiziert (siehe Tabelle 2).

KATEGORIE	DATENQUALITÄTSDIMENSION	KURZBESCHREIBUNG
Inhärent/intrinsisch	Fehlerfreiheit/Korrektheit	Übereinstimmung der Daten mit der Realität
	Glaubwürdigkeit	Wahrnehmung der Daten als real, wahr und glaubwürdig
	Objektivität	Sachlichkeit und Wertfreiheit der Daten
	Hohes Ansehen/Reputation	Vertrauenswürdigkeit der Daten, ihrer Inhalte und der Datenquelle
Kontextuell/extrinsisch	Aktualität	Zeitliche Nähe des Datenabbilds zur Realität
	Angemessener Umfang	Angemessenheit der Datenmenge im Hinblick auf die gestellten Anforderungen bzw. die zu bewältigenden Aufgaben
	Relevanz	Nutzen der Daten für Anwender:innen (Interessantheit, Anwendbarkeit)
	Vollständigkeit	Lückenfreiheit der Daten
	Wertschöpfung/Mehrwert	Beitrag der Nutzung der Daten zur Steigerung einer monetären Zielfunktion
Darstellungsbezogen	Konsistenz/Einheitliche Darstellung	Einheitlichkeit der Repräsentation der Daten (Format, Kompatibilität)
	Eindeutige Auslegbarkeit/ Interpretierbarkeit	Eindeutigkeit der Auslegung der Daten (Verständlichkeit, Einheiten, Definitionen)
	Übersichtlichkeit	Prägnanz der Darstellung der Daten (Kompaktheit, Angemessenheit des Formats)
	Verständlichkeit	Klarheit und Nachvollziehbarkeit der Darstellung der Daten
Systemunterstützt	Bearbeitbarkeit	Verwendbarkeit bzw. Einsatzbereitschaft der Daten
	Zugänglichkeit	Abrufbarkeit bzw. Verfügbarkeit der Daten

Tabelle 2: Die 15 Dimensionen von Datenqualität nach Wang und Strong (1996). Die Übersetzung ins Deutsche ist angelehnt an Rohweder et al. (2008)

Diese Qualitätsdimensionen können entlang von vier Kategorien klassifiziert werden:

- **Inhärente** bzw. intrinsische **Datenqualitätsdimensionen** beruhen auf Eigenschaften, die den Daten selbst innewohnen, wie die Fehlerfreiheit/Korrektheit oder Glaubwürdigkeit.
- **Kontextuelle** bzw. extrinsische **Datenqualitätsdimensionen** beruhen auf Eigenschaften, die nur im Zusammenhang des Anwendungszwecks Bedeutung erhalten, wie die Vollständigkeit oder Relevanz.
- **Darstellungsbezogene Datenqualitätsdimensionen** betreffen das Format oder die Präsentation von Daten, wie z. B. die Konsistenz bzw. einheitliche Darstellung und Verständlichkeit.
- Die **systemunterstützten Datenqualitätsdimensionen** Zugänglichkeit und Bearbeitbarkeit betreffen die Verfügbarkeit und Einsatzbereitschaft der Daten.

Die Deutsche Gesellschaft für Informations- und Datenqualität e. V. empfiehlt seit 2007 die Nutzung dieses Rahmenkonzepts von Wang und Strong.[4]

Die Autorinnen haben ihr Rahmenkonzept zu einem späteren Zeitpunkt zu einem Produkt- und Serviceleistungsmodel für Informationsqualität (PSP/IQ, product and service performance model for information quality) weiterentwickelt (Strong und Kahn 1997; Kahn et al. 1997, 2002). Auch weitere Rahmenkonzepte für Datenqualität wurden durch eine ähnliche Herangehensweise entwickelt, so z. B. Rahmenkonzepte für Informationsqualität von persönlichen bzw. individuellen Webseiten (Katerattanakul und Siau 1999) und Webportalen (Calero et al. 2008).

Theoretisch basierte Rahmenkonzepte. Ein Nachteil der empirischen Herangehensweise ist, dass die Richtigkeit und Vollständigkeit der Resultate nicht anhand fundamentaler Prinzipien überprüft werden kann (Wang und Strong 1996). Relevante Datenqualitätsdimensionen werden zwar identifiziert, sind aber unter Umständen nur vage definiert, gegebenenfalls mehrdeutig und werden nicht von einer soliden Theorie gestützt. Parallel zu den empirischen Untersuchungen werden daher seit den 1990er-Jahren auch theoriebasiert Rahmenkonzepte entwickelt, z. B. auf ontologischer Grundlage (Wand und Wang 1996) oder semiotischer Grundlage (Shanks und Darke 1998). Das semiotische Rahmenkonzept ist Kernelement der ISO 8000-8: „Data quality – Part 8: Information and data quality: Concepts and measuring" (ISO 8000-8:2015). Es bettet Datenqualität in den semiotischen Rahmen der Syntax, Semantik und Pragmatik ein (⟶ siehe „Arbeitsdefinition: Inhärente, systemunterstützte und pragmatische Datenqualität"):

- **Syntaktische Datenqualität** legt den Schwerpunkt auf die Form bzw. Struktur der Daten wie beispielsweise Anforderungen zum Datenformat aus den Metadaten.
- **Semantische Datenqualität** gibt an, ob der Inhalt und die Bedeutung der Daten tatsächlich das abbildet, was sie repräsentieren.
- **Pragmatische Datenqualität** stellt ein Maß für die Zweckeignung der Daten dar, ihr Kern liegt in der Nutzung bzw. der Nutzbarkeit der Daten im Anwendungskontext.

Ein Hauptmerkmal des semiotik basierten Rahmenkonzepts ist, dass es die Ziele im Hinblick auf Datenqualität von den Methoden, sie zu erreichen, entkoppelt.

Rahmenkonzepte für bestimmte Anwendungen. Rahmenkonzepte für bestimmte Anwendungsfelder oder Anwendungszwecke sind in der Regel an den bereits vorgestellten Rahmenkonzepten für Datenqualität angelehnt. Verfügbar sind unter anderem Rahmenkonzepte für Datenqualität in den Bereichen Produktion (Su und Jin 2007), IoT (Zhang et al. 2021), Prozessindustrie (Wiedau et al. 2021), Finanzen (Amicis und Batini 2004) und Gesundheitswesen (Pezoulas et al. 2019; Daniel et al. 2019) sowie Big Data (Ramasamy und Chowdhury 2020), Data Warehouses (Helfert und Herrmann 2002; Nemani und Konda 2009) und Maschinelles Lernen im industriellen Kontext (Timocin 2020). Darüber hinaus ist im Rahmen der ISO/IEC 25012 „Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model" ein Rahmenkonzept für Datenqualität festgelegt, das auf 15 Datenqualitätsmerkmalen beruht, welche zwei Kategorien zugeordnet werden: inhärente Datenqualität und systemabhängige Datenqualität (ISO/IEC 25012:2008).

4 https://www.competence-site.de/informationsqualitaet-15-dimensionen-4-kategorien/ (Abruf am 08.03.2022).

Harmonisierung von Rahmenkonzepten. Eine vergleichende Analyse diverser anwendungsfallspezifischer Rahmenkonzepte zeigt, dass die berücksichtigten Qualitätsdimensionen sehr variabel sind (Cichy und Rass 2019). Einige Dimensionen treten häufig auf, andere dagegen nur vereinzelt. Am gängigsten sind die Dimensionen Korrektheit, Vollständigkeit, Konsistenz, Aktualität und Zugänglichkeit. Die Arbeitsgruppe Datenqualität der niederländischen Sektion der Data Management Association International (DAMA NL) hat sich intensiv mit der Harmonisierung existierender Rahmenkonzepte auseinandergesetzt, um eine Basis für einen Standard für die Dimensionen von Datenqualität und ihre Definitionen zu schaffen. Im Rahmen des 2020 erschienenen Verzeichnisses für Datenqualitätsdimensionen 3DQ (Dictionary of Data Quality Dimensions) listet sie 60 Dimensionen mit nach ISO 704 standardisierten Definitionen und kategorisiert sie entlang eines Konzeptsystems (Black und van Nederpelt 2020). Das Konzeptsystem geht dabei auf verschiedene formale Eigenschaften von Daten ein, wie z. B. die Unterscheidung zwischen Format, Datenwert, Attribut oder Metadaten. Das Verzeichnis soll bei der Auswahl von Qualitätsdimensionen unterstützen sowie den Austausch bzw. die Kommunikation über sie erleichtern und fördern.

ARBEITSDEFINITION: INHÄRENTE, SYSTEMUNTERSTÜTZTE UND PRAGMATISCHE DATENQUALITÄT

Viele Datenqualitätsdimensionen werden in der Literatur wiederkehrend als relevant identifiziert, die Einordnung in die verschiedenen Kategorien hängt jedoch von der jeweiligen Perspektive ab. Der Wahrnehmung der Datenkonsumentinnen und -konsumenten nach, die in Wang und Strong (Wang und Strong 1996) wiedergegeben wird, ist etwa der ausreichende Umfang eines Datensatzes eine kontextuelle Qualitätseigenschaft, während seine Reputation als intrinsisches Qualitätsmerkmal eingestuft wird. Der eher technisch ausgerichteten ISO 8000-8 nach ist die Zuordnung gerade andersherum: Der Umfang ist als syntaktische Eigenschaft objektiv feststellbar und wird somit als inhärente Datenqualitätsdimension eingeordnet, die Reputation hingegen ist subjektiv und wird somit der kontextuellen, pragmatischen Datenqualität zugerechnet (ISO 8000-8:2015).

Diese Studie konzentriert sich auf Szenarien des Data Sharing oder Datenhandels im Rahmen von Datenwertschöpfung, bei denen ein Abgleich zwischen Beschaffenheitsangaben datengebender Unternehmen und Nutzungsanforderungen datennehmender Unternehmen stattfindet. Daher folgt die vorliegende Studie dieser Arbeitsdefinition:

Inhärente Datenqualität umfasst die syntaktische und semantische Datenqualität und ist unabhängig vom Anwendungsfall objektiv einschätzbar. Neben den gängigen Qualitätsdimensionen Korrektheit, Aktualität, Konsistenz und Vollständigkeit sind auch qualitätsrelevante Eigenschaften wie der Umfang eines Datensatzes oder Maße für die Streuung der Datenwerte in diesem Sinne inhärente Datenqualitätseigenschaften.

Systemunterstützte Datenqualität beschreibt die technische Nutzbarkeit von Daten und umfasst die ebenfalls gängigen Qualitätsdimensionen der Zugreifbarkeit und Bearbeitbarkeit von Daten (Wang und Strong 1996).

Pragmatische Datenqualität beschreibt Qualitätsdimensionen, die relativ zum Datenkonsumenten oder datenwirtschaftlichen Anwendungsfall sind. Dies umfasst:

- Subjektiv wahrgenommene Qualitätsdimensionen wie etwa die Relevanz oder Glaubwürdigkeit von Daten.
- Der wirtschaftliche Mehrwert von Daten, der nur im Zusammenhang einer konkreten Anwendung objektivierbar ist.
- Die rechtssichere Verwendbarkeit von Daten, die relativ zum lokalen Rechtsrahmen und den Vertragsmodalitäten der Datenüberlassung bestimmt werden kann.

2.2.3 DIE SPEZIFIKATION UND MESSUNG VON DATENQUALITÄTSMETRIKEN

Um Datenqualität in der Praxis zu messen, müssen einzelne Dimensionen und Unterdimensionen messbar gemacht werden. Diese Operationalisierung erfolgt schrittweise. Erst werden die relevanten Qualitätsdimensionen identifiziert: Was soll gemessen werden und warum? Dann werden mit den Datenqualitätsmetriken die hierfür erforderlichen Messwerkzeuge spezifiziert: Wie soll gemessen werden?

Die Forschungsliteratur zu Metriken beschränkt sich bisher im Wesentlichen auf ihre Spezifikation für die gängigen intrinsischen Qualitätsdimensionen Korrektheit, Konsistenz, Aktualität und Vollständigkeit. Diese messen syntaktische oder semantische Qualitätsmerkmale der Daten unabhängig vom jeweiligen Datennutzenden und Anwendungszweck, weswegen solche Metriken auch als „objektive Metriken" bezeichnet werden. Manche Qualitätsdimensionen sind jedoch pragmatisch, d. h. abhängig vom Anwendungskontext wie die Verständlichkeit oder Glaubwürdigkeit von Daten. Sie können nicht objektiv gemessen, sondern nur über „subjektive Metriken" validiert werden, z. B. durch Benutzerbefragungen.

Die objektiven Metriken für intrinsische Qualitätsdimensionen können zum Teil regelbasiert evaluiert werden und sind mittlerweile die Basis vieler kommerzieller und akademischer Werkzeuge zur automatisierten Hebung der Datenqualität (Data Correction) und zur automatisierten Datenreparatur (Data Repairing). Neben solchen Metriken für die gängigen Dimensionen der Datenqualität werden in diesem Zusammenhang oft auch statische Kennzahlen wie Mittelwerte und Streuungsmaße als deskriptive Metriken hinzugezogen, die im Rahmen von Qualitätsprüfungen von Bedeutung sein können. Zum Teil erfordert aber auch die Messung intrinsischer Datenqualität die konkrete Spezifizierung einer Evaluationsmethodik im Einzelfall (⟶ siehe „Die Spezifikation von Metriken und Evaluationsverfahren").

Für den Weg von der Dimension zur Metrik existiert in der Literatur kein generalisierbares Vorgehensmodell. Dies liegt u. a. daran, dass sich die Metriken aufgrund des weiten Spektrums an Datenqualitätsdimensionen in ihrer Natur unterscheiden. Trotz der Unterschiede lässt sich ein Basisschema mit drei Ebenen für die Spezifikation von Metriken aus der Literatur ableiten:

- Festlegung und Priorisierung der Anforderungen an die Metrik
- Definition der Metrik
- Festlegung der Evaluationsmethodik

Die Anforderungen an eine Metrik ergeben sich aus dem Nutzungskontext. Heinrich et al. (2018a) haben für einen wirtschaftlichen Kontext fünf generelle Anforderungen identifiziert. Hierzu gehören:

- Die Existenz eines Minimums und eines Maximums.
- Die Intervallskalierbarkeit der Metrikwerte, die eine Normierung auf das Intervall [0,1] ermöglichen. Dies vereinfacht die Interpretation und ermöglicht eine Aggregation verschiedener Metriken (z. B. als KPI, Key Performance Indicator).
- Die Anwendbarkeit der Metrik auf unterschiedliche Granularitätsebenen von Daten (Einzelwerte, Attribute, Tuples/Samples, Gesamtdatensatz) sowie die konsistente Aggregierbarkeit über die Granularitätsebenen hinweg. Dies ist insbesondere im Hinblick auf die Nutzung für spätere Entscheidungsfindungsprozesse wichtig.

Diese drei Anforderungen betreffen die Definition der Metrik selbst, welche sowohl eine qualitative Beschreibung als auch eine quantitative Beschreibung in Form einer mathematischen Formel beinhaltet. Die weiteren Anforderungen hingegen betreffen die Evaluationsmethodik:

- Aus dem Nutzungskontext ergeben sich Anforderungen bezüglich der Zuverlässigkeit der Metrikwerte, welche z. B. durch die Kombination mehrerer Messmethoden erhöht werden kann.
- Die ökonomische Effizienz stellt die fünfte Anforderung an eine Metrik dar, d. h. die Kosten, die bei der Erhebung anfallen, sollten in Relation zum Nutzen stehen.

Im Rahmen der Spezifikation einer Metrik müssen die Anforderungen nicht nur festgelegt, sondern gegebenenfalls auch iterativ abgewogen und priorisiert werden (siehe „Die Spezifikation von Metriken und Evaluationsverfahren“).

DIE SPEZIFIKATION VON METRIKEN UND EVALUATIONSVERFAHREN

Beispiel: In einem Onlineshop soll ein Algorithmus aufgrund der Haarfarbe von Kundinnen und Kunden Kaufempfehlungen für Kleidungsstücke generieren. Daher muss die Korrektheit der unter „Haarfarbe“ hinterlegten Daten in den Kundenprofilen gemessen werden (⟶ siehe Abbildung 5).

Da die Metrik als Teil eines KPI genutzt werden soll, ergeben sich die von Heinrich et al. (2018b) formulierten formalen Anforderungen an die Metrik (Existenz von Minimum und Maximum, Intervallskalierbarkeit und Aggregierbarkeit). Für Nominaldaten wie Haarfarbe, die keine natürliche Reihenfolge oder Abstände zueinander haben, kann die Korrektheit über das Verhältnis korrekter Werte zu inkorrekten Werten berechnet werden. Für andere Daten wie beispielsweise die Körpergröße von Kundinnen und Kunden könnte Korrektheit auch über eine Formel gemessen werden, die anhand einer Distanzfunktion (wie viele Zentimeter liegt der Wert neben dem tatsächlichen Wert?) berücksichtigt, wie inkorrekt ein Wert ist.

Als Evaluationsmethodik für die Qualitätsdimension Korrektheit bieten sich laut Rapp (2020) grundsätzlich vier Methoden an:

- der Abgleich mit Referenzdaten,
- die Verwendung von Geschäftsregeln,
- die Ableitung von automatisierten Regeln (Assoziationsregeln) über Data Mining
- sowie die automatisierte, KI-basierte Gewinnung von Referenzdaten.

