technische universität dortmund

IT in Produktion und Logistik

Systematische Analyse von Datenqualitätsmetriken im Kontext von Prozessen in produzierenden Unternehmen

Masterarbeit zur Erlangung des Grades M. Sc.

Vorgelegt von:	Lucas Hupe
Matrikelnummer:	185433
Studiengang:	Maschinenbau

Ausgabedatum:	14.07.2023
Abgabedatum:	25.12.2023

Erstprüfer:	Dr.-Ing. Anne Antonia Scheidler
Zweitprüfer:	Florian Hochkamp

Technische Universität Dortmund

Fakultät Maschinenbau

Fachgebiet IT in Produktion und Logistik

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

CRISP-DM	Cross Industry Standard Process for Data Mining

Formelverzeichnis

1 Einleitung

Im ersten Kapitel dieser Arbeit wird zunächst in das Thema eingeführt. Es wird verdeutlicht welche Relevanz das Thema Metriken in Bezug auf Datenqualität besitzt. Dabei wird neben der Forschungsfrage auch der Aufbau und die Vorgehensweise der Arbeit erläutert.

Mit der wachsenden Digitalisierung von Produktionsprozessen und der Annäherung an die Industrie 4.0 stehen zunehmend umfassende Datensätze zur Verfügung, die detaillierte Einblicke in die Produktionsabläufe ermöglichen (Trunzer et al. 2019). Die zu verarbeitenden Datenmengen nehmen dabei kontinuierlich an Umfang und Vielfalt zu. Ihre Art und Struktur differenziert sich verstärkt, wodurch die Anforderung, diese Daten möglichst in Echtzeit zu verarbeiten, zunehmend erschwert wird (McAfee und Brynjolfsson 2012). Neben diesen Herausforderungen nimmt die strategische Bedeutung der Datenanalyse für produzierende Unternehmen jedoch stetig zu und entwickelt sich zu einem zentralen Wettbewerbsfaktor (Dilda et al. 2017). Abgesehen von der Verbesserung bestehender Prozesse und Produkte eröffnet der Einsatz fortschrittlicher Datenanalysemethoden die Möglichkeit, nicht nur bestehende Positionen zu stärken, sondern auch neue Wettbewerbsvorteile zu schaffen (Dilda et al. 2017). Die erhaltenen Daten können als wertvoller Rohstoff zur Gewinnung entscheidender Wirtschaftsfaktoren und Wissen betrachtet werden (Plaue 2021). Die dadurch gewonnenen operativen Vorteile für die Qualität und Kosten spielen eine entscheidende Rolle in der Effizienzsteigerung produzierender Unternehmen (Porter und Heppelmann 2014). Die Nutzung dieser Vorteile auf beiden Ebenen ist von hoher Wettbewerbsrelevanz, da sie dazu beiträgt, die Gesamtleistungsfähigkeit und Wirtschaftlichkeit nachhaltig zu verbessern (Krechting 2021). Es wird geschätzt, dass Probleme mit der Datenqualität allein in den USA zu Kosten in Milliardenhöhe führen (Batini und Scannapieco 2006). Würthele (2003) wies dabei auf die besonders problematisch, nicht präzisen zu quantifizierbaren finanziellen Auswirkungen hin, die entstehen, wenn aufgrund einer mangelhaften Datenlage fehlerhafte Entscheidungen getroffen werden. Bei schlechter Qualität der zugrunde liegenden Daten liefern die Analysen ein verzerrtes Bild der realen Situation. Die darauffolgenden Entscheidungen können dadurch zu fehlerhaften operativen und sogar gravierenden strategischen Maßnahmen führen. Somit hängt der Erfolg eines Unternehmens maßgeblich von einer qualitativ hochwertigen Datenbasis ab (Würthele 2003).

Um die Qualität der Daten im Laufe der Zeit messen und vergleichen zu können, werden Datenqualitätsmetriken genutzt (Sebastian-Coleman 2012). Mit Hilfe einer Datenqualitätsmetrik lassen sich spezifische Aspekte der Qualität eines Datensatzes analysieren und Maßnahmen zur Verbesserung der Qualität quantifizieren (Sebastian-Coleman 2012). Angesichts dieser Tatsachen ist es durchaus bemerkenswert, dass sich bisher in der Wissenschaft und Praxis noch keine etablierten Ansätze zur Messung der Datenqualität durchgesetzt haben (Heinrich und Klier 2009).

Das Ziel dieser Arbeit besteht in einer umfassenden Analyse der in der Literatur beschriebenen Datenqualitätsmetriken zur Charakterisierung der Datenqualität in produzierenden Unternehmen. Zur Erreichung dieses Zieles werden weitere Teilziele definiert. Das erste Teilziel umfasst die Ableitung des erforderlichen Grundverständnisses, welches für den weiteren Verlauf der Arbeit notwendig ist. Dabei sollen die Begrifflichkeiten Hierarchie dies Wissens, Wissensentdeckung in Datenbanken, Datenvorverarbeitungsverfahren sowie Datenqualität in Bezug auf produzierende Unternehmen erläutert werden. Weiterhin wird eine Auflistung der in der Literatur beschriebenen Datenqualitätsdimensionen angefertigt. Im Anschluss wird auf die Bedeutung der Datenqualitätsmetriken eingegangen. Nachdem dieses Teilziel erreicht wurde, wird auf Basis der beschriebenen Datenqualitätsdimensionen eine systematische Literaturrecherche durchgeführt. Hier wird das Ziel verfolgt passende Datenqualitätsmetriken für die Dimensionen zu identifizieren. Im Anschluss sollen die Datenqualitätsmetriken bewertet und es eine Handlungsempfehlung ausgesprochen werden in Bezug auf die Nutzung in produzierenden Unternehmen. Darauf aufbauend kann das dritte Teilziel bearbeitet werden. Hier wird an einem Fallbeispiel exemplarisch ausgewählte Datenqualitätsmetriken errechnet. Durch die Durchführung exemplarisch gewählter Datenvorverarbeitungsverfahren soll der Einfluss der Verfahren auf die Datenqualitätsmetriken evaluiert und diskutiert werden. Durch die Durchführung exemplarischer Datenvorverarbeitungsschritte sollen Erkenntnisse darüber gewonnen werden, wie diese Verfahren die definierten Datenqualitätsmetriken beeinflussen können.

Um die in diesem Kapitel definierten Ziele zu erreichen, führt die Arbeit zunächst in die Domäne der Datenwissenschaft und in den Bereich der Wissensentdeckung in Datenbanken ein. Hierzu wird im ersten Abschnitt des zweiten Kapitels anhand der Wissenstreppe die in Beziehung stehenden Begriffe Zeichen, Daten, Informationen und Wissen voneinander abgegrenzt. Im Anschluss daran wird im nächsten Abschnitt die Bedeutung von Daten für produzierende

Unternehmen erläutert und die aktuellen Nutzungsfelder aufgezeigt. Weiterhin befasst sich Abschnitt 2.3 mit den Herausforderungen von produzierenden Unternehmen mit Daten in Bezug auf Industrie 4.0 und Big Data. Abschnitte 2.4 sowie 2.5 befassen sich, in Vorbereitung auf Kapitel 4, mit der Datenvorverarbeitung. In diesem Zusammenhang wird der Prozess der Wissensentdeckung in Datenbanken anhand des Vorgehensmodells Cross Industry Standard Process for Data Mining (CRISP-DM) erläutert. Zum Abschluss des zweiten Kapitels wird in Abschnitt 2.6 näher auf den Qualitätsbegriff eingegangen. Hierfür wird zuerst der Begriff Datenqualität definiert und im Anschluss werden Dimensionen vorgestellt, die zur Charakterisierung von Datenqualität genutzt werden. Hierfür wird eine Liste von in der Literatur beschriebenen Datenqualitätsdimensionen erstellt. Anschließend werden typische Datenqualitätsmängel in produzierenden Unternehmen präsentiert sowie der Begriff Datenqualitätsmetriken definiert. Weiterhin wird in diesem Abschnitt ein Modell vorgestellt, dass die Nutzung von bestimmten Datenqualitätsdimensionen vorschlägt. Aufbauend auf den identifizierten Datenqualitätsdimensionen aus Abschnitt 2.6.1, werden im dritten Kapitel verschiedene Datenqualitätsmetriken anhand einer systematischen Literaturrecherche herausgearbeitet. In Abschnitt 3.1 werden die identifizierten Datenqualitätsmetriken vorgestellt. Im folgenden Abschnitt 3.2 werden Datenqualitätsmetriken, passend zu den Dimensionen aus dem vorgestellten Modell aus Abschnitt 2.6, gegenübergestellt und für die Nutzbarkeit in produzierenden Unternehmen bewertet. Im vierten Kapitel der Arbeit schließt sich ein Vergleich von Datenqualitätsmetriken bei unterschiedlichen Datenvorverarbeitungsverfahren an. Dazu werden an einem Fallbeispiel die wesentlichen Schritte der Datenvorverarbeitung verschiedener Verfahren beschrieben und durchgeführt. Dazu werden die Datenqualitätsmetriken nach der Durchführung der Verfahren ermittelt und anschließend verglichen und kritisch eingeordnet. In Kapitel **Fehler! Verweisquelle konnte nicht gefunden werden.** wird die Arbeit kritisch diskutiert und eine Handlungsempfehlung zur Nutzung von Datenqualitätsmetriken für produzierende Unternehmen ausgesprochen. Abschließend gibt es einen kurzen Ausblick, der auf den Ergebnissen dieser Arbeit basiert und Impulse für zukünftige Forschungsbereiche der Datenwissenschaft aufzeigt.