Für den Onlineshop werden zwei Evaluationsmethoden in Betracht gezogen. KI-Modelle, mit denen die Haarfarbe der Kundinnen und Kunden aus ihrem Profilfoto abgeleitet werden kann, stellen eine Geschäftsregel dar. Diese Methode ist zwar günstig, aber ungenau. Wenn zusätzliche Referenzdaten erhoben werden, kann die Zuverlässigkeit der Metrik erhöht werden. Es könnte etwa eine E-Mail an die Nutzenden gesendet werden, bei denen der KI-Algorithmus einen inkorrekten Eintrag vermutet, um nachzufragen, ob die Haarfarbeninformation

stimmt und so die Anzahl fälschlich als inkorrekt eingestufter Datenwerte zu verringern. Dies wäre allerdings mit Kosten verbunden. Zudem könnten sich Kundinnen und Kunden gestört oder in ihrer Privatsphäre verletzt fühlen.

Die Auswahl der Evaluationsmethodik wird durch die von (Heinrich et al. 2018b) formulierten Anforderungen an die Genauigkeit der Metrik und die Kosteneffizienz ihres Einsatzes geleitet. Referenzdaten zur Messung der Korrektheit scheiden in vielen Fällen aus, da sie nicht vorliegen oder ihre Beschaffung mit einem hohen Aufwand verbunden ist. Um zu entscheiden, ob sich die Erhebung lohnt, wäre im Beispiel eine Abschätzung der Umsatzverluste notwendig, die dem Unternehmen aufgrund inkorrekter Augenfarbendaten entstehen. Ein Ansatz, um die wirtschaftlichen Auswirkungen mangelhafter Datenqualität unter Nutzung konkreter Metriken abzuschätzen, wird beispielhaft in BVDW 2018 dargestellt.

Die Spezifikation von Metriken und Evaluationsmethoden sowie die Entscheidung, ob sich die Erhebung lohnt, erfordern somit technische, formale und wirtschaftliche Abwägungen.

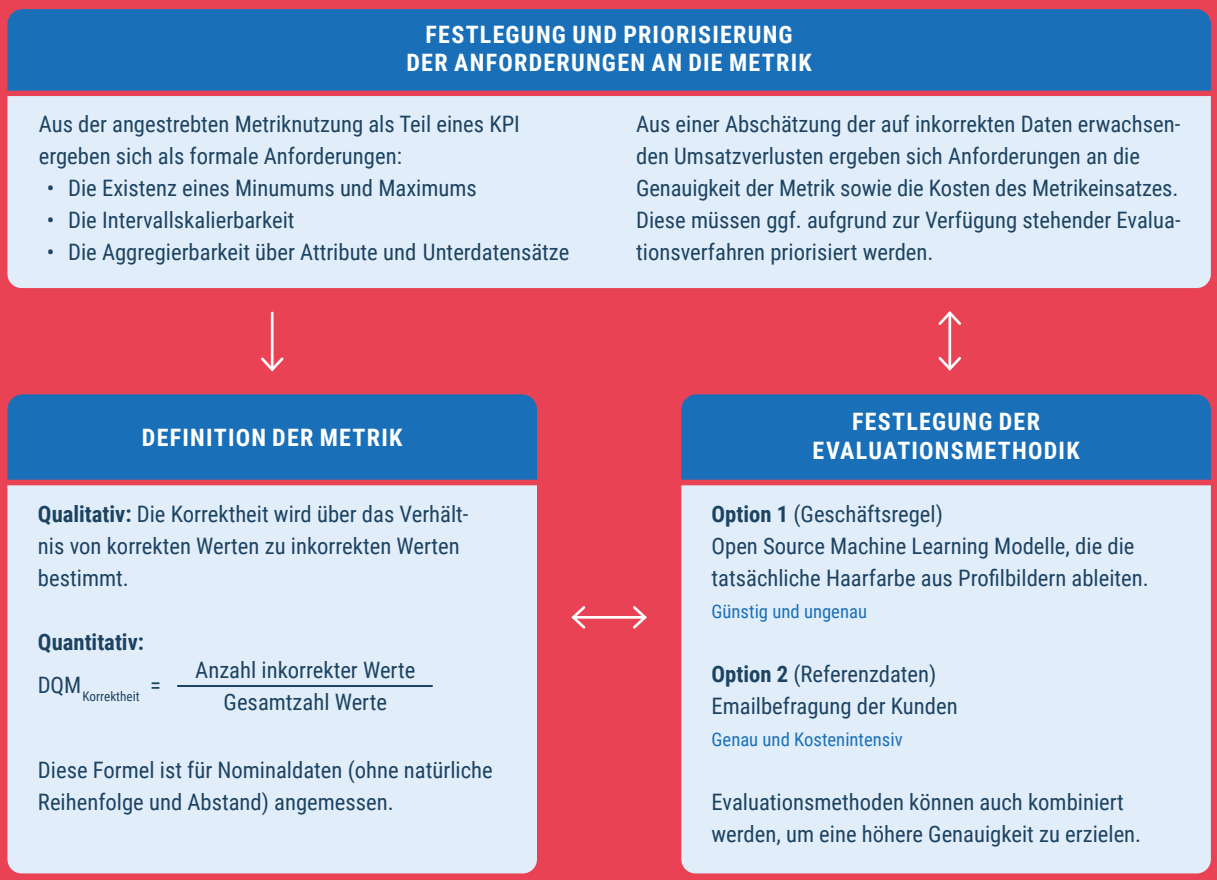

Abbildung 5: Illustration der Spezifikation einer Metrik für die Datenqualitätsdimension „Korrektheit“, angelehnt an das von Rapp (2020) aufgestellte Evaluationsschema und den von Heinrich et al. (2018b) dargelegten Prozess zur Festlegung der Anforderungen

Im konkreten Anwendungsfall wird Datenqualität meist nicht nur anhand einer Qualitätsdimension gemessen, sondern es werden mehrere Qualitätsdimensionen und Unterdimensionen als für den Anwendungsfall relevant identifiziert. Datenqualität wird somit durch Sammlungen von Metriken und mehrdimensionale Metriken charakterisiert (⟶ siehe „Mehrdimensionale Metriken und Überschneidung von Metriken“). Die Bestimmung der relevanten Dateneigenschaften, Qualitätsdimensionen und konkreten Metriken hängt von den Dateninhalten und dem Anwendungsfall ab und ist somit nicht allgemeinhin standardisierbar.

MEHRDIMENSIONALE METRIKEN UND ÜBERSCHNEIDUNG VON METRIKEN

Rapp hat eine umfassende und detaillierte Übersicht zu Metriken für die Qualitätsdimensionen Korrektheit, Vollständigkeit, Aktualität und Konsistenz sowie entsprechenden Evaluationsmethoden zusammengestellt. Datenqualitätsdimensionen wie die Korrektheit und die Konsistenz setzen sich aus mehreren Unterdimensionen zusammen, weswegen auch die entsprechenden Metriken häufig mehrdimensional sind. Im Fall der Korrektheit kann z. B. zwischen der syntaktischen und der semantischen Korrektheit unterschieden werden. Die Konsistenz gliedert sich in die Integrität, die repräsentative Konsistenz und die semantische Konsistenz. Setzt sich eine Metrik aus mehreren Untermetriken zusammen, muss gegebenenfalls festgelegt werden, wie die Untermetriken aggregiert werden, z. B. über Mittelwertbildung oder Gewichtung (siehe das Beispiel in Abbildung 6).

Unterdimensionen von Metriken können sich zudem überschneiden: Im Beispiel ist ein fehlerhafter Eintrag in der Datenbank hinterlegt, der sowohl einen Mangel der semantischen Konsistenz darstellt (es kann keine zwei Städte in Deutschland geben, die die gleiche Vorwahl haben) als auch einen Mangel der semantischen Korrektheit (Rom ist keine Stadt in Deutschland). Beide können teils mit denselben Regeln identifiziert werden. Bei der Spezifikation von Metriksammlungen und mehrdimensionalen Metriken sind solche möglichen Überlappungen zwischen Metriken bzw. Untermetriken zu berücksichtigen, z. B. indem eine Metrik vernachlässigt wird (Rapp 2020).

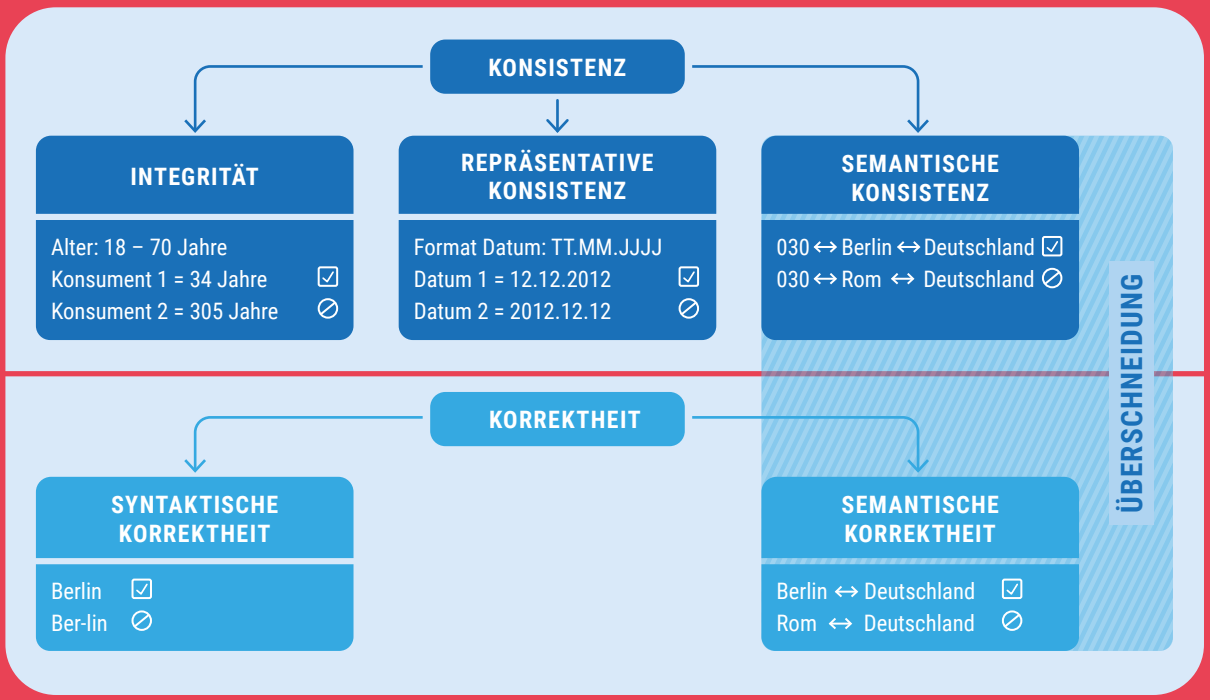

Abbildung 6: Überschneidung mehrdimensionaler Metriken am Beispiel der Dimensionen Konsistenz und Korrektheit, in Anlehnung an (Rapp 2020)

2.2.4 ANFORDERUNGEN AN EIN RAHMENKONZEPT FÜR DATENQUALITÄT IM DATENHANDEL

Bisher hat sich noch kein Rahmenkonzept für Datenqualität im Datenhandel etabliert. Auch wenn das Thema bereits an einigen Stellen in der Literatur diskutiert wird (Lawrenz et al. 2022; Zhang et al. 2018), besteht hier eine Forschungslücke. Künftige Entwicklungen auf diesem Gebiet könnten auf den im vorherigen Abschnitt diskutierten Rahmenkonzepten aufbauen, insbesondere auf dem konsolidierten Rahmenkonzept der DAMA NL (Black und van Nederpelt 2020). Dabei sollten jedoch einige Besonderheiten des Datenhandels berücksichtigt werden, die in ähnlicher Form so auch für Data Sharing und Open Data gelten.

Eine implizite Grundannahme der meisten Rahmenkonzepte für Datenqualität ist, dass die Konsumenten der Daten gleichzeitig ihre Halter und Erzeuger sind. Das Datenqualitätsmanagement erfolgt in diesem Szenario entlang etablierter Richtlinien, Aufbauorganisationen und Data-Governance-Frameworks und als fortlaufender, top down-gesteuerter Prozess. Im Datenhandel besteht allerdings eine Informationsasymmetrie zwischen Datennehmenden und Datengebenden, die eine eigene Prüfung wichtiger Qualitätsmerkmale für die Datenkonsumierenden schwierig oder unmöglich macht. Der Bereitstellung von Metriken kommt somit eine Schlüsselrolle in der Präsentation von Datenangeboten zu.

IM DATENHANDEL BESTEHT EINE INFORMATIONSASYMMETRIE ZWISCHEN DATENNEHMENDEN UND DATENGEBENDEN, DIE QUALITÄTSPRÜFUNGEN ERSCHWERT.

Welche Datenattribute und Qualitätsdimensionen wichtig sind und welche Anforderungen diesbezüglich bestehen, ist jedoch anwendungsfallspezifisch. Es gibt bisher kaum Literatur, die darlegt, wie Metriken und Sammlungen von Metriken im konkreten Anwendungsfall ausgewählt werden können, wie Messmethoden spezifiziert werden können und welche Metriken sich für bestimmte Anwendungsgebiete eignen. Einen Ansatzpunkt hierfür bietet eine Studie von Heinrich et al. zu ökonomisch motivierten Metriken, in der untersucht wird, in welchen praktischen Szenarien welche Anforderungen für Datenqualität besonders relevant sind (Heinrich et al. 2018a). Auch wurden Werkzeuge für manche Anwendungsgebiete entwickelt, wie das „Data Quality for AI Toolkit" von IBM[5,6], das verschiedene Metriken für Datenqualität speziell für KI-Anwendungen definiert. Dort sind beispielsweise Maße für Eigenschaften wie die Klassenüberlappung, die Klassenparität oder die Reinheit der Labels (d. h. der Werte eines Datenattributs, das ein maschinelles Lernmodell einschätzen soll) definiert, welche einen Einfluss auf die Leistung von KI-Modellen haben und somit im Zusammenhang von KI-Nutzung als Qualitätseigenschaften zu betrachten sind. In vielen Bereichen haben sich jedoch noch keine Standards oder Verfahren etabliert, anwendungsbezogene Datenqualität zu messen, weswegen aus dem SDW-Projekt Future Data Assets heraus gefordert wird, dass Metriksammlungen für verschiedene Datentypen und Anwendungen entwickelt werden (Band et al. 2022).

5 https://future-data-assets.de/ (Abruf am 08.03.2022).
6 https://developer.ibm.com/apis/catalog/dataquality4ai--data-quality-for-ai/Introduction (Abruf am 08.03.2022).

Erhöhte Bedeutung der pragmatischen Datenqualitätsdimensionen. Im Datenhandel rücken Qualitätsdimensionen in den Fokus, die in der Literatur bisher noch nicht oder kaum berücksichtigt werden. Bei der Sichtung von Angeboten im Datenhandel stellt sich in der Regel nicht als Erstes die Frage, wie korrekt oder konsistent die Daten sind, sondern ob sie im Hinblick auf den angedachten Anwendungszweck überhaupt von Interesse bzw. nützlich sind. Es ist zu erwarten, dass angesichts der Informationsasymmetrie zwischen Datengebenden und Datennehmenden pragmatische Dimensionen wie Glaubwürdigkeit und Relevanz eine zentrale Rolle spielen werden. Datenkonsumenten schreiben ihnen ohnehin auch jetzt schon in empirischen Studien allgemein eine hohe Bedeutung zu (Wang und Strong 1996).

IM DATENHANDEL WERDEN PRAGMATISCHE DIMENSIONEN WIE GLAUBWÜRDIGKEIT UND RELEVANZ EINE ZENTRALE ROLLE SPIELEN.

Für die Erhebung von pragmatischen Qualitätsdimensionen gibt es allerdings noch keine anerkannten Metriken, die sich aus der Literatur ableiten lassen. Insbesondere die Dimension Glaubwürdigkeit wird in der Literatur kontrovers diskutiert: Es herrscht kein Konsens bezüglich der Unterdimensionen, die ihr zugeordnet sind. Erste Ansätze zur Erhebung der Glaubwürdigkeit sind vorhanden (Prat und Madnick 2007; Moossavizadeh et al. 2012). Datenqualitätsdimensionen wie der Standardisierungsgrad und die Konformität bzw. Auditierbarkeit (Compliance) haben auch einen Einfluss auf die Glaubwürdigkeit der Daten, ebenso wie deren Vertraulichkeit und die Transparenz zur Herkunft und möglichen Transformationsschritten, die die Daten unterlaufen haben (Provenance). Für Glaubwürdigkeit sowie die Unterdimensionen existieren noch keine anerkannten Metriken.

Beachtung der Wahrnehmungsebene von Datenqualität. Zudem muss im Datenhandel sehr viel mehr als bei der internen Datennutzung die Beschaffenheit der Daten für Außenstehende nachvollziehbar dargestellt sein. Damit erhalten die sogenannten Wahrnehmungsebenen der Datenqualität eine größere Bedeutung. Daten werden selten direkt, sondern in der Regel über entsprechende Abfrageschnittstellen und Benutzeroberflächen konsumiert, die einen Einfluss auf die Wahrnehmung der Datenqualität haben. Laut Redman und Fürber sind im Hinblick auf Datenqualität grundsätzlich vier Wahrnehmungsebenen zu unterscheiden (Redman 2001; Fürber 2016):

- Die Datenebene umfasst die reinen Daten in Form von Werten.
- Auf der Datenmodellebene ist das Datenschema u. a. mit Struktur der Daten, Integritätsbedingungen, Inferenzregeln und Metadaten verortet.
- Die Präsentationsebene umfasst die Elemente, die den Nutzenden die Inhalte der Datenebene oder der Datenmodellebene vermitteln, wie z. B. die Benutzeroberfläche.
- Die Zugangsebene enthält alle Elemente, die den Zugang der Nutzenden regeln, z. B. über Autorisierungen.

Dimensionen der Datenqualität können auf den verschiedenen Ebenen zu unterschiedlichen Qualitätsproblemen führen: Die Korrektheit der Daten auf Datenebene ist beispielsweise etwas anderes als die Korrektheit der Metadaten auf der Datenmodellebene. Rahmenkonzepte für Datenqualität sollten daher nicht nur Qualitätsdimensionen benennen, sondern auch definieren, auf welchen Wahrnehmungsebenen die jeweiligen Dimensionen verortet sind. Es ist zudem auszugehen, dass beim Datenhandel die Präsentationsebene einen großen Einfluss auf die Wahrnehmung von

Datenqualität haben wird. Obwohl dies bislang noch nicht empirisch unterlegt ist, könnten daher bisher wenig berücksichtigte Dimensionen wie Ästhetik, Bequemlichkeit und Interaktivität an Relevanz gewinnen.

Hohe Bedeutung von Repurposed Data. Nicht zuletzt stellt die häufig vorkommende Verwendung von Datensätzen für nicht von vornherein vorhergesehene Zwecke (Repurposed Data) den Datenhandel vor Herausforderungen. In diesem Fall muss die Datenqualität zu einem Zeitpunkt hergestellt werden, an dem die Nutzungsanforderungen, zumindest teilweise, noch unbekannt sind. Das steht der Anforderung entgegen, Datenangebote optimal auf den Nutzungszweck zuzuschneiden. Zhang et al. (2019) haben hierfür einen vom semiotischen Rahmenmodell abgeleiteten Datenqualitätsmanagement-Ansatz entwickelt, der bottom-up vorgeht. Er befasst sich mit Datenqualität zunächst ausschließlich auf der Ebene der Syntax und anschließend auf der Ebene der Semantik. Auf beiden Ebenen können Aspekte der Datenqualität (z. B. Korrektheit, Konsistenz) in Unkenntnis der Nutzungsanforderungen bewertet werden. Eine zweckbezogene, pragmatische Beschäftigung mit Datenqualität erfolgt erst dann, wenn die Daten einem konkreten Anwendungszweck zugeführt werden. Im Datenhandel könnten Elemente eines solchen Vorgehens ebenfalls sinnvoll sein, um auch auf die Möglichkeit der zweckfremden Nutzung eines Datenangebots einzugehen.

Rechtssicherheit als Qualitätseigenschaft. Auch die Rechtssicherheit des Einsatzes von gekauften Daten ist eine Qualitätseigenschaft, der im Datenhandel erhöhte Bedeutung zukommen wird. Der folgende Abschnitt geht daher vertieft auf die rechtlichen Rahmenbedingungen der Datenwirtschaft ein.

2.3 Rechtliche Rahmenbedingungen

Die Möglichkeiten und Grenzen der Datenwirtschaft werden maßgeblich durch den Rechtsrahmen vorgegeben. Die bestehende Rechtsordnung sieht zwar kein Eigentumsrecht oder ein vergleichbares absolutes Recht an Daten vor, dennoch gibt es eine Reihe von Vorschriften, die den Umgang und die Nutzungsmöglichkeiten von Daten beeinflussen und damit beim Handel mit Daten berücksichtigt werden müssen. Neben dem Datenschutzrecht sind insbesondere die Vorgaben des Urheberrechts und die Vorschriften zum Schutz von Geschäftsgeheimnissen relevant. Wurden diese Vorgaben bei der Erstellung eines Datensatzes berücksichtigt, sodass die Nutzung rechtssicher ist, ist das eine wichtige Qualitätseigenschaft von Daten (Black und van Nederpelt 2020) und stiftet Vertrauen. Diese Qualitätseigenschaft ist nicht messbar, sie kann nur durch Expertinnen und Experten eingeschätzt oder durch Datengebende rechtlich zugesichert werden. Zum anderen bietet der Rechtsrahmen auch Raum für Regelungen zum Umgang mit Datenqualität. Tritt ein Schadensfall aufgrund mangelhafter Datenqualität auf, stellt sich die Frage, welcher Akteur einer Datenwertschöpfungskette haftbar ist. Um dies vorab zu klären, bietet das Vertragsrecht einen weiten Gestaltungsspielraum, was sich ebenfalls positiv auf das Vertrauensverhältnis im Datenhandel auswirken kann. Dieser Abschnitt fasst die wichtigsten Punkte der für die Datenqualität in der Datenwirtschaft relevanten Gebiete der Gesetzgebung zusammen.

2.3.1 DATENSCHUTZRECHT

Das Datenschutzrecht dient dem Schutz natürlicher Personen bei der Verarbeitung von personenbezogenen Daten. Das in Art. 8 der EU-Grundrechtscharta verankerte Recht auf Datenschutz wird vorrangig durch die Vorschriften der EU-Datenschutz-Grundverordnung (DSGVO) vermittelt. Die datenschutzrechtlichen Bestimmungen finden Anwendung bei der Verarbeitung von personenbezogenen Daten. Als personenbezogen gilt eine Information, die sich auf eine identifizierte oder identifizierbare natürliche Person bezieht.[7] Die datenschutzrechtlichen Vorschriften wirken sich beschränkend auf den Datenhandel aus und müssen auf jeder Stufe der Wertschöpfungskette von der Erhebung, Weitergabe, Weiterverarbeitung bis hin zur Löschung berücksichtigt werden.

— DAMIT DATEN RECHTSKONFORM NUTZBAR SIND, MÜSSEN ALLE AKTEURE ENTLANG DER WERTSCHÖPFUNGSKETTE DIE EINHALTUNG DER DATENSCHUTZRECHTLICHEN VORGABEN SICHERSTELLEN. —

Personenbezogene Daten müssen zunächst rechtmäßig erhoben werden. Notwendig ist eine Rechtsgrundlage in Form einer Einwilligung oder eines anderen gesetzlich normierten Erlaubnistatbestands. Werden Daten auf Grundlage einer Einwilligung gehandelt, so hängt die Rechtmäßigkeit der Datenverarbeitung vom Bestand der Einwilligung ab. Wird die Einwilligung widerrufen, ist der Verantwortliche der Datenverarbeitung zur Löschung der personenbezogenen Daten verpflichtet, sofern er die Weiterverarbeitung nicht auf eine andere Rechtsgrundlage stützen kann.

Die Weiterverarbeitung, Aufbereitung, aber auch die Auswertung und Weitergabe von personenbezogenen Daten ist zudem durch den Zweckbindungsgrundsatz limitiert. Dieser sieht vor, dass die Verarbeitung bzw. Weiterverarbeitung von personenbezogenen Daten im Einklang mit dem ursprünglichen Erhebungszweck stehen muss. Dies hat zur Folge, dass eine Sekundärnutzung von personenbezogenen Daten in den meisten Fällen ausgeschlossen ist.

Zudem muss der bzw. die für die Datenverarbeitung Verantwortliche gewährleisten, dass die datenschutzrechtlichen Vorgaben auch bei den Empfängern der personenbezogenen Daten erfüllt werden. Notwendig ist, dass alle Akteure entlang der Wertschöpfungskette die Einhaltung der datenschutzrechtlichen Vorgaben sicherstellen.

2.3.2 URHEBERRECHT

Das Urheberrecht dient dem Schutz von schöpferischen Leistungen und gewährt dem Rechte-Inhabenden ein zeitlich befristetes ausschließliches Verwertungsrecht an seinem Werk. Voraussetzung für den Urheberrechtsschutz ist das Vorliegen einer persönlichen geistigen Schöpfung.[8] Notwendig ist hierfür ein menschlicher Schaffensprozess. Aus diesem Grund können Urheber:innen maschinengenerierter Daten keinen Urheberrechtsschutz beanspruchen. Gleiches gilt für Einzeldaten. Daten können jedoch in Gestalt von Datenbankwerken Urheberrechtsschutz erlangen. Zudem besteht ein Leistungsschutzrecht zugunsten von Datenbankherstellern, sofern der Aufbau und die Unterhaltung der Datenbank mit einer erheblichen Investition verbunden sind.[9] Die ausschließlichen Nutzungsrechte an geschützten Datenbanken und Datenbankwerken stehen dem jeweiligen Rechteinhaber zu. Ein Handel mit urheberrechtlich geschützten Datenbanken bzw.

7 Art. 4 Nr. 1 DSGVO.
8 § 2 Abs. 2 UrhG.
9 § 87a UrhG.

Datenbankwerken ist in der Folge nur mit Zustimmung des Rechteinhabers möglich. Urheberrechtswidrige Verwertungshandlungen können untersagt werden und weitergehende Ansprüche aufseiten des Rechteinhabers begründen.

2.3.3 SCHUTZ VON GESCHÄFTSGEHEIMNISSEN

Datenhandel kann gerade im Geschäftskunden-Bereich auch sensible Unternehmensinformationen betreffen, wie z. B. die Betriebsdaten einer Maschine, die unter Umständen Rückschlüsse auf die Produktivität und interne Abläufe zulassen. Fallen die in Daten verkörperten Informationen in den Anwendungsbereich des Geschäftsgeheimnisschutzgesetzes (GeschGehG), können sich Beschränkungen hinsichtlich der Datennutzung ergeben. Datengebende Unternehmen müssen sicherstellen, dass die bereitgestellten Daten aufseiten des Datenempfängers durch angemessene Geheimhaltungsmaßnahmen geschützt werden. Datenempfangende Unternehmen müssen gewährleisten, dass die Nutzung und gegebenenfalls Weitergabe der Daten zu keiner unzulässigen Offenlegung von Geschäftsgeheimnissen führt, da ansonsten aufseiten des geschädigten Unternehmens Abwehransprüche entstehen können. Der Handel mit sensiblen Unternehmensdaten muss daher im Einvernehmen mit dem datengebenden Unternehmen erfolgen.

2.3.4 VERTRAGSRECHT

Abseits der dargestellten Beschränkungen ist die vertragliche Überlassung von Daten problemlos möglich. Auch ohne ein gesetzlich normiertes Schuldrecht steht es den Akteuren in datenbasierten Wertschöpfungssystemen frei, Art und Umfang der Datenüberlassung vertraglich festzulegen.

ES STEHT AKTEUREN IN DER DATENWIRTSCHAFT FREI, ART UND UMFANG DER DATENÜBERLASSUNG VERTRAGLICH FESTZULEGEN.

Dabei lassen sich konkrete Geschäftsmodelle für den Datenhandel (z. B. Festpreis, Pay-per-Use, Freemium usw.) zumindest mittelbar auf die im Bürgerlichen Gesetzbuch vorgesehenen Vertragstypen übertragen. Die Einordnung zu einem der gesetzlich vorgesehenen Vertragstypen ist insofern von Bedeutung, als dass sich hieraus konkrete Leistungspflichten ergeben und im Falle einer Pflichtverletzung unterschiedliche Rechtsfolgen vorgesehen sind. Bei der klassischen Form des Datenhandels (z. B. Kauf einer Adressdatenbank) wird in der Regel die dauerhafte Überlassung von Daten gegen Zahlung eines Kaufpreises vereinbart („Datenkauf"). Zwar beziehen sich die Regelungen des Kaufrechts auf Sachen. Daten gelten jedoch als „sonstige Gegenstände", auf die die Vorschriften über den Kauf entsprechend Anwendung finden.[10] Liegt der Schwerpunkt der Datenüberlassung in einer nur vorübergehenden Überlassung von Daten (sogenannte Data-as-a-Service-Angebote), sind hingegen die pachtrechtlichen Vorschriften des BGB sachnäher. Darüber hinaus können die Vertragsparteien aber auch einen Vertrag sui generis wählen, also eine von den gesetzlichen Leitbildern losgelöste vertragliche Vereinbarung.

2.3.5 HAFTUNG FÜR DATENQUALITÄT

Die Qualität von Daten ist auch aus juristischer Perspektive ein hochrelevantes Themenfeld, das vor allem dann Konfliktpotenzial birgt, wenn eine bestimmte Datenqualität nicht erreicht wird und in der Folge Nachteile aufseiten des datenempfangenden Unternehmens entstehen. Geht es um die Zu-

10 § 453 Abs. 1 BGB.

sicherung von bestimmten Qualitätskriterien, haben die Vertragsparteien naturgemäß gegenläufige Interessen. Datengebende Unternehmen wollen in der Regel nicht oder nur begrenzt für die Qualität ihrer Daten einstehen. Demgegenüber hat das datenempfangende Unternehmen ein Interesse an der Einhaltung von Qualitätsstandards, da hiervon der weitere Verwertungsprozess abhängt.

In zivilrechtlicher Hinsicht kann ein Unterschreiten der Datenqualität nachteilige Rechtsfolgen für das datengebende Unternehmen nach sich ziehen. Das datengebende Unternehmen ist verpflichtet, die vertragsgegenständlichen Daten frei von Sach- und Rechtsmängeln zu verschaffen.[11] Eine Sache ist frei von Sachmängeln, wenn sie bestimmten (gesetzlich geregelten) Anforderungen entspricht.[12] Unterschieden werden subjektive und objektive Anforderungen. Eine Sache entspricht den subjektiven Anforderungen, wenn sie die (vertraglich) vereinbarte Beschaffenheit hat.[13] Abzustellen ist also auf den Willen der Vertragsparteien, der in einer Beschaffenheitsvereinbarung in Erscheinung tritt. Im Rahmen einer solchen Beschaffenheitsvereinbarung können positive oder negative Leistungsstandards festgelegt werden (Sassenberg et al. 2020). Als positive Leistungsstandards können u. a. Anforderungen hinsichtlich der Datenqualität festgelegt werden. Mögliche Qualitätskriterien sind u. a. Fehlerfreiheit, Vollständigkeit, Datenkonsistenz, Vertrauenswürdigkeit, Aktualität, Zugänglichkeit, Compliance, Vertraulichkeit, Effizienz, Genauigkeit, Rückverfolgbarkeit, Verständlichkeit, Verfügbarkeit, Übertragbarkeit, Wiederherstellbarkeit der Daten (Auer-Reinsdorff und Conrad 2019). Ein Datensatz entspricht den subjektiven Anforderungen, wenn die Ist-Beschaffenheit der vereinbarten Soll-Beschaffenheit entspricht.

IM RAHMEN EINER BESCHAFFENHEITSVEREINBARUNG KÖNNEN POSITIVE ODER NEGATIVE LEISTUNGSSTANDARDS FESTGELEGT WERDEN.

Die Vertragsparteien können somit selbst definieren, wann und unter welchen Gegebenheiten Daten als mangelhaft einzustufen sind. Für die Feststellung eines Mangels bietet es sich an, neben den Datenqualitätskriterien auch Metriken zur Festlegung der Datenqualität zu bestimmen. Durch die Festlegung von Metriken zur Datenqualität lässt sich ein Unterschreiten von Standards leichter feststellen. Dies führt insgesamt zu mehr Klarheit und Rechtssicherheit für die Vertragsparteien. Neben der vertraglichen Festlegung von konkreten Qualitätsstandards entspricht eine Kaufsache auch dann den subjektiven Anforderungen, wenn sie sich für die nach dem Vertrag vorausgesetzte Verwendung eignet.[14] Abzustellen ist dabei auch die Eignung zum Vertragszweck. Erweisen sich daher die überlassenen Daten zur Erfüllung des Vertragszwecks als ungeeignet, liegt ein Mangel vor. Ein Beispiel wäre die Überlassung von fehlerhaften Trainingsdaten für eine KI-Anwendung.

Neben den subjektiven Anforderungen sieht das Kaufrecht auch objektive Kriterien zur Bestimmung der Mangelfreiheit vor. Abzustellen ist dann insbesondere auf die gewöhnliche Verwendung und die Beschaffenheit, die bei Sachen derselben Art üblich sind. Insgesamt lassen sich die objektiven Anforderungen der Mangelfreiheit in Bezug auf Daten schwieriger bestimmen. Gerade

11 §§ 453, 433 Abs. S. 2 BGB.
12 § 434 Abs. 1 BGB.
13 § 434 Abs. 2 Nr. 1 BGB.
14 § 434 Abs. 2 Nr. 2 BGB.

im Hinblick auf die „übliche Beschaffenheit" von Daten fehlt es an entsprechenden Vergleichsstandards für konkrete Anwendungsbereiche. Denkbar wäre aber die Heranziehung der bereits erläuterten ISO-Normen zur Datenqualität (⟶ siehe Abschnitt 2.2.3) bzw. die Anforderungen der Richtigkeit von Daten nach den datenschutzrechtlichen Vorschriften. Das Zurückfallen auf die objektiven Anforderungen dürfte in den meisten Fällen aber der Interessenslage der Vertragsparteien zuwiderlaufen und zu einer nicht unerheblichen Rechtsunsicherheit beitragen. Gerade vor dem Hintergrund der gesetzlich vorgesehenen Haftungsansprüche liegt es allermeist im Interesse der Vertragsparteien, die Datenqualitätsmerkmale selbst zu bestimmen.