2 Grundlagen der Datenanalyse und Datenqualität

Im folgenden Kapitel wird das theoretische Grundverständnis dieser Arbeit hergeleitet und erläutert. Die untereinander in Verbindung stehenden Begriffe Zeichen, Daten, Informationen und Wissen werden anhand der Wissenstreppe nach North et al. (2021) definiert und voneinander abgegrenzt. Im Anschluss werden in Abschnitt 2.2 Herausforderungen und Nutzungfeldern von produzierenden Unternehmen erläutert. Dazu werden in Abschnitt 2.3 noch Charakteristiken von Daten in Bezug produzierenden Unternehmen im Hinblick auf Industrie 4.0 und Big Data vorgestellt. Darüber hinaus wird in Abschnitt 2.4 und 2.5 ein Überblick über Wissensentdeckung in Datenbanken mit speziellem Fokus auf den Aspekt der Datenvorverarbeitung gegeben. Dazu wird anhand des CRISP-DM die Wissensentdeckung in Datenbanken beschrieben und im Anschluss die Durchführung verschiedener Datenvorverarbeitungsverfahren vorgestellt. In Abschnitt 2.6 wird die Relevanz von Datenqualität vermittelt und der integrierte Terminus *Datenqualität* ausführlich analysiert, wobei verschiedene Definitionen aus der einschlägigen Literatur gegenübergestellt werden. Zudem werden die häufig in der Fachliteratur zitierten prägnanten Dimensionen der Datenqualität näher beleuchtet. Weiterhin wird auf den Begriff Datenqualitätsmängel eingegangen und erläutert. Im letzten Abschnitt des zweiten Kapitels wird das Thema Datenqualitätsmetriken zur Messung von Datenqualitätsdimensionen aus wissenschaftlicher Sicht beleuchtet

2.1 Hierarchie des Wissens

Um ein fundiertes Verständnis von Daten, Informationen und den sich daraus ableitenden Prozessen zur Wissensgewinnung zu erlangen, ist es zunächst erforderlich, eine Definition und Einordnung unterschiedlicher Termini aus dem Feld der Datenwissenschaft vorzunehmen. Um die Abgrenzung zwischen den Begriffen Zeichen, Daten, Informationen und Wissen zu schaffen, wird in dieser Arbeit die Wissenstreppe nach North et al. (2021) gewählt. North (2021) stellt die Zusammenhänge zwischen den oben genannten Begriffen mithilfe der sogenannten Wissenstreppe dar. Die wesentlichen Grundelemente der Wissenstreppe sind in Abbildung 1, die im Folgenden präsentiert wird, veranschaulicht.

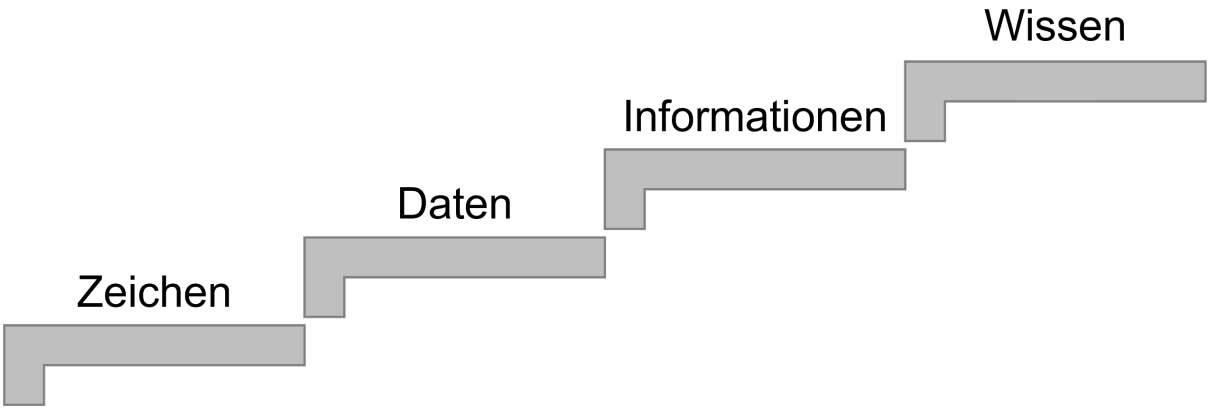

Abbildung 1: Wissenstreppe (i. A. an North 2021, S. 37)

Die erste Stufe der Wissenstreppe beschäftigt sich mit dem Begriff Zeichen. Zeichen können dabei aus Buchstaben, Ziffern oder Sonderzeichen bestehen und werden durch eine Ordnungsregel, beispielsweise einer Syntax, zu Daten (North 2021). Daten sind dabei Symbole, also eine beliebige Zeichenfolge, die zu diesem Zeitpunkt noch nicht interpretierbar sind. Diesen Daten muss zuerst eine Bedeutung zugeordnet werden, wodurch im Anschluss Informationen entstehen (North 2021). Werden diese gewonnenen Informationen nun in einen Kontext gesetzt oder mit bereits gewonnen Erfahrungen kombiniert, entsteht Wissen (North 2021). North (2021) beschreibt Wissen weiterhin als einen Prozess, der zur zweckdienlichen Vernetzung von Informationen dient. Dadurch entsteht, dass Informationen durch das Bewusstsein verarbeitet wird (Bode 1997). Probst, Raub und Romhardt (2012) argumentieren dafür, dass Wissen sämtliche Kenntnisse und Fertigkeiten umfasst, die für die Bewältigung von Problemen erforderlich sind. Über weitere Stufen, die in Abbildung 1 explizit nicht dargestellt worden sind, kann die Nutzung von Wissen über Handeln und Kompetenz zur Wettbewerbsfähigkeit führen (North 2021).

Die Disziplin der Datenwissenschaft beschäftigt sich genau mit diesem Aspekt der Wissensnutzung (Probst et al. 2012). Der Begriff Data Science wird seit den 1990er Jahren zunehmend genutzt und hat sich seitdem zu einer eigenständigen wissenschaftlichen Disziplin entwickelt (Smith 2006). Weiterhin erläutert Smith (2006), dass die Datenwissenschaft die Lehre von der Erfassung von Daten, deren Analyse, Metadaten, schnellen Abruf, Archivierung, Austausch, Mining zum Auffinden von unerwartetem Wissen und Datenzusammenhängen, Visualisierung in zwei und drei Dimensionen einschließlich Bewegung und Management beinhaltet. Plaue (2021) erweitert dies und fasst die zentrale Aufgabe der Datenwissenschaft als „*Erfassung,*

Verarbeitung, Interpretation und Kommunikation von Daten mit dem Ziel der Gewinnung von belastbarem und nutzbringendem Wissen“ (Plaue 2021) zusammen.

2.2 Herausforderungen produzierender Unternehmen im Umgang mit Daten

Nachdem dem im vorherigen Abschnitt die grundlegenden Konzepte und Grundlagen zum Thema Daten geschaffen wurde, wird nun der Bezug zu produzierenden Unternehmen hergestellt.

Die Domäne der Datenwissenschaft beschäftigt sich im Hinblick auf produzierende Unternehmen mit ebendieser Ausschöpfung des Wissens und insbesondere mit der datengestützten Wissensgenerierung (Schuh et al. 2019). Gerade im Hinblick auf die einhergehenden Digitalisierung im Zuge der Industrie 4.0 treten neue Optionen zur Datengewinnung auf (Eigner et al. 2016). Diese Entwicklung legt den Grundstein für die industrielle Anwendung im Bereich der Industrial Data Science (Eickelmann et al. 2015). Viele produzierenden Unternehmen haben die Bedeutung und die daraus resultierende Vorteile erkannt, jedoch kommen diese oftmals nicht über ein Planungsstadium hinaus (Eickelmann et al. 2015). Ursachen hierfür liegen neben der fehlenden Kompetenz meist eher, wie in **Fehler! Verweisquelle konnte nicht gefunden werden.** und Abbildung 3 dargestellt, in der Bewahrung der Firmengeheimnisse und finanziellen Restriktionen (Bange und Janoschek 2014).

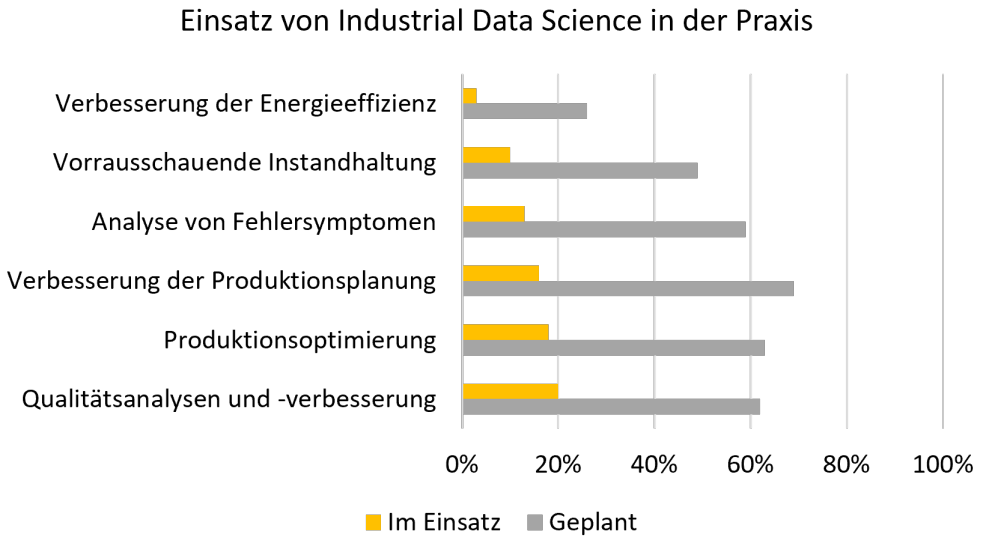

Abbildung 2: Einsatz von Industrial Data Science in der Praxis (i. A. an Bange und Janoschek 2014, S. 34)

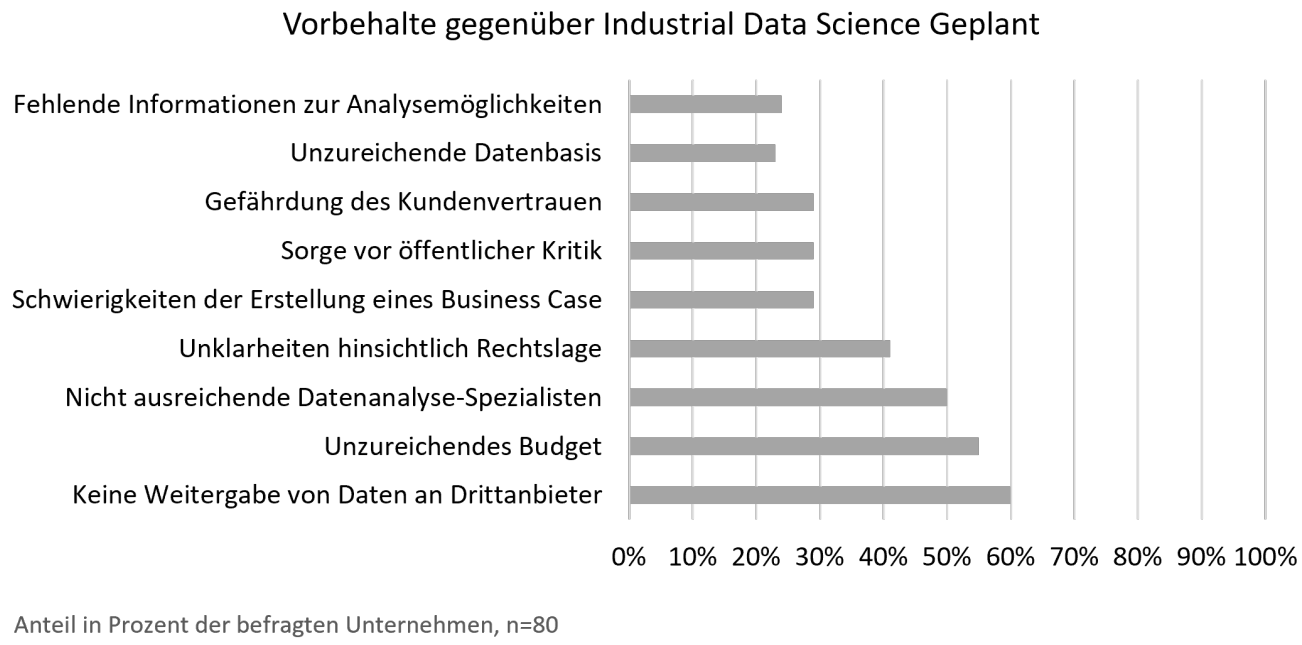

Abbildung 3: Vorbehalte von Industrial Data Science in der Praxis (i. A. an Bange und Janoschek 2014, S. 34)

Neben den Vorbehalten kommen noch weitere Herausforderungen hinzu, die gelöst werden müssen, um eine erfolgreiche Umsetzung datengetriebener Entscheidungsunterstützung sicherzustellen. Diese kann man in technische und unternehmenskulturelle Herausforderungen untergliedern (Schmitt et al. 2020).