2.3.6 DIE HAFTUNGSRECHTLICHE POSITION VON INTERMEDIÄREN

Intermediäre nehmen in der Datenwertschöpfungskette eine vermittelnde Rolle ein. Sie beteiligen sich als verarbeitende Intermediäre selbst am Wertschöpfungsprozess, indem sie Daten aus unterschiedlichen Quellen verarbeiten und in Form von Datendiensten oder Datenprodukten vertreiben. Demgegenüber handelt es sich bei ermöglichenden Intermediären um Unternehmen, die keine eigenen Dienste oder Produkte anbieten. Ihre Leistung besteht darin, die Datenwertschöpfung zu ermöglichen oder zu unterstützen, und sie stellen hierfür die notwendigen Infrastrukturen bereit.

Aus den unterschiedlichen Betriebsmodellen ergeben sich naturgemäß auch verschiedene Leistungspflichten. Verarbeitende Intermediäre nehmen Einfluss auf die Form von Datendiensten oder Datenprodukten. Sie treten am Markt als Anbietende dieser Dienste oder Produkte auf. Bei der Überlassung von fehlerhaften Datenbeständen haften sie nach den kauf- bzw. pachtrechtlichen Vorschriften. Die dort vorgesehenen Mängelrechte, z. B. Nacherfüllung oder Minderung, entstehen verschuldensunabhängig. Das bedeutet, dass der Intermediär für den Mangel einzustehen hat, unabhängig davon, ob er den Mangel zu vertreten hat. Anders verhält es sich, wenn der Schwerpunkt der Leistungspflicht nicht in der Überlassung von Daten, sondern in der Erbringung einer Beratungsdienstleistung, z. B. eines konkreten Datendienstes, liegt. In diesem Fall wird ein Dienstleistungsvertrag anzunehmen sein. Anders als im Kaufrecht- oder Pachtvertragsrecht sind im Dienstvertragsrecht keine Mängelgewährleistungsansprüche vorgesehen. Der verarbeitende Intermediär haftet damit nach den Vorschriften des allgemeinen Schuldrechts.[15] Kommt es infolge einer fehlerhaften Beratungsdienstleistung zu Schäden, kann der Intermediär auf Schadensersatz in Anspruch genommen werden.

Grundlegend anders verhält es sich beim ermöglichenden Intermediär. Seine Leistungspflicht besteht lediglich in der Vermittlung und Bereitstellung von Daten Dritter. Er haftet folglich auch nur für Schlechtleistungen im Zusammenhang mit der (technischen) Bereitstellung von Daten. Anders als der verarbeitende Intermediär macht er sich Daten Dritter nicht „zu eigen" und muss folglich auch nicht für deren Fehlerfreiheit einstehen.

15 § 280 ff. BGB.

2.3.7 GESETZLICHE REGELUNGEN ZUR DATENQUALITÄT

Datenqualität wird aus rechtlicher Perspektive primär über das Vertragsrecht festgelegt. Daneben gibt es eine Reihe von gesetzlichen Regelungen, welche ebenfalls Anforderungen an die Qualität von Daten stellen:

- In Bezug auf digitale Inhalte hat der Gesetzgeber infolge der Digitale-Inhalte-Richtlinie[16] ein bereichsspezifisches Schuldrecht geschaffen. Es betrifft Verträge, die digitale Inhalte zum Gegenstand haben. Bei digitalen Inhalten handelt es sich um Daten, die in digitaler Form erstellt und bereitgestellt werden.[17] Nach § 327e Abs. 1 BGB ist ein digitales Produkt frei von Mängeln, wenn es den subjektiven und objektiven Anforderungen und den Anforderungen an die Integration entspricht. Bei den subjektiven Anforderungen wird auf die vereinbarte Beschaffenheit bzw. auf die Eignung für die im Vertrag vorausgesetzte Verwendung abgestellt. Im Hinblick auf die objektiven Anforderungen wird u. a. an die Beschaffenheit einschließlich der Kompatibilität und der Zugänglichkeit von digitalen Produkten gleicher Art und Güte angeknüpft. Zu beachten ist jedoch, dass die schuldrechtlichen Bestimmungen über digitale Inhalte im Geschäftskunden-Bereich keine Anwendungen finden.
- Der aktuelle Entwurf der EU-KI-Verordnung sieht für sogenannte Hochrisiko-KI-Systeme eine Reihe von Mindestanforderungen vor, die durch die Anbietenden der Anwendung umgesetzt werden müssen. Hierzu gehören auch Anforderungen an die Qualität der Trainingsdatensätze. Insbesondere sollen die Trainings-, Validierungs- und Testdatensätze im Hinblick auf die Zweckbestimmung des Systems hinreichend relevant, repräsentativ, fehlerfrei und vollständig sein (Europäische Kommission 2021).
- In Bezug auf personenbezogene Daten sieht Art. 5 Abs. 1 lit. d) DSGVO vor, dass diese sachlich richtig sein müssen. Das bedeutet, dass die über eine Person gespeicherten Informationen mit der Realität übereinstimmen müssen. Die betroffene Person hat das Recht, von dem bzw. der datenschutzrechtlich Verantwortlichen die Berichtigung der unrichtigen personenbezogenen Daten zu verlangen.[18]

16 Richtlinie (EU) 2019/770.
17 vgl. § 327 Abs. 2 BGB.
18 Art. 16 DSGVO.

03

3 DIE PRAXIS

Auf Basis der Literaturrecherche wurden von Juli bis Dezember 2021 elf leitfadengestützte Interviews mit Experten aus der Praxis geführt. Besprochen wurde zunächst die heutige Situation von Datenwirtschaft und Datenhandel in Deutschland, um auch das Umfeld in den Blick zu nehmen. Hauptthema war dann der gegenwärtige Umgang mit Datenqualität und Qualitätsmetriken in diesem Feld. Erfragt wurden auch Zukunftserwartungen, jeweils für Datenwirtschaft und Datenqualität. Die Ergebnisse wurden in einem Workshop im Januar 2022 mit mehreren Interviewpartnern sowie dem Leiter der Arbeitsgruppe Data Economy des KI Bundesverbands validiert und vertieft.

3.1 Datenwirtschaft und Datenhandel

Die meisten Experten beschrieben die gegenwärtige Situation in der Datenwirtschaft als wenig zufriedenstellend, sahen aber auch positive Tendenzen wie eine relevante Nachfrage nach Daten und bereits etablierte Teilmärkte.

3.1.1 DATENSILOS UND FEHLENDE DATENKULTUR

Insbesondere jene Befragten, die für oder mit Unternehmen des produzierenden Gewerbes, der Logistik oder des Handels arbeiten, berichteten, dass in den IT-Systemstrukturen der datengebenden Unternehmen Datensilos gängig sind, was schon die technische Integration verschiedener Daten in die Wertschöpfungskette erschwere. Viele Datengebende scheuten die Aufwände und Kosten, ein modernes Datenmanagementsystem aufzusetzen. Das erschwere in der Praxis den effizienten Umgang mit den meist heterogenen Datensätzen und wurde von der Mehrheit der Experten als wirtschaftliches Versäumnis der datengebenden Unternehmen eingestuft.

„Datenökonomie ist auch ein Weg, auf den man sich machen muss und der viel Zeit in Anspruch nimmt."

Tobias Manthey, Evotegra GmbH und Leiter der Arbeitsgruppe Data Economy im KI Bundesverband

Damit einhergehend sei in vielen datengebenden oder potenziell datengebenden Unternehmen eine fehlende „Datenkultur" zu beobachten. Dies betreffe zum einen die mangelnde Bereitschaft, relevante, aber sensible Daten mit Externen zu teilen – seien es Dienstleister, Zulieferer, Lieferanten oder Handelspartner. Zum anderen berichteten alle Interviewpartner, dass es in vielen Unternehmen den Entscheidungstragenden und Domänenexperten an Datenkompetenz (Data Literacy) mangele, was es den Unternehmen erschwere, Wertschöpfungspotenziale zu erkennen und zu nutzen. Gerade im Mittelstand sei der Umgang mit Daten „unbedarft" und „restriktiv" und der Wert von Daten werde häufig nicht erkannt. Unter Hinweis auf interne Regelungen zur Vertraulichkeit der Geschäftsprozesse und zur Informationssicherheit würde das Teilen von Daten als Basis der Datenwertschöpfung oft begrenzt, vor allem das Teilen von Rohdaten. Die Daten müssen aufgrund der Regelungen oft lokal verbleiben; selbst eine interne Cloud-Nutzung würde von vielen Unternehmen abgelehnt. Ein Dienstleister berichtete sogar, dass seine Auftraggeber den Informationszugriff selbst für datenbasierte Dienstleistungen stark einschränken. Um auf derlei Sicherheitsbedenken und die Kosten einer Offenlegung einzugehen, bieten sowohl die IT-Dienstleister als auch die Plattformanbieter unter den Befragten technische Lösungen an, die auch lokal beim Datengebenden (on-premise) eingesetzt werden können. Die Aufwände, die bei der Erschließung aufgrund des restriktiven Umgangs mit den Daten und aufgrund mangelhafter Nutzbarkeit anfallen, würden sich negativ auf das Wertschöpfungspotenzial der Bestandsdaten auswirken.

„Das ganze Thema Datenaustausch und -handel im Maschinenbau und der Automobilindustrie ist extrem restriktiv. Meistens kriegen Sie von Ihrem Kunden kaum etwas. Es gehen keine Daten raus und rein kommt ihr schon gar nicht!"

Stephan Boch, KEB Automation KG

3.1.2 RECHTLICHE HERAUSFORDERUNGEN

Wenig überraschend nannten die Experten die bekannten rechtlichen Rahmenbedingungen (⟶ siehe Abschnitt 2.3) als wichtige Einflussfaktoren auf die Datenwirtschaft. Einige benannten in diesem Zusammenhang explizit den Schutz von Personendaten und die bestehenden Einschränkungen hinsichtlich der Nachnutzung. Ein Standardansatz für die Erfüllung der rechtlichen Datenschutzvorgaben sei die Anonymisierung der Daten, die zudem auch zum Schutz von sensiblen Betriebsdaten, etwa von Maschinendaten im produzierenden Gewerbe, herangezogen werden könne. Allerdings komme es durch die Anonymisierung auch zu Informationsverlusten.

Unter Hinweis auf die Datenschutz-Grundverordnung (DSGVO) der EU wiesen einige Interviewpartner mit unterschiedlichen Einschätzungen auf die internationale Situation zum Schutz von Personendaten hin. Ein verarbeitender Intermediär unterstrich, dass in Ländern mit geringeren Datenschutzvorgaben wie etwa den USA die Hürden, um beispielsweise neue KI-Lösungen auf der Basis personenbezogener Daten zu entwickeln, niedriger seien. Ein anderer verarbeitender Intermediär gab jedoch zu bedenken, dass es auch im Ausland wachsende Bestrebungen für mehr Datenschutz gebe. Der deutsche und europäische Datenschutz biete Rechtssicherheit hinsichtlich der künftigen Handelbarkeit von Daten und sei daher ein wirtschaftlicher Vorteil. Die meisten Befragten sehen die existierenden Regelungen insgesamt als einen notwendigen Faktor für den Schutz der Betroffenen sowie für die Rechtssicherheit an, obgleich die Einschränkungen für konkrete Anwendungsfälle erheblich sein können.

Eine weitere Herausforderung sei die besondere rechtliche Stellung von Daten als immaterielles Gut ohne Eigentumsrechte. Ein Plattformanbieter unter den Experten nannte als Beispiel maschinengenerierte Daten einer Industrieanlage, die etwa in Datenprodukte für das Benchmarking von Produktionsprozessen fließen könne. Hier müssten sich Hersteller und Betreiber einigen, wer wem welche Daten überlässt. Wenn mehrere Parteien an der Erstellung von Daten beteiligt seien, müsse meist eine individuelle vertragsrechtliche Lösung zur Datenüberlassung vereinbart werden, was Aktivitäten in der Datenwirtschaft zum Teil erheblich erschwere.

Thematisiert wurden auch Unsicherheiten bei der Haftung von Dienstleistungen in der Wertschöpfungskette. Manche Plattformbetreiber bieten keine eigenen Dienste zur Abschätzung der Datenqualität an, weil sie sonst rechtlich auch für die Dateninhalte haften müssten. Solche Angebote könnten zwar als Dienst für Datengebende auf Plattformen bereitgestellt werden, deren Nutzung würde jedoch spezialisiertes Wissen zum Datenmanagement erfordern, das nicht in jeder Branche bzw. jedem Unternehmen vorhanden sei und deshalb selten nachgefragt würde.

„Wir lassen [als Plattformbetreiber] jedem die Freiheit, seine Angebote und Qualität zu bestimmen und zu beschreiben. Naheliegend ist zu fragen: Sollte ein Marktplatz nicht Features bieten, die das zum Beispiel automatisiert überprüfen? Hier ist technisch einiges möglich. Es gibt jedoch rechtliche Fallstricke, die den Handlungsspielraum für Plattformbetreiber einschränken und die bei der genauen Ausgestaltung berücksichtigt werden müssen."

Sebastian Wiemann, T-Systems International GmbH

3.1.3 MANGEL AN DATENANGEBOTEN

Bei der Betrachtung des heutigen Datenhandels zeigten sich die Experten einig: Ein offener Datenhandel, etwa über Datenmarktplätze, sei in den meisten Branchen und Anwendungsdomänen noch nicht etabliert, insbesondere, wenn es nur um ein ergänzendes Nebengeschäft gehe. Wiederkehrende Motive in den Interviews waren fehlende Anreize zur Beteiligung an Data Sharing und Datenhandel, besonders für Datengebende.

Als einen Grund dafür nannten Interviewpartner, die in Branchen mit einem geringen Digitalisierungsgrad aktiv sind, dass für viele potenziell handelbare Datensätze noch die Anwendungsfälle fehlen würden, die das vorhandene Wertschöpfungspotenzial verdeutlichen. Der initiale Aufbau der Datenbasis sei kostspielig, generiere aber aus Sicht der Datengebenden zunächst keinen oder nur wenig Mehrwert, insbesondere wenn das Potenzial nicht für das Unternehmen, sondern nur für den eigenen Bereich betrachtet werde. Ein weiteres Argument war das mangelnde Vertrauen in die gegenwärtigen Datenplattformen. Die Bereitstellung eigener Daten für Dritte über offene Plattformen mit teilweise unbekannten Akteuren gehöre bisher nicht zur deutschen Unternehmenskultur.

Die Plattformbetreiber unter den Experten stellten heraus, dass der bilaterale Datenaustausch und der Datenaustausch innerhalb existierender Wertschöpfungsketten eine verhältnismäßig häufige Form sei, Datentransaktionen durchzuführen, teils zum gegenseitigen Nutzen, teils mit monetären oder nicht-monetären Vergütungen beim Datentausch, sodass der Übergang zum Datenhandel fließend sei (so auch die Befragungsergebnisse von Lindner et al. 2021). Bei der technischen Umsetzung werde zum Teil auf Plattformangebote zurückgegriffen.

Neben dem Datenaustausch in geschlossenen Netzwerken ist laut den Befragten auch Open Data eine verhältnismäßig häufige Form, Daten für die Wertschöpfung anzubieten, sowohl durch Unternehmen als auch durch öffentliche Anbietende. Gerade Letztere wurden von verarbeitenden Intermediären unter den Interviewpartnern, die als Konsumenten solcher Angebote auftreten, als ebenfalls problematisch thematisiert. Zwar stünden aus der Forschung und aus Behörden aufgrund von Auflagen zur Offenlegung zahlreiche potenziell wertvolle Datensätze zur Verfügung. Allerdings würden solche Daten häufig signifikante Qualitätsmängel aufweisen und seien nicht oder nur mit großen Aufbereitungsaufwänden erschließbar, da die Offenlegung nur erfolge, um Auflagen zu erfüllen, und keine Anreize bestünden, die Daten so aufzubereiten, dass eine Nachnutzung ermöglicht oder vereinfacht wird.

Auf bereits existierenden Marktplätzen sowie auch in etablierten Primärmärkten, d. h. Märkten mit einem bilateralen Austausch zwischen datenanbietenden und datenkaufenden Unternehmen, sei das Datenangebot zudem oft unübersichtlich. Vielfach würden unzureichend aufbereitete oder veraltete Produkte angeboten, Letzteres vor allem im Open-Data-Bereich. Preise kommerzieller Angebote wurden insbesondere von verarbeitenden Intermediären, die als Nachfragende im Datenhandel agieren, mehrfach als willkürlich und nicht angemessen eingestuft. Zudem gebe es gelegentlich zu wenig Bereitschaft, Datenangebote und Preise flexibel an den oft spezifischen Nutzungsanforderungen auszurichten. Gerade wenn Daten für nicht vorhergesehene Zwecke verwendet werden (Repurposed Data), sei das aber von hoher Bedeutung.

3.1.4 GROSSE DATENNACHFRAGE

Dem insgesamt ungenügenden und defizitären Angebot gegenüber steht laut Experten eine große, noch ungedeckte Nachfrage, die ihrer Vermutung nach in kommenden Jahren noch wachsen wird und als zentrale Triebkraft des Datenhandels zu betrachten ist. Generell gebe es mehr Marktpotenzial für auf konkrete Anwendungsfälle zugeschnittene Informationsangebote, die sich in der Preissetzung am Mehrwert bei den Kunden orientieren könnten, als für Datenangebote, die noch den Wertschöpfungsschritt der Informationsgewinnung durchlaufen müssen, um genutzt werden zu können. Solche Datenangebote würden künftig vermutlich kostenfrei oder zu verhältnismäßig geringeren, an den Kosten der Erstellung orientierten Preisen angeboten. Eine Produktpalette mit diversen datenbasierten Informations- und Lösungsangeboten hätte zudem für datengebende Unternehmen den Vorteil, dass ein Datensatz in vielen Marktnischen – zu an den jeweiligen wirtschaftlichen Nutzen angepassten Preisen und Anforderungen – immer wieder neu monetarisiert werden kann, da sich Daten uneingeschränkt teilen und nachnutzen lassen.

„Jedes der großen US-Unternehmen ist ein in sich geschlossener Wertschöpfungskreislauf, was sowohl die Erhebung als auch die Nutzung der Daten angeht. [...] Der krasse Gegensatz ist Europa. Durch die mittelständische Struktur im Bereich der KI und bei den Industrieunternehmen [...] gibt es einen natürlichen Anreiz zum Austausch und Handel von Daten."

Tobias Manthey, Evotegra GmbH und Leiter der Arbeitsgruppe Data Economy im KI Bundesverband

Zudem sahen die Experten auch die Nachfrage nach Datenangeboten, um sie neuen, vom Datenanbietenden nicht vorgesehenen Nutzungszwecken zuzuführen (Repurposed Data), als sehr marktrelevant an. Gerade bei Anwendungen für die Datenanalyse, wie etwa für die Beschaffung von Trainingsdaten bei der Entwicklung von KI-Anwendungen, sei das sehr relevant. Für dieses Nachfragesegment, aber nicht nur dort, sehen die Interviewpartner einen sekundären Datenhandel über Marktplätze, auf denen diverse Datenangebote bereitgestellt werden, als eine grundsätzlich sinnvolle Marktform an. Die verarbeitenden Intermediäre unter den Befragten, die sich selbst als potenzielle Nachfragende solcher Marktplätze einschätzten, stellen sich die Datenmarktplätze in der Zukunft domänenspezifisch vor, beispielsweise als Plattformen für Bild- und Mediendaten zur Nutzung in der Computerspiel- und Unterhaltungsbranche oder als Plattformen, die Sensordaten für Industrieanwendungen anbieten. Neben Trainingsdaten erwarten die Experten auch vorgelernte KI-Modelle und synthetische Daten als relevante Angebote im sekundären Datenhandel.

„Im Moment träumen wir alle von einer Datenökonomie, in der man nicht mehr zwingend aus der gleichen Nische kommen muss, um wechselseitig auf Daten zugreifen zu können. [...] Das schaffen wir nicht, indem wir nur darauf schauen, wie bereits existierende, vollkommen in sich geschlossene Märkte kommunizieren, dafür brauchen wir offene Standards."

Dennis Weber, Advaneo GmbH

„Wie strukturiert der Datenhandel ist, kommt sehr darauf an, in welche Ecke man schaut. […] Neben den neuen innovativen Analysefeldern gibt es auch sehr etablierte Geschäftsfelder für strukturierten Datenhandel, in denen man direkt operational etwas mit den Daten macht, wie klassischerweise der Adressdatenhandel für den Briefversand im Dialogmarketing."

Sebastian Wiemann, T-Systems International GmbH

3.1.5 BEREITS ERFOLGREICHE DATENMÄRKTE

Die Befragten wiesen darauf hin, dass neben diesen Zukunftsbetrachtungen in Teilbereichen der Datenwirtschaft bereits heute ein etablierter und wirtschaftlich tragfähiger Datenhandel zu beobachten sei. Das sei vor allem dann der Fall, wenn die Standardisierung von Daten und Anwendungsszenarien in den Domänen bereits weit fortgeschritten sei. Zu nennen seien insbesondere der Handel mit Adressdaten, Wetterdaten, Finanztransaktionsdaten und Bilddaten. Hier hätten sich bereits Primärmärkte gebildet.

Als weiteren Bereich großer, bereits erschlossener Datenmärkte führten mehrere Experten Unternehmen an, die Dienste über Plattformen anbieten und die entstehenden Betriebsdaten als Informationsdienste monetarisieren. Das betreffe nicht nur die bekannten großen Internetplattformen, sondern etwa auch Finanzdienstleister, die Transaktionsstatistiken verkaufen, oder Maschinenhersteller, die die Daten ihrer Maschinen bei den Kunden anonymisiert bündeln, um gegen Entgelt einen Benchmarking-Dienst anzubieten. Die Befragten waren sich einig, dass dieser Bereich der Datenwirtschaft in der öffentlichen Debatte weitgehend „unsichtbar" sei, weil die Unternehmen weder Interesse noch Bedarf an der Offenlegung ihrer Datennutzung sowie der Geschäftsstrategien und -prozesse haben, da sie zumeist die Datenwertschöpfung rein intern betreiben und daher nicht auf einen Datenhandel oder Datenaustausch angewiesen sind. Die datenwirtschaftliche Bedeutung werde daher oft unterschätzt. Kleineren Unternehmen würden jedoch oft die Ressourcen und die Reichweite fehlen, um in ähnlicher Weise hochwertige Dienste anzubieten. Für solche, in Unternehmen anfallenden Betriebsdaten wäre ein offener Datenhandel über spezialisierte Marktplätze sowie ein offener Umgang mit entsprechenden Daten und Standards wichtig, damit auch KMU sich in diesem Segment der Datenwirtschaft engagieren könnten.[19]

„Man wird einige Akteure kaum zum offenen Gespräch bekommen. Wo Big Data strukturiert und in Verticals monetarisiert und ausgebeutet wird, braucht man nicht gesprächsbereit zu sein, weil das gut funktioniert. Hier ist mehr Transparenz häufig nicht gewünscht."

Kai Meinke, deltaDAO AG

3.2 Die Motivation von Datenqualitätsprüfungen

Entsprechend der Komplexität der Datenwertschöpfung beschrieben Experten verschiedene Zusammenhänge, in denen sie Prüfungen der Datenqualität vornehmen:

- Mit der Qualitätsprüfung erkennen Datengebende oder Datennehmende Qualitätsmängel in den Datensätzen, um sie dann zielgerichtet zu beheben.
- Mit der Qualitätsprüfung bewerten Datennehmende die Daten vor oder bei der ersten Sichtung hinsichtlich der funktionalen Eignung sowie des potenziellen wirtschaftlichen Nutzens. Dies dient ihnen zur Entscheidung über die Nutzung eines Datensatzes und zur operativen Planung.
- Mit der Qualitätsprüfung legen Datengebende objektivierbare Eigenschaften von Datenangeboten fest. Das dient beim Data Sharing und im Datenhandel dazu, die Informationsasymmetrie zwischen Datengebenden und Datennehmenden abzubauen und so bei Datennehmenden Vertrauen in die Angebote herzustellen.

19 In mehreren SDW-Projekten wird an der Konzeption solcher Marktplätze für Daten, Datenprodukte und Datendienste gearbeitet. Dazu gehören das Projekt DE4L, mit einer Plattform für Bewegungs-, Sensor- und Adressatendaten, die bei Kurier- und Expressdiensten anfallen (www.de4l.io), und das Projekt EVAREST, mit einer Plattform für Datendienste aus und für die Lebensmittelproduktion (www.evarest.de).

- Mit Qualitätsprüfungen schaffen Unternehmen die Grundlage, ihre Daten für die Unternehmensberichterstattung als Wirtschaftsgut zu bewerten oder diese Bewertung sogar für die Unternehmensbilanz zu verwenden.

3.3 Relevante Datenqualitätsdimensionen und ihre Herausforderungen

Bei der Einschätzung, welche Qualitätsdimensionen in der Praxis wichtig seien, bestätigten die Interviews weitgehend die Literatur. Laut den Experten gibt es keine dominante Qualitätsdimension, sondern jeweils „eine Reihe an Dateneigenschaften", die in Abhängigkeit vom konkreten Anwendungsfall zusammenwirken. Die Befragten waren sich einig, dass Qualitätsanforderungen an Daten immer vom wirtschaftlichen und technischen Zusammenhang der Nutzung abgeleitet werden, gleich ob die Qualitätsdimensionen inhärent, systemunterstützt oder pragmatisch sind. Besonderer Handlungsbedarf bestehe daher darin, auf konkrete Anwendungsfälle zugeschnittene Metriken und Messverfahren auszuwählen und zu spezifizieren.

3.3.1 INHÄRENTE DATENQUALITÄT

Fast alle Experten sahen den Kern der Datenqualität durch gängige inhärente Qualitätsdimensionen wie Korrektheit, Aktualität, Vollständigkeit und Konsistenz abgebildet. Unter Qualitätsmetriken verstanden einige Interviewpartner zunächst nur solche etablierten Metriken für syntaktische und semantische Datenqualität, für die es auch Messverfahren gibt, welche regelbasiert und unabhängig vom Anwendungsfall angewendet werden können, wie die Konsistenz des Datenformats. Die Erhebung solcher Metriken kann oft mit Software-Tools durchgeführt werden und die Mängel in Einzeldaten können gegebenenfalls gleich behoben werden, z. B. mithilfe von Korrekturfaktoren.

Insbesondere verarbeitende Intermediäre unter den Experten, welche Lösungen oder Informationen für datennutzende Unternehmen anbieten, stellten jedoch heraus, dass eine aussagekräftige Messung der inhärenten Datenqualität nicht zu jedem Zeitpunkt und nicht durch jeden Akteur erfolgen könne. Beispielsweise könnten Prüfungen oder Hebungen der semantischen Korrektheit oder die Erkennung von Verzerrungen („Bias") oft nur früh in der Datenwertschöpfungskette, d. h. bei oder unmittelbar nach der Erhebung, oder nur durch Domänenexperten mit Expertenwissen und Zugang zu allen Informationen zum Messprozess durchgeführt werden. Deswegen sei eine lückenlose und gründliche Dokumentation aller Verarbeitungsschritte in der Wertschöpfungskette häufig essenziell, um im weiteren Verlauf Glaubwürdigkeit und Verständlichkeit der Daten sicherzustellen.

Auch Dateneigenschaften wie die Streuung (Varianz) bestimmter Datenattribute oder der Umfang der Datensätze (Volumen) wurden als relevante inhärente Qualitätsdimensionen benannt. Es wurde mehrfach bekräftigt, dass es gelegentlich schwierig sei, im Vorfeld genaue Anforderungen an den Datenumfang zu formulieren. Dies läge u. a. daran, dass das ausreichende Volumen der Daten mit anderen Qualitätsdimensionen interagiert, über die keine Angaben gemacht werden: Ein größerer Umfang könne im Einzelfall andere Qualitätsprobleme kompensieren, z. B. das Rauschen in Messdaten oder die zu geringe Auflösung bei Bilddaten. Auf der anderen Seite könne eine hohe Qualität der einzelnen Datenpunkte zu geringeren Anforderungen an den Umfang der Datensätze führen.

3.3.2 SYSTEMUNTERSTÜTZTE DATENQUALITÄT

Neben den üblichen Qualitätsdimensionen Korrektheit, Aktualität, Konsistenz und Vollständigkeit wurden vielfach auch die ebenfalls gängigen systemunterstützten Qualitätsdimensionen Bearbeitbarkeit und Zugänglichkeit als für die Datenwertschöpfung zentral benannt. Insbesondere

Experten, die in Anwendungsbereichen aktiv sind, in denen der Standardisierungsgrad von Daten und die Digitalisierung der Unternehmen noch nicht weit fortgeschritten sind, sahen diese Qualitätsdimension als relevant an. Ein verarbeitender Intermediär berichtete, dass 80 Prozent der Aufwände in die Nutzbarmachung von Rohdaten fließen und dementsprechend systemunterstützte Qualitätsdimensionen für ihn von hoher Bedeutung seien. Als Metriken bzw. Werkzeuge zur Messung dieser Dimensionen erwähnten einige Interviewpartner, die als verarbeitende Intermediäre mit Sensordaten aus der industriellen Produktion arbeiten, selbstentwickelte Checklisten, die datengebende Kundinnen bzw. Kunden im Rahmen der Geschäftsanbahnung ausfüllen und die zur Abschätzung der Aufwände, die zur Datenerschließung notwendig sind, sowie zur Bewertung der Anwendungspotenziale herangezogen werden. In den Checklisten werden ähnliche Inhalte abgefragt wie sie auch mit etablierten Metriken zur Bewertung des digitalen Reifegrads von Unternehmen geprüft werden.

3.3.3 PRAGMATISCHE DATENQUALITÄT

Insgesamt zeigte sich in vielen Interviews mit dem anfänglichen Fokus auf die inhärenten und die systemunterstützten Qualitätsdimensionen zunächst eine eher technisch-mathematische Sicht auf Datenqualität, in der Datenqualität unabhängig von der späteren Nutzung mit Metriken erhoben werden kann. Die Experten waren sich jedoch einig, dass auch die pragmatischen, am Gebrauch orientierten Qualitätsdimensionen für die Datenwirtschaft von hoher Bedeutung sind.

Relevanz und Glaubwürdigkeit. Relevanz und Glaubwürdigkeit wurden meist zuerst aufgeführt, wenn es um pragmatische Datenqualität ging. Es sei allerdings unter Umständen schwierig, die Relevanz eines Datensatzes vorab zu beurteilen, wenn kein ausreichender Zugang zu den Daten bestehe bzw. keine ausreichenden Metadaten angeboten werden. Als eine Metrik für die grundsätzliche Relevanz eines Datensatzes im Zusammenhang von Analyse und KI-Anwendungen wurden entropie basierte Maße des statistischen Informationsgehalts erwähnt, welche zwar nicht die Eignung zum konkreten Zweck widerspiegeln, aber für eine Abschätzung der möglichen generellen Aussagekraft von Daten herangezogen werden können. Ratings, Reviews und Nutzungsstatistiken auf Datenmarktplätzen wurden als mögliche Metriken genannt, mit deren Hilfe die Nutzungsrelevanz direkt abgebildet werden könne.

Vielfach wurde auch die Glaubwürdigkeit von Daten als wichtige Qualitätsdimension benannt. Die Reputation von Datengebenden und Angeboten wurde als wichtiger Faktor der Glaubwürdigkeit und wahrgenommenen Datenqualität beschrieben, der ebenfalls über Ratings und Reviews auf Datenmarktplätzen messbar gemacht werden könne. Der Reputation wurde besondere Bedeutung beigemessen, da sie als Indikator von Datenqualität insgesamt herangezogen werden könne. Weiterhin sahen die Befragten Transparenz über die Herkunft von Daten sowie über mögliche vorgenommene Verarbeitungsschritte, die die Daten im Vorfeld durchlaufen haben, als Faktoren, die eine Einschätzung der Glaubwürdigkeit von Daten erleichtern würden.

„Es ist wirklich relevant: Wie glaubwürdig ist die Datenquelle? Wie glaubwürdig ist der Kurator, der die Daten dann ja auch einordnet und sagt, das ist mit dieser bestimmten wissenschaftlichen Methode gemacht."

Daniel Abbou, KI Bundesverband

Wertschöpfung/Mehrwert. Große Herausforderungen sahen die Experten in der unzureichenden Betrachtung der Qualitätsdimension Wertschöpfung/Mehrwert in der Datenwirtschaft. Einige von ihnen formulierten den Wunsch, dass mehr Metriken zum Einsatz kommen, die den wirtschaftlichen Wert von Daten und Datenqualität im konkreten Nutzungskon-

text quantitativ abschätzen. Solche Metriken werden schon in der Literatur beschrieben (BVDW 2018) und von manchen der Experten auch bereits eingesetzt. Als ein Beispiel wurde die datenbasierte Abschätzung der Kosten genannt, die potenziellen Kunden durch konkrete Qualitätsmängel der inhärenten Datenqualität wie Datendubletten in den internen Datenbeständen entstehen. Diese Abschätzung soll Kaufanreize für datennahe Dienstleistungen generieren. Die Entwicklung solcher Metriken erfordere ein umfangreiches Verständnis der Anwendungsfälle inklusive der operativen Verwendung der Daten im datennutzenden Unternehmen.

Rechtssicherheit. Im Zusammenhang mit der datenwirtschaftlichen Nutzbarkeit von Datensätzen wurde häufig auch die rechtssichere Verwendbarkeit von Daten, etwa unter Berücksichtigung des Urheberrechts oder der Datenschutz-Grundverordnung, als relevante Eigenschaft eines Datensatzes genannt. Es sei eine wichtige Aufgabenstellung, diese in der datenwirtschaftlichen Praxis stärker als Qualitätseigenschaft zu berücksichtigen, insbesondere beim Datenhandel, auch wenn hier keine herkömmlichen Metriken definiert werden können. Ein Plattformanbieter unter den Befragten sieht etwa Dienstleistungsangebote, um Daten oder KI-Modelle hinsichtlich der Beachtung von Datenschutzauflagen zu auditieren, als für den Datenhandel sinnvoll an.