Im Zentrum der technischen Herausforderungen steht die Erfassung von Rohdaten aus verschiedenen Domänen (Schmitt et al. 2020). Dies beinhaltet einerseits die Erfassung und Integration unterschiedlicher Datentypen, sei es strukturiert oder unstrukturiert, ohne dabei relevante Kontextinformationen zu verlieren (Schmitt et al. 2020). Andererseits birgt die Extraktion von Daten aus verschiedenen heterogenen Datenquellen selbst einige Schwierigkeiten. Das Aachener Internet of Production sieht hierbei vor, eine sogenannte Middleware+ zur Verwaltung des Datenzugriffs auf verschiedene proprietäre Systeme einzusetzen, um einen Data Lake als Grundlage für anwendungsspezifische Analysen zu etablieren (Schuh et al. 2017). Zusätzlich kann es erforderlich sein, veraltete Technologie oder Fremdmaschinen nachzurüsten, um gezielt fehlende Daten auf dem Shopfloor zu erfassen (Schmitt et al. 2020).

Neben den technischen Herausforderungen ergeben sich ebenso organisatorisch bedeutende Herausforderungen, insbesondere im Zusammenhang mit der Integration der Mitarbeiterinnen

und Mitarbeiter (Schmitt et al. 2020). Ein erfolgreicher Einsatz von datengetriebenen Entscheidungsunterstützungen erfordert ein Umdenken bei den Mitarbeiterinnen und Mitarbeitern (Schmitt et al. 2020). Dieses Umdenken kann nur durch gezielte Maßnahmen zur Förderung des Vertrauens in die Leistungsfähigkeit der Systeme erreicht werden (Schmitt et al. 2020). Zusätzlich zur Förderung der Akzeptanz bei den Mitarbeiterinnen und Mitarbeitern müssen unternehmensintern oft auch monetäre Vorteile solcher Systeme nachgewiesen werden, um die Investitionen zu rechtfertigen (Schmitt et al. 2020).

Im externen Kontext des Unternehmens stellen sich die größten Herausforderungen bei der Einbindung von Kunden und Lieferanten (Schmitt et al. 2020). Bei der länderübergreifenden Nutzung von Daten muss oft der Kunde überzeugt werden, seine Daten zu teilen, beispielsweise durch die Aufklärung über die ihm entstehenden Mehrwerte. In diesem Zusammenhang müssen auch rechtliche Fragen im Hinblick auf die gesammelten Daten geklärt werden: Wer ist der Eigentümer der Daten? Wer darf die Daten zu welchen Zwecken verwenden? Ähnlich wie bei den eigenen Mitarbeitern erfordert dies eine gezielte Förderung des Vertrauens in datengetriebene Unterstützungssysteme beim Kunden, insbesondere bei der Einführung neuer Geschäftsmodelle (Schmitt et al. 2020).

2.3 Charakteristiken von Daten produzierender Unternehmen

Die Begriffe *Industrie 4.0* und *Big Data* werden immer bedeutsamer in der dynamischen Landschaft produzierender Unternehmen (Gröger 2015)**.** Dieser Begriff markiert einen signifikanten Paradigmenwechsel in der Art und Weise, wie Unternehmen Daten betrachten und nutzen und birgt damit auch neue Herausforderungen (Gröger 2015). *Big Data* repräsentiert in diesem Kontext eine Datenlandschaft, die durch fünf zentrale Merkmale geprägt ist. Sie sind als die 5Vs bekannt: *Volume, Velocity, Variety, Veracity und Value (Bauer et al. 2017; Frehe et al. 2016; Seufert 2016)*. Neben diesen Charakteristika werden in der Literatur noch weitere Merkmale genannt (Holland 2020). Im weiteren Verlauf dieser Arbeit werden jedoch lediglich die 5Vs als Charakteristiken und die daraus folgenden als Herausforderung betrachtet.

Das Merkmal *Volume* beschreibt die produzierte Menge an Daten und die es aus Sicht von Big Data zu lokalisieren sowie zu sichern gilt (Baars und Kemper 2021). Hieraus resultieren für Unternehmen die Herausforderungen, die enorme Anzahl der aufgenommen Daten zu verarbeiten (Gröger 2015; Frehe et al. 2016). Das Kriterium *Velocity* liegt der Schwerpunkt auf

der Geschwindigkeit der Identifikation und Auswertung verfügbarer Daten (Baars und Kemper 2021). Es erfordert eine präzise Analyse der Anforderungen je nach Anwendungsfall und die entsprechende Maßnahmenableitung (Quix 2021). Die Frage nach der Notwendigkeit von Echtzeitdaten oder der Akzeptanz einer längeren Verarbeitungszeit steht dabei im Zentrum (Quix 2021). Die Dynamik der Datenänderung spielt ebenfalls eine entscheidende Rolle, insbesondere im Hinblick auf die Geschwindigkeit sich wandelnder Beziehungen zwischen Daten oder der Veränderung der Daten selbst (Quix 2021). Es ist daher von essenzieller Bedeutung, die Anpassungsfähigkeit der Datenverarbeitung an die spezifischen Dynamiken und Anforderungen des Anwendungsfalls anzupassen (Quix 2021). Die Dimension der *Variety* fokussiert sich auf die Breite der Datenformate sowie die unterschiedlichen Grade der Strukturierung (strukturiert, halbstrukturiert, unstrukturiert), die in der Datenverarbeitung berücksichtigt werden müssen (Baars und Kemper 2021). Dies stellt die eingesetzten Technologien vor erhebliche Herausforderungen, da Daten, um nutzbar zu sein, in bestimmte Formate überführt werden müssen (Quix 2021). Zusätzlich führt die rasante Entwicklung von Systemen oft dazu, dass ältere Versionen nicht mehr kompatibel sind und daher nicht mehr verwendet werden können (Quix 2021). Darüber hinaus erschwert die Vielfalt von Sprachen und Programmiersprachen in den einzelnen Anwendungen die Analyse erheblich (Quix 2021). Die Dimension der *Veracity* legt den Fokus auf die Wahrhaftigkeit und Genauigkeit der Ergebnisse, die durch die Verarbeitung von Big Data entstehen (Baars und Kemper 2021). Die Qualität der Daten spielt hierbei eine zentrale Rolle, da die Glaubwürdigkeit der abgeleiteten Ergebnisse stark von der Qualität der zugrunde liegenden Daten abhängt (Frick et al. 2021). Eine schlechte Datenqualität kann somit die Verlässlichkeit der erzielten Ergebnisse erheblich beeinträchtigen (Frick et al. 2021). Das letzte Merkmal *Value* beschreibt den Anspruch der Generierung eines Mehrwertes aus der Analyse der Daten (Frehe et al. 2016). Die Herausforderungen hierbei umfassen die Balance zwischen dem Anspruch der Generierung eines Mehrwerts aus der Analyse der Daten und der Notwendigkeit, strenge Standards für Privatsphäre und Datenschutz zu wahren (Frehe et al. 2016).

2.4 Wissensentdeckung in Datenbanken

Nachdem die vorherigen Abschnitte die grundlegenden Begriffe aus dem Feld der Datenwissenschaft erklärt und der Zusammenhang zu produzierenden Unternehmen dargestellt worden ist, befasst sich dieser Abschnitt mit der Datenvorverarbeitung. Hierzu wird zuerst der

Prozess der Wissensentdeckung in Datenbanken anhand des Vorgehensmodells CRISP-DM erläutert.

Der CRISP-DM wurde gewählt, da dieses Vorgehen zwar ein frühes aber gleichzeitig noch häufig genutztes Vorgehensmodell in der Industrie darstellt (Lieber et al. 2013). Er wurde durch Chapman et al. (2000) als standardisierter Prozess zur Umsetzung von Data-Mining Prozessen detailliert auf mehreren Ebenen beschrieben. Diese verschiedenen Ebenen werden im Folgenden je nach Relevanz für diese Arbeit unterschiedlich ausführlich erläutert, wobei das Hauptaugenmerk auf der Datenvorverarbeitung liegt.

Das CRISP-DM-Modell besteht aus einem Zyklus, der sechs Phasen umfasst und in Abbildung 4 illustriert wird (Azevedo und Santos 2008). Dabei werden die wichtigsten Beziehungen und Wechselwirkungen zwischen den Phasen mithilfe von Pfeilen angedeutet, wobei nicht alle Beziehungen dargestellt werden können, da auch untergeordnete Aufgaben und Prozesse wiederum miteinander vernetzt sein können.