3.4 Der tatsächliche Einsatz von Qualitätsmetriken

Die Praxis der Qualitätsmessung von Daten hängt allerdings nach übereinstimmender Einschätzung der Experten deutlich hinter dem zumindest unter Fachleuten etablierten Wissensstand zurück. Derzeit würden Metriken vor allem im unternehmensinternen Kontext eingesetzt, z. B. bei der Datenerhebung, -sichtung und -aufbereitung oder in Form von KPI in großen Unternehmen. Manche verarbeitende Intermediäre unter den Interviewpartnern nutzen Metriken auch in der Kommunikation mit ihren Kunden, die aus eigenen Daten Wert schöpfen möchten. Anhand der Qualitätsmetriken können sie den Kunden das Datenwertschöpfungspotenzial darlegen. Als explizite Grundlage beim Data Sharing oder im Datenhandel, z. B. über eine Aufnahme in den Vertrag bei einem Datenkauf, würden Metriken aktuell jedoch noch nicht genutzt. Als Verantwortliche wurden sowohl Datengebende als auch Datennehmende benannt.

3.4.1 UNKLARE ODER ZU HOHE ANFORDERUNGEN AN DIE DATENQUALITÄT BEI DATENNUTZENDEN

Die Experten waren sich einig, dass in der Praxis die Anforderungen an die Daten oft nicht klar sind. Als größtes Manko erachteten die Befragten in diesem Zusammenhang, dass datennutzende Unternehmen häufig keine, unklare oder nicht nachvollziehbare Anforderungen formulieren. Grund sei die oft noch unzureichende Konzeption der datenwirtschaftlichen Anwendungsfälle und ihrer wirtschaftlichen Potenziale. In der Praxis würde bei der Anforderungsdefinition der wirtschaftliche Mehrwert eines Anwendungsfalls nicht ausreichend oder im Extremfall gar nicht beachtet. Ergebnis seien häufig willkürliche oder überhöhte Anforderungen an die Datenqualität, die in keinem ausgewogenen Verhältnis zum tatsächlichen Mehrwert der Datennutzung stehen. Das würde es datengebenden Unternehmen erschweren, relevante Dateneigenschaften zu identifizieren und geeignete Metriken für Datenqualität als Metadaten zu Datenangeboten bereitzustellen. Oft würden Qualitätsanforderungen auch nachträglich angepasst oder überhaupt erst zu einem späteren Zeitpunkt formuliert und Mängel müssten in der Datenqualität durch aufwendige Maßnahmen behoben oder kompensiert werden. Eine fehlerbehaftete oder unvollständige Erfassung und Aufbereitung der Daten am Anfang der Datenwertschöpfungskette verursache oft hohe Folgekosten.

In diesem Zusammenhang wurde von einem verarbeitenden Intermediär auch kritisiert, dass endnutzende Unternehmen Datenwertschöpfungsprojekte häufig wie klassische IT-Projekte behandeln, die durch hohe Qualitätsanforderungen an das Ergebnis charakterisiert sind. Daher werde oft ein unangebrachter Perfektionismus an den Tag gelegt. Bei Datenwertschöpfungsprojekten könne jedoch oft schon mit suboptimalen Lösungen ein wirtschaftlicher Mehrwert erzielt werden. Zudem könnten die Daten im Nachhinein verhältnismäßig leicht durch äquivalente, aber bessere Datensätze ersetzt werden, um die Ergebnisqualität zu heben.

3.4.2 SONDERFALL: VARIABLE ANFORDERUNGEN BEI DATENANALYSEN UND KI

Allerdings wurde auch ausdrücklich ein Sonderfall benannt, bei dem die Variabilität von Qualitätsanforderungen unvermeidlich sei: im Rahmen der Nutzung für Datenanalysen oder KI. Anschaulich wurde dies am Beispiel des Trainings von KI-Modellen beschrieben. Es gebe einen ständigen Wechsel („hin und her") zwischen der Anforderungsdefinition an die Daten, die etwa aus dem angestrebten Anwendungsfall und der Wahl des KI-Modells erwachse, und der Anforderungsdefinition an die Verarbeitung, die aus den zur Verfügung stehenden Daten erwachse. Dies ließe sich allerdings zum Teil erst nachträglich feststellen, etwa anhand der Bewertung eines gelernten Modells. Wenn Daten sich als unzulänglich für einen Anwendungsfall herausstellten, würden adaptiv Anpassungen der Datengrundlage vorgenommen, etwa durch die zielgerichtete Erweiterung der Eingangsdaten für die Modelle in Kollaboration mit den Datengebenden, durch nachträgliches „Labeling" der Daten durch Domänenexperten oder spezialisierte Dienstleister, durch das Hinzuziehen oder die Synthese weiterer Datensätze oder durch eine Neuerhebung. Dieses iterative Vorgehen stünde teilweise einer Festigung von Qualitätsstandards entgegen, zumal in diesem Prozess Daten häufig für nicht vorhergesehene Zwecke nachgenutzt werden (Repurposed Data). Auch hier bestätigen die Experten die Fachliteratur.

3.4.3 KEIN BEWUSSTSEIN FÜR DATENQUALITÄT BEI DATENGEBENDEN

Den oft unklaren und dynamischen Anforderungen der Datennutzenden gegenüber stehen Datenangebote, die von den Befragten als häufig nicht transparent und von geringer Qualität beschrieben wurden. Datengebende Unternehmen, die die Datenwertschöpfung im Nebengeschäft betreiben, würden der Datenqualität keine ausreichende Beachtung schenken. Es würden zudem häufig kein ausreichender Datenzugang und keine ausreichenden Metadaten zur Entstehungsgeschichte von Datensätzen bereitgestellt, um die Datenqualität einschätzen oder prüfen zu können.

„Das größte Hemmnis ist, dass die Datenqualität oft vergessen wird. Unternehmen kommen dann auf uns [als Dienstleister] zu, wenn sie auf das Problem der Datenqualität gestoßen sind. Das Verständnis ist noch nicht da."

Dan Follwarczny, zum Zeitpunkt des Interviews Uniserv GmbH, jetzt ecovium GmbH

Sinnvolle und bereits etablierte Metriken bzw. Werkzeuge für die Messung und Hebung von Datenqualität würden von den Datengebenden der Datenwirtschaft häufig nicht eingesetzt. Sowohl unternehmensübergreifende Datenplattformen als auch Datenkataloge für die unternehmensinterne Datenhaltung würden zum Teil bereits den Einsatz von Metriken und Werkzeugen unterstützen, diese stießen jedoch bei den Datengebenden auf wenig Nachfrage.

3.4.4 KEINE DURCHGEHENDE QUALITÄTSBEWERTUNG

Selbst wenn es bei datengebenden und datennehmenden Unternehmen schon ein Bewusstsein für Datenqualität gebe, fehle in Datenwertschöpfungsketten mit mehreren beteiligten Akteuren trotzdem vielfach noch eine durchgängige Qualitätsbewertung, die einen Abgleich der Beschaf-

fenheit von Daten mit den anwendungsspezifischen Nutzungsanforderungen ermögliche. Eine derartige durchgängige Bewertung scheitere oft daran, dass sich in vielen Bereichen noch keine Qualitätsstandards etabliert haben und keines der an einer Datenwertschöpfungskette beteiligten Unternehmen alle notwendigen Informationen habe, um einen solchen Abgleich durchzuführen. Unternehmen, welche die gesamte Datenwertschöpfungskette intern implementieren, wie etwa große Internetplattformen, seien im Vorteil, da sie dieses Problem nicht haben.

3.4.5 OFFENE DISKUSSION: STANDARDMETRIKEN FÜR DATENQUALITÄT?

Die Experten waren sich uneins, inwieweit eine Standardisierung von Qualitätsmetriken oder eine Automatisierung des Umgangs mit Datenqualität durch den Einsatz von Tools zur Messung von Metriken möglich oder sinnvoll seien. Eine Gruppe von verarbeitenden Intermediären unter den Interviewpartnern, die primär als Dienstleistende und Anbietende von Individuallösungen aktiv sind, stellte die Variabilität der Anforderungen und die Einzigartigkeit von Anwendungen in den Vordergrund und befürwortete zum Teil händische Ansätze, bei denen Data Scientists ad hoc Qualitätsanforderungen definieren und prüfen, z. B. eine gewisse Varianz in einem bestimmten Datenattribut. Sie sahen zudem in händischem Vorgehen die einzige Möglichkeit, Verzerrungen in Daten zu erkennen und zu vermeiden. Diese Gruppe sah es als wünschenswert an, dass Akteure der Datenwirtschaft flexibel mit variablen Anforderungen der Nachfrageseite umgehen und Individualisierungen ermöglichen.

Andere verarbeitende Intermediäre unter den Befragten sahen solch händisches Vorgehen als ineffizient an. Sie drückten die Einschätzung aus, dass menschliche Expertise bei der Datenqualitätsprüfung in der Community eine überhöhte Wertschätzung erhalte. Es stünden Metriken und Tools zur Verfügung, die in der Praxis nicht häufig genug zum Einsatz kämen. Um Transparenz in der Datenwirtschaft zu schaffen, sollten Metriksammlungen in den verschiedenen Anwendungsbereichen spezifiziert, offengelegt und etabliert werden. Neben den gängigen Metriken für die Qualitätsdimensionen Korrektheit, Konsistenz, Aktualität und Vollständigkeit wurden in diesem Zusammenhang vereinzelt auch auf KI-Trainingsdaten ausgerichtete Metriken benannt. Auch der Einsatz von KI zur Hebung der Datenqualität, z. B. zur nachträglichen Annotierung (Labeling) oder Bereinigung von Daten, sei sinnvoll und Befürchtungen einer möglichen Verfälschung oder Verzerrung von Daten in diesem Prozess seien oft unbegründet.

„[Datenqualitätsmetriken] sind entscheidend für die Datenökonomie von morgen. In dem Augenblick, wo ich die erste Metrik habe, veröffentliche und operationalisierbar mache, kriege ich Anreize rein, dass andere sich danach ausrichten, dass ich danach gemessen werden kann. So nimmt das Ganze Fahrt auf."

Kai Meinke, deltaDAO AG

Es zeichnete sich allerdings das Bild ab, dass diese beiden Sichten nicht unbedingt im Gegensatz zueinander stehen, sondern durch bestimmte Faktoren in den Anwendungsfällen beeinflusst sind. Die Experten gaben zu bedenken, dass beim Umgang mit großen Datenvolumen automatisierte Verfahren notwendig seien, um eine Skalierung der Aufwände zu erreichen. Bei Anwendungsfällen mit hoher Prozessreife und Anwendungsdomänen mit etablierten Branchenstandards für Daten führten die Befragten zudem häufiger strukturierte Prozesse und sinnvolle Metriken an, die genutzt werden sollten, um Datenbestände vor der Bearbeitung im Wertschöpfungsprozess zu bewerten. Inwieweit eine Standardisierung von Datenqualitätsmetriken und darauf aufbauend eine Automatisierung der Datenqualitätsmessung möglich und sinnvoll sind, hängt somit vom Einzelfall wie auch vom Standardisierungsgrad der Daten und Anwendungsfälle sowie von der Größe der zu bewertenden Datensätze ab.

3.4.6 SONDERFALL: QUALITÄTSPRÜFUNG IM DATENHANDEL

Nach Einschätzung der Experten kommt der Datenqualität im Datenhandel eine zentrale Rolle zu. Gängige Praxis der Datenangebotspräsentation seien beispielhafte Auszüge aus den Daten und die Bereitstellung von Metadaten. Die Quantität und Aussagekraft der Metadaten habe dabei eine zentrale Bedeutung, um die Informationsasymmetrie zwischen Anbietenden und Nachfragenden zu mindern. Neben transparenten Angaben zur Beschaffenheit, z. B. zum Datenmodell und zum Datenvolumen sowie zur Herkunft und Verarbeitungsgeschichte der Daten, seien dabei laut einigen Interviewpartnern insbesondere auch Metriken für die Datenqualität von entscheidender Bedeutung.

„Wenn man Richtung Datenmarktplatz denkt, braucht man früher oder später so etwas wie eine dritte Meinung zur Werthaltigkeit eines Datensatzes. [...] Es wird also im Wesentlichen darauf hinauslaufen, dass man einen quantitativen Fingerprint der Daten erstellt, aus semantischen Metriken, eventuell sogar aus den wirtschaftlichen."

Georg Wittenburg, Inspirient GmbH

Andere Experten, die ihre Angebote primär an KMU der industriellen Produktion vertreiben, äußerten allerdings die Einschätzung, dass beim Datenhandel wichtige Qualitätsprüfungen vor einer Kaufentscheidung nur durch einen weitreichenden Datenzugriff oder die Kenntnis des Erhebungsprozesses möglich seien. Die Bereitstellung von Metadaten, Datenauszügen oder etwa durch Anonymisierung stark transformierte Daten seien nicht ausreichend, um die Datenqualität sicher einschätzen zu können. Der Zugang zu den Volldaten werde aber von den Datengebenden verständlicherweise erst nach dem Kauf geöffnet. Der Zugriff zu Rohdaten könne oft überhaupt nicht gewährt werden, weil sonst etwa Geschäftsgeheimnisse oder Personendaten offengelegt würden. Ein möglicher Lösungsansatz läge in der Nutzung von On-premise-Verfahren, d. h. von Diensten, die vor Ort beim Datengebenden ausgeführt werden, beispielsweise um für einen bestimmten Anwendungsfall wichtige Qualitätsmetriken zu berechnen, und die nur Informationen herausgeben, die keine Rückschlüsse auf die Datengrundlage zulassen. Aktuelle, auf die Datensouveränität ausgerichtete Initiativen, wie das Industriekonsortium IDSA und das europäische Projekt GAIA-X, sehen die Bereitstellung solcher Dienste als wichtigen Teil ihrer technischen Systemarchitekturen an.

„[Eine wichtige Rahmenbedingung ist] die Anonymisierung […] z. B. wo [in der Industrieanlage] der Kompressor steht. Das wäre natürlich besser zu wissen, aber da würde man aus den Daten Rückschlüsse auf Geschäftsdaten ziehen. […] Ich befürchte, dass viele Betreiber dieses Risiko nicht eingehen werden."

Andreas Herzog, Fraunhofer-Institut für Fabrikbetrieb und -automatisierung IFF

Generell galt für die Befragten, dass es einfacher sei, aussagekräftige Qualitätsmetriken für Informationsangebote zu spezifizieren, die auf einen Anwendungsfall zugeschnitten sind, als für Datenangebote, die noch nicht für einen konkreten Anwendungsfall aufbereitet sind. Im Handel mit Informationsangeboten könne davon ausgegangen werden, dass sich Qualitätsstandards parallel zu den Anwendungsfällen in Branchen und Domänen etablieren.

Die Tatsache, dass in der Datenwirtschaft häufig Daten neue Nutzungszwecke zugeführt werden (Repurposed Data) und hierfür Datentransaktionen, z. B. im Rahmen des Trainings von KI-Modellen, häufig nur einmalig stattfinden, führe jedoch zu neuen Herausforderungen im Datenhandel, da die Datengebenden das Datenangebot inklusive der aussagekräftigen Präsentation der Datenqualität in teilweiser Unkenntnis der Anforderungen entwickeln müssen. Notwendig sei es, durch eine vergleichende Analyse praxisnaher Qualitätsmetriken über einzelne Anwendungsfälle hinweg, Metriken für den Datenhandel mit einer anwendungsübergreifenden Relevanz zu identifizieren. Weiterhin sei es wichtig, anpassbare Angebots- und Preismodelle zu entwickeln, um auf Datennehmende mit unterschiedlichen Qualitätsanforderungen und unterschiedlichem wirtschaftlichen Mehrwert aus der Datennutzung eingehen zu können.

04

4 HANDLUNGSEMPFEHLUNGEN AN UNTERNEHMEN UND WISSENSCHAFT

Die Experten, die für diese Studie befragt wurden, sehen durchgehend in der Messung und der Transparenz von Datenqualität eine Grundvoraussetzung für eine breite Durchsetzung der Datenwirtschaft. Eine über Metriken nachgewiesene Datenqualität schafft Transparenz zu Datenangeboten und damit Entscheidungssicherheit in allen Teilschritten der Wertschöpfungskette. Im Datenhandel ist das von besonderer Bedeutung, da vor der Transaktion eine grundlegende Informationssymmetrie über die Beschaffenheit der Daten zwischen Datenanbietenden und Datenkaufenden besteht. Die Anbietenden kennen das Datenangebot im Detail, die Kaufenden aber nicht. Erst mit für alle Seiten nachvollziehbaren Aussagen zu den zentralen Eigenschaften der Datenprodukte können sich in der Datenwirtschaft offene Märkte bilden, in denen über den Preis eine optimale Ressourcenallokation erreicht wird. Datenqualitätsmetriken sind ein wichtiger Teil solch einer nachvollziehbaren Beschreibung von Datenangeboten, können die Informationsasymmetrie zwischen Datenanbietenden und Datenkaufenden verringern und so die Markttransparenz steigern.