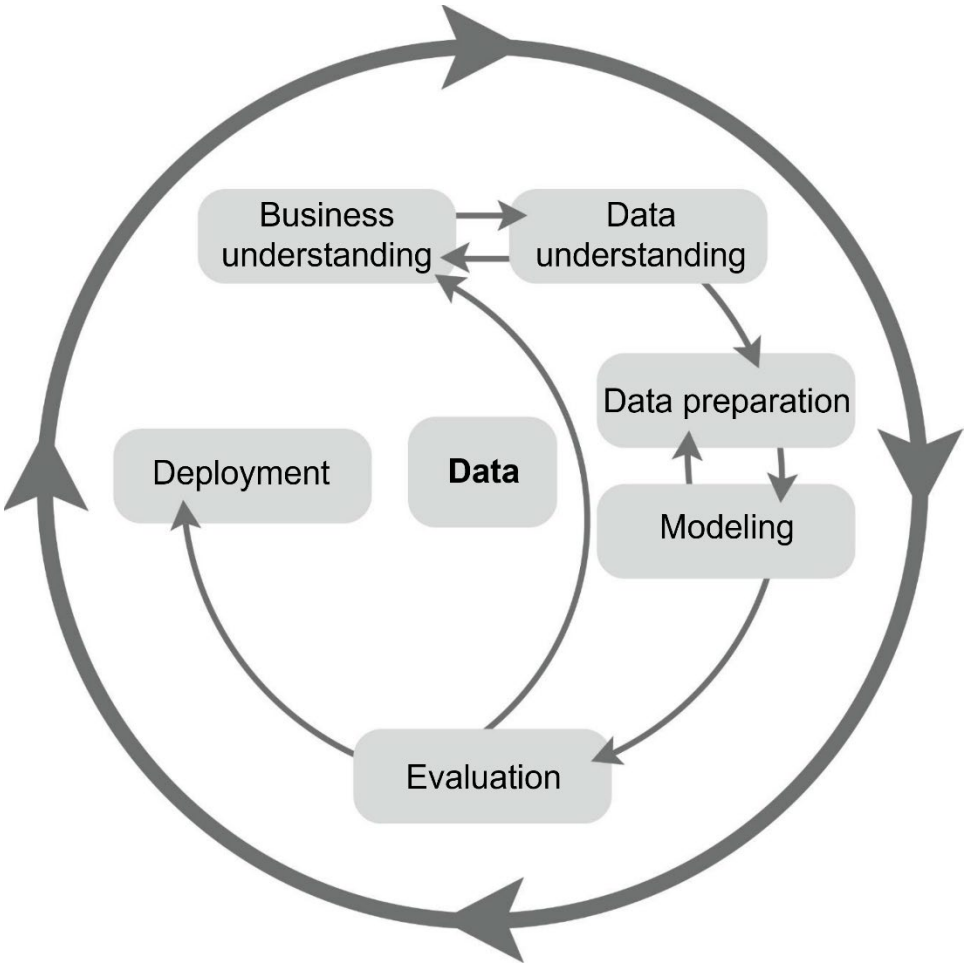

Abbildung 4: CRISP-DM-Vorgehensmodell (i. A. an Chapman et al. 2000, S.13)

Die erste Phase befasst sich mit dem *Business understanding* (Domänenverständnis) (Chapman et al. 2000), in welcher die Geschäftssituation bewertet werden soll. Sie erlaubt einen Überblick und ein Verständnis über die verfügbare und benötigte Ressourcen zu erhalten (Schröer et al. 2021). Weiterhin wird in dieser Phase des Projektes die Problemstellung abgeleitet, die Zielsetzungen sowie ein Projektplan definiert (Schröer et al. 2021). Im Kontext der produzierenden Unternehmen dient diese Phase des *Business understanding* dazu, Geschäftsanforderungen in datenbasierte Lösungen umzuwandeln. Sie führen zu einer Verbesserung der betrieblichen Abläufe, zur Steigerung der Effizienz sowie zur Maximierung der Wettbewerbsfähigkeiten (Azevedo und Santos 2008).

Die zweite Phase, die *Data understanding* (Datenverständnis), beschreibt das Sammeln von Daten und darauf aufbauend das Vertraut machen mit den Daten, um ein allgemeines Datenverständnis zu erlangen (Chapman et al. 2000). Weiterhin wird in dieser Phase die Qualität der Daten geprüft und Probleme identifiziert (Azevedo und Santos 2008). Auf die Qualität der Daten wird in Abschnitt 0 näher eingegangen. Bezogen auf produzierende Unternehmen sollen in dieser Phase Einblicke erworben werden, die zur Steigerung der operativen Effizienz, Verbesserung der Produktqualität sowie der Senkung der Kosten beitragen (Schröer et al. 2021).

Die dritte Phase, *Data preparation* (Datenvorverarbeitung), umfasst in diesem Modell alle Aktivitäten zur Auswahl und Konstruktion des endgültigen Datensatzes aus den bestehenden Rohdaten (Azevedo und Santos 2008). Dieser Schritt beinhaltet dabei nicht nur die Auswahl des Datensatzes sowie der relevanten Attribute, sondern auch die Transformation und Beseitigung von Dateninkonsistenzen (Schröer et al. 2021). Auf diese Phase wird am Ende dieses Kapitels noch einmal eingegangen.

Die vierte Phase, das *Modeling* (Modellierung), beschäftigt sich mit der Auswahl der Modellierungstechnik, die Erstellung der Testcases sowie das Erstellen des Models (Schröer et al. 2021). In dieser Phase werden häufig Iterationen sowie die Rückkehr zu früheren Schritten verwendet, da verschieden Modellierungsmethoden mit unterschiedliche Anforderungen und Lösungsmöglichkeiten existieren (Chapman et al. 2000).

In der fünften Phase, der *Evaluation* (Auswertung), werden das Modell, die daraus entstehenden Ergebnisse sowie die Durchführung der vorherigen Schritte bewertet (Chapman et al.

2000). Anschließend werden die Ergebnisse mit den in der ersten Phase definierten Unternehmensziele verglichen (Schröer et al. 2021).

In der letzten Phase des CRISP-DM, der *Deployment*-Phase (Umsetzung), werden die zuvor gewonnen Ergebnisse und das generierte Wissen in der Praxis eingebracht und umgesetzt (Chapman et al. 2000). Dies kann von der reinen Dokumentation der Erkenntnisse bis hin zur Implementierung von veränderten Prozessen reichen (Chapman et al. 2000)

Nachdem nun das CRISP-DM-Modell eingeführt worden ist, wird im Folgenden ein Überblick über den Aufbau von Datenbanken gegeben, in dem wichtige Begriffe definiert werden. Eine *Entität*, auch oft als Tabellenname bezeichnet, repräsentiert einen spezifischen Themenbereich, in dem Elemente mit ähnlichen Merkmalen oder Eigenschaften gruppiert sind. Die *Entitätsmenge*, die auch als Datensätze bezeichnet wird, umfasst sämtliche Werte, die zu den Merkmalen einer bestimmten Entität gehören. Dies schließt alle gespeicherten Datensätze innerhalb einer Tabelle ein, die auch als *Relation* bezeichnet wird und aus einer Entität und ihrer zugehörigen Entitätsmenge besteht. Bei der Betrachtung dieser Tabelle werden alle Aspekte berücksichtigt, einschließlich ihrer Entitätsbezeichnung, Attributen (auch als Spalten bezeichnet) und Tupeln. Jedes Merkmal oder jeder Wert eines Elements innerhalb dieser Tabelle wird als *Tupel* oder Datensatz bezeichnet. Die Gesamtheit aller Tupel einer Entität bildet wiederum die Entitätsmenge. Ein *Attribut* (der Name einer Spalte) repräsentiert ein spezifisches Merkmal eines Tupels und kennzeichnet eine bestimmte Eigenschaft innerhalb der Entitätsmenge. Der *Attributwert* eines Attributs ist ein konkreter Datenwert, der die Eigenschaft eines Tupels beschreibt. Eine *Datenbasis* umfasst alle vorhandenen Tabellen und somit sämtliche gespeicherte Daten. Die *Datenbank* wiederum beinhaltet nicht nur die Datenbasis, sondern auch das Datenbankverwaltungssystem (Steiner 2017).

2.5 Datenvorverarbeitung

Nachdem im vorherigen Abschnitt ein genereller Überblick über die Wissensentdeckung in Datenbanken gegeben worden ist, wird nun im Hinblick auf produzierende Unternehmen der Fokus auf die Datenvorverarbeitung gelegt. Bei dem Prozess der Datenvorverarbeitung, der Bestandteil des in Abschnitt 2.4 beschriebenen CRISP-DM-Modells ist, werden die Daten im Sinne der Datenselektion, Datenbereinigung und Datentransformation vorverarbeitet (Deuse et al. 2014). Die Datenvorverarbeitung ist eine essenziele Phase der Wissensentdeckung und

nimmt bis zu 80 % der zeitlichen, technischen und personellen Ressourcen in Anspruch (Gabriel et al. 2009). Aufgrund dessen werden diese in der Praxis häufig vernachlässigt (Kureljusic und Karger 2022).

In Abbildung 5 werden die zur Vorverarbeitung definierten Aufgaben mit beispielhaft ausgewählten Aktivitäten sowie Ergebnissen präsentiert und im Anschluss erläutert.

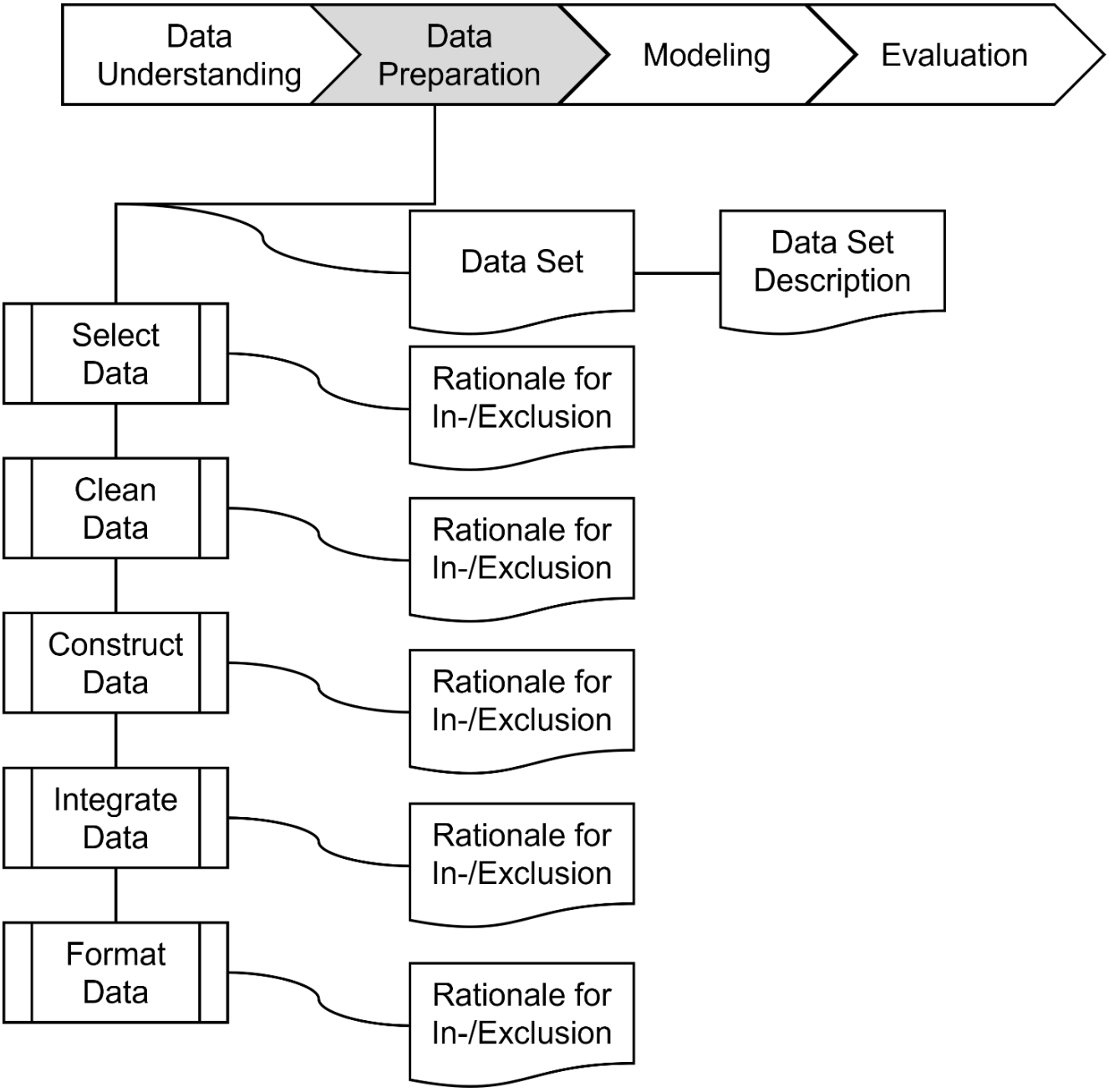

Abbildung 5: Aufgaben und Ergebnisse der Datenvorverarbeitung nach CRISP-DM (i. A. an Chapman et al. 2000, S. 16)

Die Datenauswahl stellt die erste Aufgabe der Datenvorverarbeitung dar. Alle weiteren Aktivitäten finden auf Grundlage dieser Auswahl statt (Chapman et al. 2000). Innerhalb von Data-Science-Projekten werden bei produzierenden Unternehmen Daten aus unterschiedlichen Ursprungsquellen herangezogen. Sie unterscheiden sich strukturell als auch in der Informationsqualität. Daher ist die Auswahl der Daten ein wichtiger Bestandteil der Datenvorverarbeitung (Saleh 2018). Weiterhin müssen produzierende Unternehmen bei der Auswahl, neben der Struktur, die Reduzierung irrelevanter Informationen, wie beispielsweise die Minimierung des Rauschen der Daten, vornehmen (García-Gil et al. 2019). Als nächster Schritt folgt die Datenbereinigung bei der ein fundierter, auswertungsfähiger Datenbestand mit einer möglichst hohen Qualität entsteht (Deuse et al. 2014). Zur Erreichung dieses Zieles werden fehlerhafte, irrelevante, redundante oder unvollständige Werte identifiziert und ersetzt, entfernt oder ergänzt (Lieber et al. 2013). Um diese Aufgaben zu erfüllen, werden Aktivitäten zum Umgang mit allgemeinen Rauschen der Daten festgelegt und angewendet (Chapman et al. 2000). Rauschen ist nach CRISP-DM nicht genauer definiert, bezieht sich aber laut Literatur meist auf Ausreißer, fehlende Merkmalswerte oder besondere Werte (Aggarwal 2015; Bramer 2016). Durch die Datenbereinigung können produzierende Unternehmen falsche Erkenntnisse infolge schlechter Datenqualität und verzerrter Analyseergebnisse vermeiden (Lieber et al. 2013). In der Datenkonstruktion, die je nach Anwendungsfall notwendig oder hilfreich sein kann, ist das Ziel, neue und für das weitere Vorgehen hilfreiche Attribute zu erzeugen oder Werte bereits existierender Attribute zu transformieren (Chapman et al. 2000).

Ein weiterer Bestandteil der Datenvorverarbeitung ist die Datenintegration (Chapman et al. 2000). Mit dem Ziel einen zusammengeführten Datensatz zu generieren, können über einzigartige Kennungen verschiedene Datenbanksysteme logisch miteinander verknüpft werden (Schuh et al. 2023). Aufgrund der Tatsache, dass bei produzierenden Unternehmen Datensätze aus verschiedenen Ursprungsquellen bereitgestellt werden, ist dies ein wichtiger Schritt.

Die abschließende Aufgabe der Datenvorverarbeitung besteht in der Datenformatierung, die zur Vereinfachung der Auswertung benötigt wird (Schuh et al. 2023). Diese Aufgabe beschreibt die Transformation der Daten auf syntaktischer Ebene. Je nach Data-Mining-Technik und -Werkzeug existieren verschiedene Anforderungen an Datentypen oder Attributsreihenfolgen (Chapman et al. 2000). Das Ergebnis dieser Aufgabe wird dementsprechend durch einen formatierten Datensatz dargestellt (Chapman et al. 2000).

Das Gesamtergebnis des ganzen Datenvorverarbeitungsprozesses ist daher der finale Datensatz, der in den weiteren Schritt des CRISP-DM zur Modellierung eingesetzt wird. Im folgenden Abschnitt wird der Fokus auf den Begriff Datenqualität gelegt.

2.6 Datenqualität

Nachdem in den vorherigen Abschnitten die Grundlagen der Wissensentdeckung erläutert und ausführlich über die Datenvorverarbeitung gesprochen worden ist, wird im folgenden Abschnitt der Qualitätsbegriff näher betrachtet. Hierfür wird zuerst der Begriff Datenqualität definiert und im Anschluss in Abschnitt 2.6.1 Dimensionen vorgestellt, die zur Charakterisierung genutzt werden. Anschließend werden in Abschnitt 2.6.2 typische Datenqualitätsprobleme in produzierenden Unternehmen präsentiert sowie ein Modell vorgestellt, welches den Zusammenhang zwischen Dimensionen, Datenqualität sowie Prozessen darstellt. Zum Abschluss wird der Begriff Datenqualitätsmetriken vorgestellt und definiert.

Der Begriff Datenqualität lässt sich durch die wissenschaftliche Disziplin der Semiotik in die syntaktische, semantische und pragmatische Ebene differenzieren (Müller 2000). Auf der syntaktischen Ebene lässt sich dabei die technische Verfügbarkeit und Nutzbarkeit der Daten einsehen. Neben den datenschutzrechtlichen Aspekten ist diese Ebene für produzierende Unternehmen aufgrund der sachgerechten und einheitliche Repräsentation des dargestellten Sachverhalts interessant (Wenzel et al. 2005). Die semantische Ebene wiederum bezieht sich auf Merkmale, die den Informationsgehalt der Daten betreffen. Diese Merkmale sind Präzision, Ausführlichkeit, Gültigkeit und Messbarkeit. Sie umfasst auch verschiedene Aspekte von empirischer und logischer Wahrheit, wie Vertrauenswürdigkeit, Fehlerfreiheit, Kohärenz und Überprüfbarkeit (Wang und Strong 1996).

In seiner ausführlichen Darlegung rückt Miller (1996) den Anwender in den Mittelpunkt seiner Betrachtung. Er betont die zentrale Bedeutung der Informationsqualität im Hinblick darauf, wie diese von den Konsumenten wahrgenommen und in ihrer Anwendung genutzt wird. Dabei definiert er die Informationsqualität anhand der Wahrnehmung verschiedener Merkmale. Die Bewertung der Informationsqualität unterteilt er in zwei aufeinanderfolgende Phasen. Zunächst ist es notwendig, die relevanten Merkmale aus Sicht der Konsumenten zu identifizieren. In einem nächsten Schritt wird analysiert, wie sich diese identifizierten Merkmale auf die indivi-

duellen Nutzer auswirken. Dabei weist Miller auf die essenzielle Rolle von Merkmalen wie Korrektheit, Aktualität, Vollständigkeit, Widerspruchsfreiheit, Format, Zugänglichkeit, Kompatibilität, Sicherheit und Gültigkeit hin. Diese Aspekte sind von besonderer Bedeutung und bilden die Grundlage für die Analyse der Informationsqualität im Kontext seiner Forschungsarbeit (Miller 1996). In Olsons (2003) Beitrag wiederum wird betont, dass die Datenqualität von zwei entscheidenden Faktoren beeinflusst, wird - dem beabsichtigten Verwendungszweck und den Daten selbst. Um den angestrebten Verwendungszweck erfolgreich zu erfüllen, ist es erforderlich, dass die vorliegenden Daten bestimmte Schlüsselmerkmale aufweisen. Diese sind die Datenqualitätsdimensionen, auf die im nächsten Abschnitt näher eingegangen wird. Neben Miller (1996) und Olsen (2003) lässt sich die folgende Definition von Gebauer und Windheuser (2021) heranziehen: "*Datenqualität umfasst sämtliche Qualitätsmerkmale eines Datensatzes in Bezug auf dessen Eignung zur Erfüllung festgelegter und erwarteter Anforderungen*".

In der wissenschaftlichen Literatur zur Datenqualität finden sich viele Übereinstimmungen, insbesondere hinsichtlich der Perspektive des Datenempfängers. Diese anwenderzentrierte Sichtweise betont die Bedeutung, die Daten für den Konsumenten haben. In der Literatur werden zahlreiche Qualitätsmerkmale identifiziert und erläutert. Bei der Betrachtung dieser Definitionen wird deutlich, dass es eine Vielzahl unterschiedlicher Merkmale gibt. Diese Merkmale wurden sowohl aus den Erfahrungen der Praxis als auch aus Expertenwissen und empirischen Studien abgeleitet.

Im Rahmen von Abschnitt 2.6.1 werden diese Merkmale genauer untersucht. Die in Abbildung 6 dargestellte Datenqualitätspyramide veranschaulicht die drei Ebenen erfolgreicher Datenoperationen. Datenqualität kann als eine übergeordnete Kategorie verstanden werden, die die zweite Ebene der Pyramide darstellt. Zur Bewertung von Datenqualitätsmerkmalen sind Datenqualitätsmetriken erforderlich, wie von Gebauer und Windheuser (2021) dargelegt. Die einzelnen Elemente der Datenqualitätspyramide werden in den kommenden Abschnitten ausführlich erläutert.