Noch lässt der praktische Umgang mit Datenqualität aber Lücken offen. Aus der Literaturrecherche, den Interviews mit den Experten und den Validierungs-Workshops ergeben sich eine Reihe von Handlungsempfehlungen zur Integration von Datenqualitätskonzepten und Qualitätsmetriken in der Datenwirtschaft. Die Empfehlungen richten sich teilweise an die Wissenschaft und Verwaltung, zum größeren Teil aber an Unternehmen, Branchenverbände und Fachverbände für Datenwirtschaft und Datenqualität.

4.1 Anwendung etablierter Konzepte

Viele klassische Datenqualitätsdimensionen wie Korrektheit, Konsistenz, Vollständigkeit und Aktualität sind theoretisch gut fundiert. In der Praxis haben sich allerdings auch für diese Dimensionen oft noch keine Qualitätsstandards und Metriken etabliert. Daraus leiten sich die folgenden Handlungsbedarfe ab:

DURCHGÄNGIGE BETRACHTUNG DER DATENQUALITÄT ENTLANG DER WERTSCHÖPFUNGSKETTE PRAKTIZIEREN

Oft handeln die verschiedenen Akteure in der Wertschöpfungskette nur jeweils aus einer eingeschränkten Perspektive, wenn sie Transformationsschritte an den Daten durchführen und deren Qualität einschätzen und zu heben versuchen. Das verhindert eine ganzheitliche Betrachtung der Datenqualität. Hieraus ergeben sich unterschiedliche Handlungsbedarfe für die Beteiligten in der Wertschöpfungskette:

- Nur datenerhebende Akteure kennen die Daten in ihrer Gänze und ihre Entstehungsgeschichte. Dies ist die Grundlage für die Messung einiger wichtiger Qualitätsdimensionen. Unternehmen, die wegen des Schutzes von Betriebsgeheimnissen oder Personendaten nur selektiv Daten preisgeben möchten, sollten aussagekräftige und nachvollziehbare Metriken mit den Daten bereitstellen bzw. vertrauenswürdigen Partnern, welche in der Datenwertschöpfungskette die Datenaufbereitung durchführen, den notwendigen Zugang zu den Daten bzw. dem Erhebungsprozess gewähren, damit diese die Datenqualität messen, Metriken bereitstellen und die Datenqualität auf ein anforderungsgerechtes Maß heben können. Behörden und Forschende, die Open Data bereitstellen sowie Unternehmen die Rohdaten anbieten, sollten neben den Daten auch vollständige und nachvollziehbare Angaben zur Herkunft und Herstellungsgeschichte der Daten bereitstellen, um Datennutzenden eine Qualitätsbewertung zu ermöglichen.

- Die Datennutzung ist der Ursprung der Anforderungen an Datenqualität. Datennutzende sollten angemessene Anforderungen an die Daten und Metriken formulieren, die sich am wirtschaftlichen Nutzen des Anwendungsfalls ausrichten, bzw. Entwicklungspartnern wie Dienstleistern, die in der Datenwertschöpfungskette die Informationsgewinnung durchführen, ausreichend Einblick in die Geschäftsprozesse gewähren. So lassen sich überzogene oder falsche Anforderungen an die Datenqualität vermeiden.
- Verarbeitende Intermediäre, d. h. technische Dienstleistende, Datenbroker oder Anbietende datenbasierter Dienste und Produkte, bringen die Daten der Datenerhebenden und die Anforderung der Datennutzenden zusammen. Wenn neue Datenwertschöpfungsketten entstehen, sind sie oftmals die einzigen Akteure mit genügend Fachexpertise, um Metriken festzulegen. Daher sollten sie zur Verfügung stehende Metriken anwenden und die kollaborative Spezifikation von Anforderungen und Metriken sowie deren Nutzung entlang der Wertschöpfungskette gemeinsam mit Datenerhebenden und Datennutzenden vorantreiben.
- Betreiber von Datenplattformen sollten die kollaborative Verwendung und Spezifizierung von Metriken sowie den Nachweis von Qualitätseigenschaften der Daten in ihren Angeboten berücksichtigen und unterstützen.

4.1.1 QUALITÄTSMETRIKEN STRUKTURIERT DEFINIEREN UND OFFENLEGEN

Qualitätsmetriken ermöglichen objektivierbare Aussagen zu Dateneigenschaften. Sie schaffen die Grundlage für informierte Entscheidungen in der Datenwertschöpfungskette und sind damit die Grundlage für transparente Märkte. Dafür müssen sie aber vom jeweiligen Datengebenden oder verarbeitenden Intermediär strukturiert erarbeitet und für die anderen Akteure offengelegt werden. Das umfasst:

- die Festlegung und Priorisierung der Anforderungen an die Metrik,
- die Definition der Metrik und
- die Festlegung der Evaluationsmethodik.

Erst so werden die Metriken nachvollziehbar und vergleichbar, was in der Praxis bisher oft nicht der Fall ist. Branchenverbände, die sich mit Datenwirtschaft bzw. Datenmanagement befassen, sollten daher die Offenlegung von Metriken für einzelne branchenspezifische Anwendungsfälle forcieren, Kataloge erstellen und somit den Grundstein für eine Etablierung von offenen Standardmetriken legen.

4.1.2 BRANCHENÜBERGREIFENDE SYNERGIEPOTENZIALE AUSSCHÖPFEN

Auch wenn in der Datenwirtschaft Datenqualität anwendungsspezifisch zu definieren ist, lassen sich viele Metriken oder Prozesse im Umgang mit Datenqualität doch gut von einer Branche auf die anderen übertragen. Um dieses Synergiepotenzial zu nutzen, sollten Branchenverbände, Fachverbände für Digitalwirtschaft bzw. Datenmanagement und Wissenschaft eine Plattform für den Austausch über Datenqualität aufbauen. Diese Plattform sollte Standards für die Spezifikation von Datenqualitätsmetriken erarbeiten, was einen offenen Umgang mit Metriken begünstigen würde. Daneben sollte die Identifikation von branchenübergreifenden Metriken und Metriksammlungen ein Ziel des Austauschs sein. Damit eröffnet sich die Chance, das Bewusstsein für Datenqualität auch in Anwendungsbereichen zu schärfen, für die es noch keine etablierten Metriken gibt.

4.1.3 DIE BESCHAFFENHEIT VON DATEN VERTRAGLICH REGELN

Qualitätsmängel in Daten können zu Nachteilen bei datennehmenden Unternehmen oder deren Kunden führen. Da in vielen Anwendungsbereichen noch keine Standards für Datenqualität etabliert sind, gibt es keine klaren objektiven Anforderungen, die zur Feststellung von Haftungsansprüchen herangezogen werden können. Um Klarheit und Rechtssicherheit bezüglich der Haftbarkeit für Datenqualität zu erhalten, sollten die Vertragsparteien der Datenwirtschaft daher die Beschaffenheit von Daten möglichst konkret festlegen. Hierzu gehört die Bestimmung eines Verwendungszwecks sowie die Definition von konkreten Qualitätsanforderungen. In Bezug auf die Feststellung eines Mangels sollten die Vertragsparteien zudem Methoden und Metriken zur Bestimmung der geschuldeten Datenqualität verbindlich regeln. Damit kann im Falle des Unterschreitens der Datenqualität ein Mangel leichter festgestellt werden. Branchen- und Fachverbände sollten Standardverträge für die Datenüberlassung definieren, um die Vielfalt an möglichen rechtlichen Regelungen zu reduzieren und damit für Akteure in der Wertschöpfungskette die notwendigen rechtlichen Prüfungen auf ein angemessenes Maß zu beschränken. Plattformbetreibende sollten solche Standardverträge mit konfigurierbaren Vereinbarungen zur Datenqualität auf der Plattform bereitstellen.

4.2 Konzeptionelle Weiterentwicklungen

Mit der zunehmenden Bedeutung unternehmensübergreifender, offener Netzwerke in der Datenwirtschaft, insbesondere des Datenhandels, ergeben sich neue Anforderungen an die Datenqualität, die in Zukunft stärker berücksichtigt werden müssen.

4.2.1 RAHMENKONZEPT FÜR DATENQUALITÄT IM DATENHANDEL ENTWICKELN

Existierende Rahmenkonzepte für Datenqualität sind in der Regel auf eine interne Verwendung von Daten ausgelegt. Branchenverbände und Fachverbände für Datenwirtschaft und Datenqualität sollten daher ein Rahmenmodell etablieren, das besonders die für das Data Sharing und den Datenhandel wichtigen Qualitätsdimensionen identifiziert, wie z. B. die Glaubwürdigkeit und die Relevanz von Daten sowie die Nutzbarkeit der Daten und die Qualität ihrer Präsentation. Um anschlussfähig an die aktuelle Fachdebatte zu sein, könnte der Ausgangspunkt das Dictionary of Data Quality Dimensions (3DQ) der Arbeitsgruppe Datenqualität in der niederländischen Sektion der Data Management Association International (DAMA NL) sein, welches die Qualitätsdimensionen in bereits existierenden Rahmenmodellen harmonisiert hat.

4.2.2 METRIKEN FÜR PRAGMATISCHE DATENQUALITÄT ENTWICKELN

Pragmatische Datenqualitätsdimensionen wie die Glaubwürdigkeit, die Relevanz oder der wirtschaftliche Mehrwert von Daten sind für das Data Sharing und den Datenhandel von besonderem Interesse. Allerdings gibt es hierfür noch keine etablierten Metriken.

Unterdimensionen, die die Glaubwürdigkeit von Daten beeinflussen und für Metriken herangezogen werden könnten, reichen von der Transparenz der Datenherkunft und Entstehungsgeschichte über die Reputation von Datengebenden und Datenangeboten zum Standardisierungsgrad und der Auditierbarkeit. Unternehmen und Wissenschaft sollten Metriken für Glaubwürdigkeit und Relevanz entwickeln und in der datenwirtschaftlichen Anwendung validieren. Für den Datenhandel kann hier von anderen Märkten gelernt werden. Beispielsweise haben sich Reputationssysteme in anderen Bereichen des E-Commerce bewährt, sofern die Sicherheit vor Manipulationen durch Marktteilnehmende gewährleistet ist.

Datennutzende Unternehmen sollten dazu übergehen, für ihre jeweiligen Anwendungsfälle immer auch den wirtschaftlichen Mehrwert abzuschätzen, den ein Datenangebot ihnen voraussichtlich liefert. Nach Möglichkeit sollten sie dafür unter Berücksichtigung der notwendigen Qualitätsmerkmale auch Metriken entwickeln und gegenüber den anderen Akteuren in der Datenwirtschaft transparent machen. Intermediäre, die häufig über einen breiteren Erfahrungsschatz verfügen, sollten ihre Kunden dabei aktiv unterstützen. Bei Bedarf sollten sie in Zusammenarbeit mit ihnen die Entwicklung der Metriken übernehmen.

4.2.3 RECHTSSICHERHEIT DER DATENNUTZUNG ZUSICHERN

Auch wenn sich die Rechtssicherheit der Datennutzung nicht über Metriken abbilden lässt, so stellt sie doch eine wichtige pragmatische Qualitätsdimension dar, die den Datennehmenden zugesichert werden muss. Datengebende und Plattformbetreibende sollten ausdrücklich und verbindlich spezifizieren, welche rechtlichen Vorgaben für die Nutzung der von ihnen angebotenen Daten gelten. Plattformbetreibende und andere Intermediäre sollten entsprechende Auditierungsverfahren entwickeln.

4.2.4 REPURPOSED DATA STÄRKER BERÜCKSICHTIGEN

Werden Datenangebote für nicht vorhergesehene Zwecke verwendet (Repurposed Data), ergeben sich häufig neue, von den Datengebenden nicht betrachtete Nutzungsanforderungen. Um die nicht vorhergesehene Nutzung von Daten zu erleichtern, sollten datengebende Unternehmen für die Datenangebote Standardmetriken bereitstellen, die anwendungsfallübergreifend Relevanz haben. Hierfür könnte auch auf Metrikenkataloge zurückgegriffen werden. Auch die Möglichkeit einer nachträglichen Bewertung von Daten, z. B. anhand der Performance eines gelernten Modells, sollte im Datenhandel berücksichtigt werden, etwa durch Definition eines entsprechend parametrisierten Preises im Datenhandelsvertrag.

Um Angebot und Nachfrage auch jenseits etablierter Anwendungsfälle zusammenzubringen, sollten Betreibende von Plattformen für das Data Sharing und den Datenhandel mit ihren Angeboten einen Rahmen schaffen, in dem Datengebende und Datennehmende hinsichtlich der Datenqualität effizient kommunizieren und flexibel agieren können. Datennehmende sollten spezifische Anforderungen an die Datenqualität übermitteln können. Datengebende sollten Angebote und Preise anpassen bzw. konfigurierbar gestalten können.

4.2.5 DATENQUALITÄTSBEWERTUNG ON-PREMISE ERMÖGLICHEN

Viele datenerhebende Unternehmen wollen einen umfänglichen Zugriff auf interne Daten von außen nicht zulassen, da dies mit einem möglichen Abfluss von Geschäftsgeheimnissen verbunden ist. Ein umfänglicher Zugriff ist allerdings in vielen Fällen für die Qualitätsbewertung auch gar nicht nötig. Die Anbietenden von Diensten zur Datenbewertung sollten Ansätze verfolgen, bei denen Auswertungen zur Einschätzung oder Bewertung lokal beim Datengebenden (on-premise) auf dem vollen Datenbestand, gegebenenfalls sogar auf den Rohdaten, ausgeführt werden können, ohne dass dieser die Daten selbst preisgeben muss. Anbietende von Plattformen für das Data Sharing und den Datenhandel sollten die Anwendung solcher Lösungen unterstützen und dabei eine enge Abstimmung mit den auf die Datensouveränität ausgerichteten Rahmenentwicklungen im Industriekonsortium International Data Spaces Association und im europäischen Projekt GAIA-X verfolgen.

4.2.6 DIE WAHRNEHMUNGSEBENEN PRÄSENTATION, NUTZBARKEIT UND ZUGANG BERÜCKSICHTIGEN

Insbesondere beim unternehmensübergreifenden Data Sharing und Datenhandel kommt der Wahrnehmung der angebotenen Daten durch die externen Partner eine wichtige Rolle zu. Die Datengebenden und die Plattformbetreibenden sollten die folgenden hieraus erwachsenden Anforderungen an die Datenqualität berücksichtigen:

- Die wahrgenommene Datenqualität wird auch durch die Präsentation der Daten und ihre Nutzbarkeit bestimmt, weswegen die Gestaltung von Datenangebot und Schnittstellen professionellen Usability-Maßstäben genügen muss.
- Die wahrgenommene Datenqualität wird auch durch die Möglichkeit des Zugriffs bestimmt, weswegen eine transparente Zugriffskontrolle implementiert werden sollte.
- Die Datenbestände sollten auch bei nicht fortwährender interner Nutzung für die Dauer des Datenangebots aktuell gehalten und gepflegt werden. Die diesbezüglich durchgeführten Maßnahmen und Schritte müssen transparent kommuniziert werden.

4.3 Umfeld

Die letzte Gruppe von Handlungsempfehlungen betrifft das weitere Umfeld der Datenwirtschaft, in der wichtige Grundlagen für die Hebung der Datenqualität geschaffen werden können.

4.3.1 ZERTIFIZIERUNGSSTRUKTUREN AUFBAUEN

Insbesondere bei wichtigen oder sicherheitskritischen Verwendungen der Datenprodukte sollte die Aufgabe der Qualitätsmessung von neutralen und glaubwürdigen Zertifizierungsdienstleistern übernommen werden, um so Vertrauen aufzubauen. Diese Rolle müsste in der Datenwirtschaft neu geschaffen werden.

4.3.2 DATENFORMATE UND SCHNITTSTELLEN FÜR DIE DATENWIRTSCHAFT STANDARDISIEREN

Der Standardisierungsgrad von Datenformaten und Datenschnittstellen wirkt sich einerseits über die Qualitätsdimension der Konsistenz und Bearbeitbarkeit direkt auf die Datenqualität aus. Andererseits gibt es auch indirekte Effekte: Die Konformität mit Standards erleichtert die Messung von Datenqualität und die Auditierung von Daten und trägt somit zur Glaubwürdigkeit von Daten und der datenwirtschaftlichen Nutzbarkeit bei. Wissenschaft, Fachverbände sowie Standardisierungs- und Normungsgremien sollten sich daher um die Etablierung von Datenstandards und Schnittstellen für die Datenwirtschaft bemühen und dabei enge Abstimmung suchen mit entsprechenden Bemühungen in beispielsweise den Domänen-Arbeitsgruppen von GAIA-X.