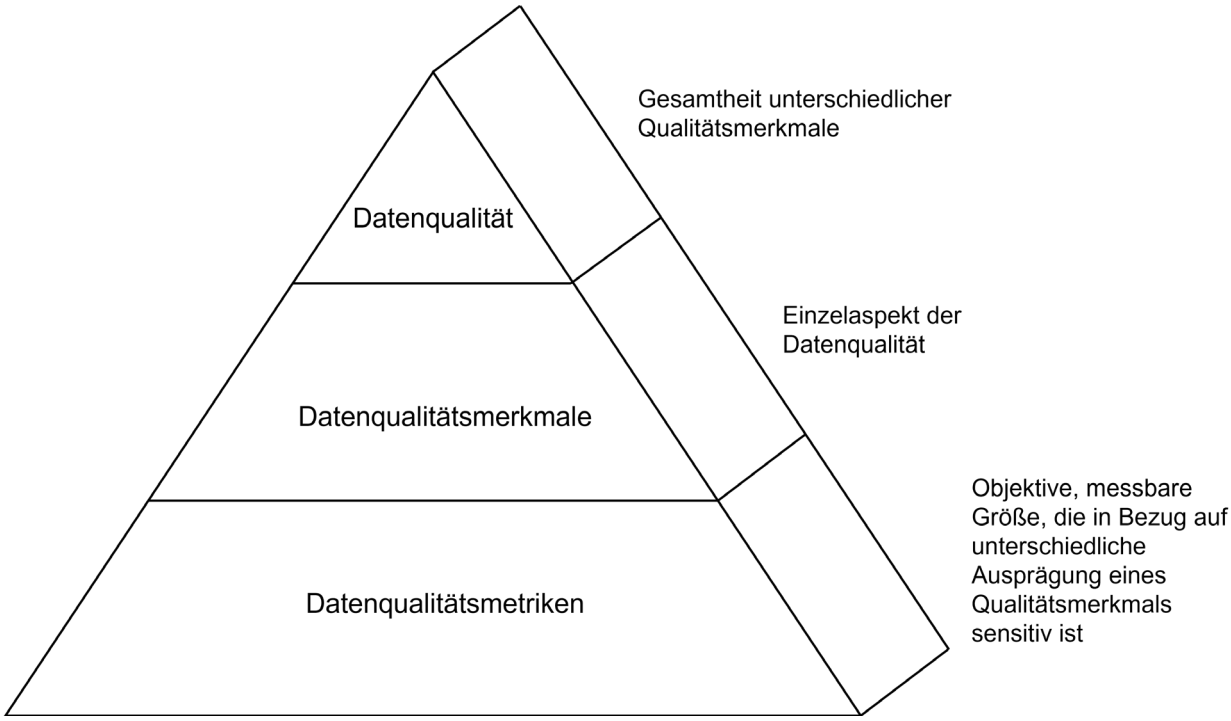

Abbildung 6: Datenqualitätspyramide (i. A. an Gebauer und Windheuser 2011, S. 88)

2.6.1 Datenqualitätsdimensionen

Durch die Nutzung von Datenqualitätsdimensionen kann die Datenqualität näher beschrieben, gemessen und bewertet werden (McGilvray 2021). Der Begriff Dimension wird dabei als Synonym für die Begriffe Merkmal oder Attribut genutzt und wird bereits einige Jahre zur Charakterisierung genutzt. Bereits Anfang der 80er Jahre wurden die ersten Attribute definiert, um ein zusammengesetztes Maß für den Informationswert zu erhalten (King und Epstein 1983). Im Verlaufe der Jahre wurden durch unterschiedliche Autoren weitere Dimensionen definiert, wobei sich viele Autoren und ihre Werke auf Wang & Strong (1996) beziehen und deren Veröffentlichung als Grundlage dient. Wang & Strong (1996) haben in ihrer Abhandlung vier Kategorien entwickelt und die von ihnen gefundenen Datendimensionen klassifiziert und eingeordnet. Die entwickelten Kategorien heißen *intrinsische*, *kontextabhängige*, *begriffliche* und *zugängliche* Datenqualität (Wang und Strong 1996). *Intrinsische Datenqualität* bezieht sich auf die Eigenständigkeit der Datenqualität. *Accuracy* ist lediglich eine von vier Dimensionen, die dieser Kategorie zugrunde liegt. Die kontextabhängige Datenqualität betont die Anforderung, die Datenqualität im Zusammenhang mit der jeweiligen Aufgabe zu betrachten. Dies bedeutet, dass die Daten relevant, zeitnah, vollständig und angemessen in Bezug auf die Menge sein

müssen, um Mehrwert zu schaffen. Begriffliche und zugängliche Datenqualität legen besonderen Wert auf die Rolle von Systemen. Systeme müssen zugänglich und sicher sein und sie sollten Daten so darstellen, dass sie interpretierbar, leicht verständlich, prägnant und konsistent sind (Wang und Strong 1996). In der wissenschaftlichen Literatur werden die Begriffe Datenqualität und Informationsqualität oft synonym verwendet, obwohl es grundlegende Unterschiede gibt (Gebauer und Windheuser 2011). In dieser Arbeit werden diese Begrifflichkeiten ebenfalls synonym verwendet. Neben den von Wang & Strong beschrieben Datenqualitätsdimensionen haben weitere Forscher zusätzliche Datenqualitätsdimensionen erarbeitet. Tabelle 1 illustriert dabei einen Ausschnitt, der am häufigsten referenzierten Dimensionen der Datenqualität und ihre Definitionen aus der Literatur (Sidi et al. 2012). Im Anhang ist die vollständige Tabelle vorzufinden.

Tabelle 1: Auszug der Datenqualitätsdimensionen (i. A. an Sidi et al. 2012)

Dimensionen	Definitionen
Accessibility (Zugänglichkeit)	Das Ausmaß, in dem Informationen verfügbar sind oder leicht und schnell abgerufen werden können (Wang und Strong 1996).
Accuracy (Genauigkeit)	Daten sind genau, wenn die in der Datenbank gespeicherten Datenwerte den realen Werten entsprechen (Batini et al. 2009; Ballou und Pazer 1985). Es bezieht sich auf den Grad, in dem Daten korrekt, zuverlässig und zertifiziert sind (Wang und Strong 1996).
Amount of data (Datenmenge)	Das Ausmaß, in dem die Menge oder das Volumen der verfügbaren Daten für die aktuelle Aufgabe angemessen ist (Wang und Strong 1996).
Appropriate amount of data (Angemessene Datenmenge)	Das Ausmaß, in dem das Datenvolumen für die aktuelle Aufgabe angemessen ist (Pipino et al. 2003)

Um im späteren Verlauf der Arbeit Datenqualitätsdimension von produzierenden Unternehmen bewerten zu können, wird ein Modell von Frehe et al. (2016) genutzt. In dieser Abhandlung wird die Nutzung verschiedener Datenqualitätsdimensionen vorgeschlagen, die die Charakteristiken adressieren und die daraus resultierenden Herausforderungen der 5Vs zu bewältigen. Das in Abbildung 7 entwickelte Modell zeigt eine umfassende Darstellung der Wechselwirkungen zwischen den identifizierten Herausforderungen im Bereich Big Data, den definierten Qualitätsdimensionen und den entsprechenden Datenqualitätsmetriken.

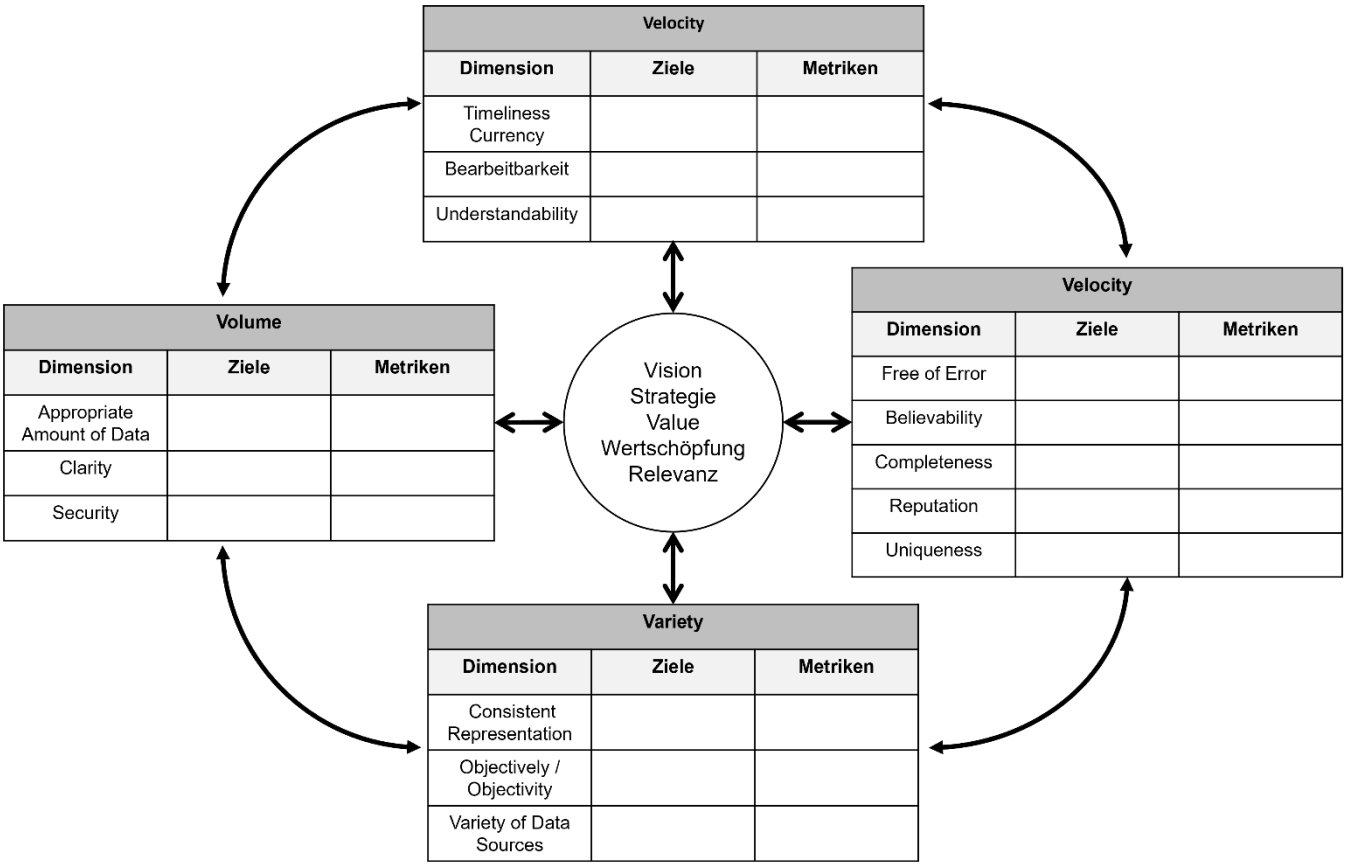

Abbildung 7: Modell zur Bewertung der Datenqualitätsmetriken (i. A. an Frehe et al. 2016, S. 151)

2.6.2 Datenqualitätsmängel

Im Folgenden wird das Thema Datenqualitätsmängel betrachtet. Bereits die anfängliche Dateneingabe ist eine der wichtigsten Ursachen für eine unzureichende Datenqualität bei produzierenden Unternehmen, wobei die Gründe für die fehlerhaft Eingabe mannigfaltig sind (Apel et al. 2015). Zusätzlich dazu ist das fehlende Bewusstsein über mögliche Folgen von Datenqualitätsmängel ein wesentlicher Grund für eine schlechte Datenqualität in der Datenerhebung (Apel et al. 2015). Ein weiterer Aspekt von Datenqualitätsmängel können technische Probleme

bei Sensoren während der Datenerfassung sein (Windelband et al. 2011). Insbesondere im Bereich der Sensoren können unzureichende Integritätsbedingungen auftreten, die die Verlässlichkeit der erfassten Daten beeinträchtigen (Windelband et al. 2011). Technische Probleme oder Fehlfunktionen während der Datenerfassung können zu inkonsistenten oder nicht vollständigen Daten führen (Windelband et al. 2011).