4.3.3 DATENKULTUR IN UNTERNEHMEN SCHAFFEN

Ein Großteil der bisherigen Handlungsempfehlungen richtet sich auch an Unternehmen, die im Nebengeschäft in der Datenwirtschaft tätig sind bzw. tätig werden wollen. Um die Empfehlungen aufzugreifen, zu reflektieren und in die Praxis umzusetzen, sollten diese Unternehmen bei sich eine entsprechende Datenkultur schaffen. Das beinhaltet die Wahrnehmung von Daten nicht nur als interne IT-Ressource, sondern auch als Wirtschaftsgut oder Handelsgut, das mit Dritten ausgetauscht werden kann. Für die Wertschöpfung sollte ein unternehmensweites Datenmanagement umgesetzt werden, das interne Datensilos auflöst und so den Wert und die Nutzbarkeit von Daten hebt. Die Datenstrategie sollte zunächst einen konkreten datenwirtschaftlichen Anwendungsfall

mit klarem wirtschaftlichem Nutzen in den Fokus nehmen. Diesen zu identifizieren, erfordert die Zusammenarbeit über Unternehmensbereiche hinweg. Bei Führungskräften und Mitarbeitenden sollte daher eine angemessene Datenkompetenz (Data Literacy) aufgebaut werden. Das schließt je nach Aufgabengebiet ein grundlegendes Verständnis von Daten als Wirtschafts- und Handelsgut ein sowie Kenntnisse zu Datenqualität, Qualitätsmetriken, Methoden zur Hebung der Datenqualität und den entsprechenden Software-Werkzeugen. Branchenverbände sollten ihre Mitglieder für die Schaffung einer solchen Datenkultur sensibilisieren.

05

5 LITERATURVERZEICHNIS

Aamodt, Agnar; Nygård, Mads (1995): Different roles and mutual dependencies of data, information, and knowledge – an AI perspective on their integration. In: Data and Knowledge Engineering (16), S. 191–222.

Amicis, Fabrizio de; Batini, Carlo (2004): A methodology for data quality assessment on financial data. In: Studies in communication sciences: journal of the Swiss Association of Communication and Media Research 4 (2), S. 115. DOI: 10.5169/seals-790977.

Auer-Reinsdorff, Astrid; Conrad, Isabell (Hg.) (2019): Handbuch IT- und Datenschutzrecht. Deutscher Anwaltverein. 3. Auflage. München: C.H. Beck (Beck-Online Bücher). Online verfügbar unter: https://beck-online.beck.de/?vpath=bibdata/komm/AuerReinsdorffConradHdbITDSR_3/cont/AuerReinsdorffConradHdbITDSR.htm

Band, Beatrice; Bauer, Margret; Blumenthal, Rolf; Cloppenburg, Frederik; Dühnen, Christopher; Froese, Thomas et al. (2022): Future Data Assets. (im Erscheinen). Hg. v. VDI/VDE/GMA, Richtlinienausschuss 7.24 „Big Data".

Black, Andrew; van Nederpelt, Peter (2020): Dictionary of dimensions of data quality (3DQ). Dictionary of 60 Standardized Definitions. DAMA NL. Online verfügbar unter: http://www.dama-nl.org/wp-content/uploads/2020/11/3DQ-Dictionary-of-Dimensions-of-Data-Quality-version-1.2-d.d.-14-Nov-2020.pdf

BVDW (Hg.) (2018): Data Economy: Datenwertschöpfung und Qualität von Daten. Online verfügbar unter: https://www.bvdw.org/fileadmin/bvdw/upload/publikationen/data_economy/BVDW_Datenwertschoepfung_2018.pdf, zuletzt geprüft am 15.02.2022.

Calero, Coral; Caro, Angélica; Piattini, Mario (2008): An Applicable Data Quality Model for Web Portal Data Consumers. In: World Wide Web 11 (4), S. 465–484. DOI: 10.1007/s11280-008-0048-y.

Cichy, Corinna; Rass, Stefan (2019): An Overview of Data Quality Frameworks. In: IEEE Access 7, S. 24634–24648. DOI: 10.1109/ACCESS.2019.2899751.

Daniel, Christel; Serre, Patricia; Orlova, Nina; Bréant, Stéphane; Paris, Nicolas; Griffon, Nicolas (2019): Initializing a hospital-wide data quality program. The AP-HP experience. In: Computer methods and programs in biomedicine 181, S. 104804. DOI: 10.1016/j.cmpb.2018.10.016.

Demary, Vera; Fritsch, Manuel; Goecke, Henry; Krotovaz, Alevtina; Lichtblau, Karl; Schmitz, Edgar et al. (2019): Readiness Data Economy - Bereitschaft der deutschen Unternehmen für die Teilhabe an der Datenwirtschaft. Eine Veröffentlichung im Rahmen des BMWi-Verbundprojektes: DEMAND - DATA ECONOMICS AND MANAGEMENT OF DATA DRIVEN BUSINESS. IDW. Köln. Online verfügbar unter: https://www.demand-projekt.de/paper/Gutachten_Readiness_Data_Economy.pdf, zuletzt geprüft am 23.02.2022.

Dewenter, Ralf; Lüth, Hendrik (2019): Datenhandel und Plattformen. ABIDA – Assessing Big Data. Online verfügbar unter: https://www.abida.de/sites/default/files/ABIDA_Gutachten_Datenplatformen_und_Datenhandel.pdf, zuletzt geprüft am 23.02.2022.

Europäische Kommission (2017): Aufbau einer europäischen Datenwirtschaft. Mitteilung der Kommission an das Europäische Parlament, den Rat, den europäischen Wirtschafts- und Sozialausschuss und den Ausschuss der Regionen. Brüssel.

Europäische Kommission (2021): Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts com/2021/206 FINAL. Artifical Intelligence Act, vom Document 52021PC0206. Online verfügbar unter: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206, zuletzt geprüft am 10.03.2022.

Fürber, Christian (2016): Data Quality Management with Semantic Technologies. Wiesbaden: Springer Fachmedien.

Geiger, Walter; Kotte, Willi (2008): Die Fachbegriffe Qualität und Fähigkeit. In: Handbuch Qualität. Wiesbaden: Vieweg, S. 67–82.

Heinrich, Bernd; Klier, Mathias; Hristova, Diana; Schiller, Alexander Paul Rudolf; Szubartowicz, Michael (2018a): Requirements for Data Quality Metrics. DOI: 10.5283/epub.36889.

Heinrich, Bernd; Klier, Mathias; Hristova, Diana; Schiller, Alexander Paul Rudolf; Szubartowicz, Michael (2018b): Requirements for Data Quality Metrics. In: Journal of Data and Information Quality (JDIQ) 9 (2), ArtNo.12. Online verfügbar unter: https://epub.uni-regensburg.de/36889/, zuletzt geprüft am 10.03.2022.

Helfert, Markus; Herrmann, Clemens (2002): Proactive data quality management for data warehouse systems. In: DMDW.

ISO 8000-8:2015, 2015: ISO 8000-8:2015 Data quality - Part 8: Information and data quality: Concepts and measuring. Online verfügbar unter: https://www.iso.org/standard/60805.html

ISO/IEC 25012:2008, 2008: ISO/IEC 25012:2008 Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model.

Kahn, Beverly; Strong, Diane M.; Wang, Richard Y. (1997): A Model for Delivering Quality Information as Product and Service. In: Diane M. Strong und Beverly Kahn (Hg.): Second Conference on Information Quality (IQ 1997): MIT, S. 80–94. Online verfügbar unter: http://mitiq.mit.edu/ICIQ/Documents/IQ%20Conference%201997/Papers/AModel4DeliverQualityInfoasProductnService.pdf, zuletzt geprüft am 10.03.2022.

Kahn, Beverly; Strong, Diane M.; Wang, Richard Y. (2002): Information quality benchmarks. In: Commun. ACM 45 (4), S. 184–192. DOI: 10.1145/505248.506007.

Katerattanakul, Pairin; Siau, Keng (1999): Measuring Information Quality of Web Sites: Development of an Instrument. In: Proceedings of the 20th International Conference on Information Systems. USA: Association for Information Systems (ICIS '99), S. 279–285. Online verfügbar unter: https://aisel.aisnet.org/icis1999/25/, zuletzt geprüft am 10.03.2022.

Koutroumpis, Pantelis; Leiponen, Aija; Thomas, Llewellyn D. W. (2017): The (Unfulfilled) Potential of Data Marketplaces (ETLA working papers, 53). Online verfügbar unter: https://www.econstor.eu/bitstream/10419/201268/1/ETLA-Working-Papers-53.pdf, zuletzt geprüft am 16.02.2022.

Lawrenz, Sebastian; Poschmann, Hendrik; Rausch, Andreas; Stein, Vera (2022): Data Trading Similarity Signature An Extended Data Trading Framework for Human and Non-Human Actors. In: Tung Bui (Hg.): Proceedings of the 55th Hawaii International Conference on System Sciences. Hawaii International Conference on System Sciences: Hawaii International Conference on System Sciences (Proceedings of the Annual Hawaii International Conference on System Sciences). Online verfügbar unter: https://scholarspace.manoa.hawaii.edu/bitstream/10125/79937/0483.pdf, zuletzt geprüft am 10.03.2022.

Leiting, Tobias; Rix, Calvin; Holst, Lennard (2022): Herausforderungen der Preisbildung datenbasierter Geschäftsmodelle in der produzierenden Industrie. In: Marieke Rohde, Kristina Peneva, Matthias Bürger und Johannes Mock (Hg.): Wie aus Daten Wert entsteht – Datenwirtschaft und Datentechnologie (im Erscheinen). Wiesbaden: Springer Vieweg.

Lindner, Maximilian; Straub, Sebastian; Kühne, Bettina (2021): How to share data? Data-Sharing-Plattformen für Unternehmen. Betriebswirtschaftliche und juristische Grundlagen, aktuelle Praxisprojekte und erste Handlungsempfehlungen. iit. Berlin. Online verfügbar unter: https://www.iit-berlin.de/wp-content/uploads/2021/04/SDW_Studie_DataSharing_ES-1.pdf, zuletzt geprüft am 23.02.2022.

Mahanti, Rupa (2019): Data Quality. Dimensions, Measurement, Strategy, Management, and Governance. Milwaukee, WI: Quality Press. Online verfügbar unter: https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=6262212.

Meisel, Lukas; Spiekermann, Markus (2019): Datenmarktplätze – Plattformen für den Datenaustausch und Datenmonetarisierung in der Data Economy. Fraunhofer ISST (ISST Berichte). Online verfügbar unter: https://www.isst.fraunhofer.de/content/dam/isst-neu/documents/Publikationen/Datenwirtschaft/2019-2_ISST-Bericht_Datenmarktplaetze-ISSN-0943-1624.pdf, zuletzt geprüft am 23.02.2022.

Moossavizadeh, Seyed Mohammad Hossein; Mohsenzadeh, Mehran; Arshadi, Nasrin (2012): A new approach to measure believability dimension of data quality. In: MSL 2 (7), S. 2565–2570. DOI: 10.5267/j.msl.2012.07.007.

Nemani, Rao R.; Konda, Ramesh (2009): A Framework for Data Quality in Data Warehousing. In: Jianhua Yang, Athula Ginige, Heinrich C. Mayr und Ralf-D Kutsche (Hg.): Information Systems: Modeling, Development, and Integration, Bd. 20. Berlin, Heidelberg: Springer Berlin Heidelberg (Lecture Notes in Business Information Processing), S. 292–297.

Oliveira, Paulo; Rodrigues, Fátima; Henriques, Pedro Rangel (2005): A Formal Definition of Data Quality Problems. In: ICIQ.

Pezoulas, Vasileios C.; Kourou, Konstantina D.; Kalatzis, Fanis; Exarchos, Themis P.; Venetsanopoulou, Aliki; Zampeli, Evi et al. (2019): Medical data quality assessment: On the development of an automated framework for medical data curation. In: Computers in biology and medicine 107, S. 270–283. DOI: 10.1016/j.compbiomed.2019.03.001.

Prat, Nicolas; Madnick, Stuart E. (2007): Measuring Data Believability: A Provenance Approach. In: SSRN Journal 12 (4), S. 5. DOI: 10.2139/ssrn.1075723.

Ramasamy, Anandhi; Chowdhury, Soumitra (2020): Big Data Quality Dimensions: A Systematic Literature Review. In: JISTEM (17). DOI: 10.4301/S1807-1775202017003.

Rapp, Jannis (2020): Datenqualitätsmetriken zur Unterstützung von Domänenexperten bei interaktiven Analysen. Abschlussarbeit (Bachelor). Universität Stuttgart. Institut für Parallele und Verteilte Systeme. Online verfügbar unter: http://dx.doi.org/10.18419/opus-11118, zuletzt geprüft am 10.03.2022.

Rea, Nick; Sutton, Adam (2019): Putting a value on data. PWC. Online verfügbar unter: https://www.pwc.co.uk/data-analytics/documents/putting-value-on-data.pdf, zuletzt geprüft am 23.02.2022.

Redman, Thomas C. (2001): Data quality. The field guide. Boston: Digital Press.

Rohweder, Jan P.; Kasten, Gerhard; Malzahn, Dirk; Piro, Andrea; Schmid, Joachim (2008): Informationsqualität – Definitionen, Dimensionen und Begriffe. In: Knut Hildebrand, Marcus Gebauer, Holger Hinrichs und Michael Mielke (Hg.): Daten- und Informationsqualität, Bd. 11. Wiesbaden: Vieweg+Teubner, S. 25–45.

Sassenberg, Thomas; Faber, Tobias; Bodungen, Benjamin von; Mantz, Reto (Hg.) (2020): Rechtshandbuch Industrie 4.0 und Internet of Things. Praxisfragen und Perspektiven der digitalen Zukunft. 2. Auflage. München: C.H. Beck; Vahlen.

Schweitzer, Heike; Peitz, Martin (2017): Datenmärkte in der digitalisierten Wirtschaft: Funktionsdefizite und Regelungsbedarf? ZEW Zentrum für Europäische Wirtschaftsforschung GmbH (Discussion Paper, 17-043). Online verfügbar unter: http://ftp.zew.de/pub/zew-docs/dp/dp17043.pdf, zuletzt geprüft am 23.02.2022.

Shanks, Graeme; Darke, Peta (1998): Understanding data quality in a data warehouse: a semiotic approach. In: Proceedings of conference on information quality: University of Massachusetts Lowell, S. 292–309.

Stein, Hannah; Groen in't Woud, Florian; Holuch, Michael; Mulryan, Dominic; Froese, Thomas; Holst, Lennard (2022): Bewertung von Unternehmensdatenbeständen: Wege zur Wertermittlung des wertvollsten immateriellen Vermögensgegenstandes. In: Marieke Rohde, Kristina Peneva, Matthias Bürger und Johannes Mock (Hg.): Wie aus Daten Wert entsteht – Datenwirtschaft und Datentechnologie (im Erscheinen). Wiesbaden: Springer Vieweg.

Strong, Diane M.; Kahn, Beverly (Hg.) (1997): Second Conference on Information Quality (IQ 1997): MIT.

Strong, Diane M.; Lee, Yang W.; Wang, Richard Y. (1997): Data quality in context. In: Commun. ACM 40 (5), S. 103–110. DOI: 10.1145/253769.253804.

Su, Zhanming; Jin, Zhanming (2007): A Methodology for Information Quality Assessment in the Designing and Manufacturing Processes of Mechanical Products. In: Latif Al-Hakim (Hg.): Information Quality Management: IGI Global, S. 190–220.

Timocin, Teoman (2020): Data Quality in the Interface of Industrial Manufacturing and Machine Learning. Abschlussarbeit (Master). Uppsala University. Faculty of Social Sciences, Department of Business Studies. Online verfügbar unter: http://uu.diva-portal.org/smash/record.jsf?pid=diva2%3A1469008&dswid=-9305, zuletzt geprüft am 08.03.2022.

Trauth, Daniel; Mayer, Johannes (2022): Grenzkostenfreie IoT-Services in den Datenmarktplätzen der Zukunft. In: Marieke Rohde, Kristina Peneva, Matthias Bürger und Johannes Mock (Hg.): Wie aus Daten Wert entsteht – Datenwirtschaft und Datentechnologie (im Erscheinen). Wiesbaden: Springer Vieweg.

Wand, Yair; Wang, Richard Y. (1996): Anchoring data quality dimensions in ontological foundations. In: Commun. ACM 39 (11), S. 86–95. DOI: 10.1145/240455.240479.

Wang, Richard Y.; Strong, Diane M. (1996): Beyond Accuracy: What Data Quality Means to Data Consumers. In: Journal of Management Information Systems 12 (4), S. 5–33. Online verfügbar unter: http://www.jstor.org/stable/40398176

Wiedau, Michael; Tolksdorf, Gregor; Oeing, Jonas; Kockmann, Norbert (2021): Towards a Systematic Data Harmonization to Enable AI Application in the Process Industry. In: Chemie Ingenieur Technik 93 (12), S. 2105–2115. DOI: 10.1002/cite.202100203.

Zhang, Dan; Wang, Hongzhi; Ding, Xiaoou; Zhang, Yice; Li, Jianzhong; Gao, Hong (2018): On the Fairness of Quality-based Data Markets. Online verfügbar unter: https://arxiv.org/pdf/1808.01624

Zhang, Lina; Jeong, Dongwon; Lee, Sukhoon (2021): Data Quality Management in the Internet of Things. In: Sensors (Basel, Switzerland) 21 (17). DOI: 10.3390/s21175834.

Zhang, Ruojing; Indulska, Marta; Sadiq, Shazia (2019): Discovering Data Quality Problems. In: Bus Inf Syst Eng 61 (5), S. 575–593. DOI: 10.1007/s12599-019-00608-0.