Datenqualitätsmängel lassen sich im Allgemeinen in zwei Klasen unterteilen (Sidi et al. 2012). Zum einen in Single-Source-Probleme, zum anderen in Multi-Source-Probleme (Rahm und Do 2000). Innerhalb dieser Gruppen können weitere Untergruppen identifiziert werden, die in Abbildung 8 dargestellt werden. Das Hauptziel bei der Klassifizierung von Datenqualitätsproblemen besteht darin, nicht standardgemäße Daten zu visualisieren und die präzise Anwendung von Daten gemäß den jeweiligen Anforderungen zu erkennen (Man et al. 2010).

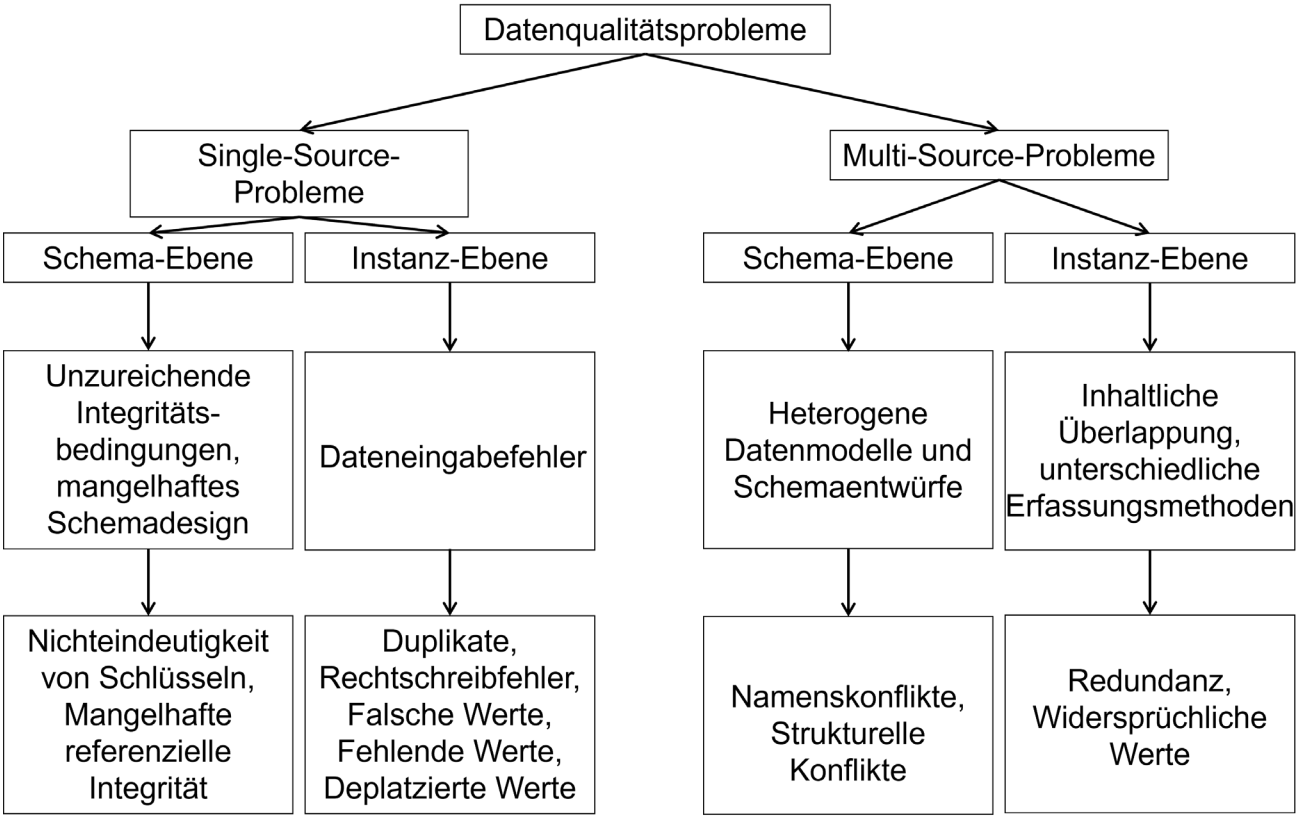

Abbildung 8: Datenqualitätsprobleme (i. A. an Man et al. 2010, S. 29)

2.6.3 Datenqualitätsmetriken

Nachdem in den vorherigen Abschnitten einige Datenqualitätsdimensionen und Datenqualitätsmängel beschrieben worden sind, wird im Folgenden der Begriff Datenqualitätsmetrik vorgestellt und definiert.

Grundsätzlich bezeichnet eine Metrik eine Vorgehensweise zu Messung einer quantifizierbaren Größe und ermöglicht hierdurch eine Objektivität (Witte 2018). 1991 schon hat Harrington beschrieben, dass Messungen der Schlüssel sind, um Daten zu verbessern (Harrington 1991).

Datenqualitätsmetriken sind Kennzahlen, die Daten zusammenfassen und in einen Zusammenhang stellen (Vollmuth und Zwettler 2021). Sie bündeln Daten zu einer aussagekräftigen Größe und stellen komplexe Sachverhalte prägnant dar (Vollmuth und Zwettler 2021). Bevor nun in Kapitel 0 eine detaillierte Übersicht von Metriken zur Bewertung der Datenqualität aufgezeigt wird, ist es von Bedeutung, die Anforderungen an diese Metriken präzise zu definieren um die Reproduzierbarkeit und der wissenschaftliche Standard eingehalten wird. Hierzu werden die in der Literatur vorgestellten Anforderungen an Metriken beschrieben und aufgeführt. Ein grundlegendes Erfordernis besteht darin, die Ergebnisse dieser Metriken zu *normieren*, um ihre Interpretierbarkeit und Vergleichbarkeit sicherzustellen (Hinrichs 2001). Die *Kardinalität* der Metriken spielt ebenfalls eine wichtige Rolle, da sie bei der wirtschaftlichen Bewertung von Maßnahmen und der Überwachung der zeitlichen Entwicklung der Ergebnisse von entscheidender Bedeutung ist (Hinrichs 2001). Dies ermöglicht es, verschiedene Ausprägungen in eine sinnvolle Rangfolge zu bringen und den Grad der Unterschiede zwischen verschiedenen Merkmalsausprägungen zu bestimmen (Azeroual 2022).

Die *Sensibiliserbarkeit* der Metriken ist ein weiteres unverzichtbares Merkmal, um sicherzustellen, dass die erlangten Ergebnisse gezielt und zweckmäßig gemessen werden können (Ehrlinger und Wöß 2022). Dies erfordert eine Anpassung der Metriken an die spezifische Anwendung und die definierten Ziele (Budach et al. 2022). Darüber hinaus können verschiedene miteinander in Beziehung stehende Objekttypen zu einem übergeordneten Objekttyp aggregiert werden, was als A*ggregierbarkeit* bezeichnet wird (Hinrichs 2001). Dies ermöglicht eine flexible Anwendung der Metriken auf verschiedenen Ebenen, sei es auf der Ebene von Attributwerten, Tupeln, Relationen oder sogar auf der Datenbankebene (Hinrichs 2001).

Eine weitere wesentliche Anforderung besteht darin, dass die Ergebnisse der Metriken *fachlich interpretierbar* sein müssen. Dies ist von entscheidender Bedeutung, da die bloße Normierung und Kardinalität allein für die praktische Anwendung nicht ausreichen (Heinrich und Klier 2008).

Neben den Anforderungen an die Metrik werden in der Literatur ebenso Empfehlungen für die Aufbereitung der Daten und Anwendung der Metriken gegeben. Die effektive Verwendung von Metriken setzt voraus, dass alle relevanten Metadaten zur Qualitätsbewertung, einschließlich Informationen wie Datum und Herkunft, verfügbar und zugänglich sind (Hinrichs 2001). In der Praxis ist dies jdeochnicht immer gewährleistet, da die Beschaffung aller erforderlichen Metadaten mit erheblichen Kosten verbunden sein kann (Hinrichs 2001). Dennoch ist die Bereitstellung hochwertiger Metadaten von entscheidender Bedeutung, um eine fundierte Bewertung der Datenqualität durchzuführen (Hinrichs 2001).

Ein etabliertes Verfahren zur Beurteilung der Qualität eines Datenbestandes ist das Hinrichs-Verfahren, das Metriken für ausgewählte Dimensionen entwickelt und anwendet (Lieber et al. 2013). Dabei beziehen sich die Metriken auf ein relationales Datenbankschema, sind aber weiterhin auch auf objektorientierte und objektrelationale Datenmodelle anwendbar (Budach et al. 2022). Im Hinrichs-Verfahren werden die Ergebnisse der Datenqualitätsmetriken effektiver skaliert, indem diese für verschiedene Granularitätsebenen entwickelt werden. Der Ansatz beginnt bei der Messung von Attributwerten und erstreckt sich über Tupel, Relationen bis hin zu Datenbanken (Hinrichs 2001). Eine Metrik auf einer bestimmten Ebene n+1 (z. B. Vollständigkeit auf der Tupel-Ebene) wird basierend auf den zuvor festgelegten Kriterien auf dieser n-Ebene (z. B. Vollständigkeit auf der Attributebene) definiert (Hinrichs 2001). Alle Metriken sind auf einen Wertebereich von]0, 1] oder [0, 1] normiert (Hinrichs 2001). Diese Normierung erleichtert die Aggregation von Metriken auf verschiedenen Ebenen und ermöglicht den Austausch von Formeln zwischen den Metriken (Hinrichs 2001). Alle Metriken verfolgen eine Wertehierarchie, bei der höhere Werte eine bessere Datenqualität repräsentieren (Hinrichs 2001). Dies ermöglicht den Einsatz von Vergleichsoperatoren wie *größer als* oder *kleiner als* (Hinrichs 2001). Die Methode zur Messung, die die Operationalisierung einer Metrik darstellt, wird schließlich der fein granularen Messung (bezogen auf Qualitätsmerkmale) zugeordnet (Hinrichs 2001). Die erfassten Messwerte werden dann auf einer gröberen Ebene gewichtet und gemäß der Metrikdefinition aggregiert (Hinrichs 2001). Die Richtigkeit von Daten kann nur

durch den Vergleich des Zustands der Diskurswelt (Sollzustand) mit dem Zustand des Informationssystems (Istzustand) bewertet werden (Hinrichs 2001). Ein anschauliches Beispiel hierfür ist die Inventur, bei der der tatsächliche Lagerbestand mit den Bestandsinformationen im Informationssystem abgeglichen wird. Die Genauigkeit wird anhand der Ähnlichkeit zwischen dem Attributwert des Datenproduktes und dem Attributwert der in der Diskurswelt repräsentierten Entität bewertet (Hinrichs 2001).

Nachdem die Grundlagen der Datenqualität erläutert wurden, erfolgt im anschließenden Kapitel eine eingehende systematische Literaturrecherche zu Datenqualitätsmetriken, deren Ergebnisse im weiteren Verlauf diskutiert werden.

3 Vergleich von Datenqualitätsmetriken anhand einer systematischen Literaturrecherche

Nach der Einführung der grundlegenden Begriffe und einem Überblick über den aktuellen Stand der Technik in Bezug auf Wissensentdeckung in Datenbanken und Datenqualität in Kapitel 2 dieser Arbeit, wird sich das folgende Kapitel nun mit dem Vergleich von Datenqualitätsmetriken befassen. Dazu wird eine systematische Literaturrecherche durchgeführt, mit dem Ziel, die identifizierten Datenqualitätsmetriken im Anschluss miteinander zu vergleichen und bewerten. Hierfür wird zuerst die Methodik zur Auswertung der Literatur kurz beschrieben bevor in Abschnitt 3.1 die in der Literatur identifizierten Datenqualitätsmetrik vorgestellt werden. In Abschnitt 3.2 werden die beschriebenen Datenqualitätsmetriken in Bezug auf die Nutzung in produzierenden Unternehmen bewertet.

Bei der systematischen Literaturrecherche handelt es sich um eine wissenschaftliche Methode (Fink 2014). Um eine erfolgreiche Recherche durchzuführen, müssen folgende Kriterien bei der Durchführung beachtet werden: Die Recherche muss einem systematischen Vorgehen unterliegen, dessen Ablauf detailliert beschrieben ist und darüber hinaus alle relevanten Veröffentlichungen umfasst (Fink 2014). Hierdurch wird die Reproduzierbarkeit der Ergebnisse gewährleistet und erfüllt damit die Ansprüche an wissenschaftliches Arbeiten (Fink 2014). Das Vorgehen dieser Arbeit richtet sich nach dem Vorgehen von Okoli aus dem Jahr 2015. In dieser Veröffentlichung wird ein einheitliches Vorgehen für systematische Literaturrecherchen vorgeschlagen, die vorherige Ansätze nicht ausschließt, sondern deren Elemente aufgreift (Okoli 2015). Der Ansatz lässt sich in vier Phasen unterteilen. (Okoli 2015).

In der ersten Phase, der Planungsphase, findet die Identifizierung und Dokumentation der Ziele der Literaturrecherche statt (Okoli 2015). In dieser Arbeit sind die Ziele der Literaturrecherche eng mit den Zielen der Arbeit verknüpft. Dadurch konzentriert sich die systematische Literaturrecherche auf die Identifikation von Datenqualitätsmetriken und kann in zwei Hauptziele unterteilt werden. Zum einen verfolgt die systematische Literaturrecherche das Ziel, die vorhandenen Datenqualitätsmetriken umfassend zu erfassen und darzustellen. Dabei geht es darum, einen Überblick darüber zu gewinnen, welche Metriken in der wissenschaftlichen Literatur vorgestellt werden. Zum anderen zielt die Literaturrecherche darauf ab, eine qualita-

tive Bewertung einzelner Metriken durchzuführen. Konkret wird untersucht, wie gut die identifizierten Metriken geeignet sind, um die Datenqualität zu beurteilen. In der zweiten Phase findet die Festlegung und Dokumentation der Auswahlkriterien statt (Okoli 2015). Dies ist ein entscheidender Schritt einer systematischen Literaturrecherche, da dadurch die eigentliche Durchführbarkeit und darüber hinaus die Reproduzierbarkeit der Ergebnisse gewährleistet wird. Okoli (2015) benennt unter anderem den Inhalt, die Sprache, die Zugriffsmöglichkeit, die Art der Veröffentlichung, bestimmte Autoren, das Setting der Studien, das Studiendesign und das Veröffentlichungsdatum als Kriterien für die Vorauswahl. In dieser Arbeit sind zur Vorauswahl der Beiträge bestimmte Suchwörter, Veröffentlichungsdaten, Sprachen und Verfügbarkeiten zur Einschränkung bestimmt worden. Um sicherzustellen, dass keine potenziell hilfreichen Veröffentlichungen ausgeschlossen werden, erfolgt keine Einschränkung hinsichtlich bestimmter Autoren, Domänen oder Veröffentlichungstypen.

Der gesuchte Inhalt und die damit verbundenen *Suchwörter* werden aus den Zielen der Literaturrecherche abgeleitet. Die gewünschten Informationen und Schlüsselbegriffe können je nach den spezifischen Fragestellungen und Zielsetzungen der Forschung erheblich variieren. Da diese Masterarbeit nicht spezifische, sondern sämtliche Datenqualitätsmetriken im Kontext der Datenvorverarbeitung für produzierende Unternehmen untersucht und deren Anwendbarkeit bewertet, erfolgt zunächst eine umfassende Literaturrecherche. Dabei wird nach sämtlichen Publikationen gesucht, die Datenqualitätsmetriken behandeln. Um Publikationen zu finden, deren Fokus auf Datenqualitätsmetriken liegt, wird nach Arbeiten gesucht, deren Titel oder Abstract bestimmte Stichwörter bzw. Stichwortkombinationen enthält. Konkret wird nach Veröffentlichungen gesucht, deren Titel oder Abstract eine Kombination der Stichwörter *Datenqualitätsmetriken*, sowie der in Tabelle 1 vorgestellten *Datenqualitätsdimensionen* bzw. der englischsprachigen Entsprechungen enthält.

Um veraltete Ergebnisse auszuschließen und die Suche auf eine begrenzte Anzahl von Veröffentlichungen zu beschränken, werden zunächst nur Studien und Publikationen im *Zeitraum* der letzten zehn Jahre berücksichtigt. Die Suche gilt dabei als abgeschlossen, falls zehn relevante Quellen je Datenqualitätsdimension gefunden worden sind. Falls dies für diesen Zeitraum nicht der Fall ist, wird anschließend eine schrittweise Erweiterung der Suche in einzelnen

Jahren vorgenommen, um festzustellen, ob ältere Ergebnisse neue Erkenntnisse liefern können. Sollten diese zusätzlichen Suchergebnisse keine neuen Erkenntnisse bieten, wird die Suche und Auswertung an diesem Punkt abgeschlossen.

Zur Durchführung einer umfassenden Literaturrecherche wird Google Scholar als Datenbank für die Suche verwendet. Google Scholar bietet den Vorteil, dass es verschiedene online verfügbare Datenbanken durchsucht und somit Veröffentlichungen aus verschiedenen wissenschaftlichen Quellen zusammenfasst (Galvagno et al. 2014). Aus Gründen der Verständlichkeit und Zugänglichkeit werden jedoch nur Veröffentlichungen in deutscher und englischer *Sprache* herangezogen. Zudem müssen diese Veröffentlichungen entweder frei *verfügbar* sein oder über den institutionellen Zugang der Universitätsbibliothek Dortmund zugänglich sein. Im nächsten Schritt werden die ausgewählten Veröffentlichungen im Detail untersucht und analysiert, um den Inhalt und die Ergebnisse jeder einzelnen Arbeit zu erfassen und zusammenzuführen.

Die dritte Phase der systematischen Literaturrecherche umfasst die eigentliche Suche anhand der zuvor definierten Einschränkungen. Zunächst wird überprüft, ob die Veröffentlichungen inhaltlich relevant sind, indem geprüft wird, ob sie das Thema Datenqualitätsmetriken behandeln. Diese Überprüfung erfolgt anhand des Abstracts oder der Kurzfassung, sofern vorhanden. Falls kein Abstract verfügbar ist, wird die Einleitung und gegebenenfalls die Methodik der Studie auf relevante Informationen hin untersucht. Die inhaltliche Relevanz des Beitrags für die Arbeit wird somit sichergestellt. In dieser Arbeit wurden während der Recherche die gesammelten Informationen in einer Excel-Datei zusammengestellt. Auf diese Weise wurde jeder einzelne Beitrag geprüft, ob er zur Erreichung des Ziels der Arbeit berücksichtigt wird.

Die vierte Phase der systematischen Literaturrecherche umfasst die Verschriftlichung der vorhergegangenen Analyse. Diese Phase wird durch die Ergebnispräsentation in Abschnitt 3.1 repräsentiert.

3.1 Vorstellung der durch systematischen Literaturrecherche ermittelten Datenqualitätsmetriken

Im folgenden Abschnitt werden die Ergebnisse der systematischen Literaturrecherche beschrieben und die identifizierten Datenqualitätsmetriken der Datenqualitätsdimensionen aus Tabelle 1 in Abschnitt 2.6.1Tabelle 1: Auszug der Datenqualitätsdimensionen (i. A. an Sidi et al. 2012) vorgestellt. Mit

Hilfe der beschriebenen Vorgehensweise konnten insgesamt 238 Veröffentlichungen identifiziert werden, deren Inhalte den gestellten Ansprüchen erfüllen. Die Erwägung, den Betrachtungszeitraum schrittweise auszuweiten, um gegebenenfalls weitere Ergebnisse zu erhalten, war in einigen Fällen sinnvoll, um ältere Veröffentlichungen zu finden. Daher werden in Folgenden die Inhalte der 238 identifizierten Veröffentlichungen aus den Jahren 2008 bis 2023 behandelt. Dabei ist jedoch anzumerken, dass nicht alle Veröffentlichungen eine Berechnung der gesuchten Metrik vorschlagen. Zur übersichtlichen Darstellung und zur einfachen Auswertung wurden die untersuchten Inhalte der Veröffentlichungen in eine Literaturtabelle übertragen, die an dieser Stelle in Tabelle 2 auszugsweise und im Anhang vollständig abgebildet sind.

Tabelle 2: Auszug aus der Literaturtabelle der systematischen Literaturrecherche

ID	Dimension	Titel	Sprache	Jahr
48	Accessibility	An Advanced Big Data Quality Framework Based on Weighted Metrics	Englisch	2022
111	Accessibility	Data measurement in research information systems: metrics for the evaluation of data quality	Englisch	2018
205	Accessibility	Open Data Quality Dimensions and Metrics: State of the Art and Applied Use Cases	Englisch	2020
20	Accessibility	A Novel Data Quality Metric for Minimality	Englisch	2019
196	Accessibility	Methodology for linked enterprise data quality assessment through information visualizations	Englisch	2019
236	Accessibility	Review of data quality indicators and metrics, and suggestions for indicators and metrics for structural health monitoring	Englisch	2022
74	Accessibility	Big Data Quality Metrics for Sentiment Analysis Approaches	Englisch	2019
127	Accessibility	Daten- und Informationsqualität	Deutsch	2021

Im Folgenden werden die in der Literatur beschriebenen Datenqualitätsmetriken in alphabetischer Reihenfolge der Datenqualitätsdimensionen aus Tabelle 1 vorgestellt. Zuerst wird die Dimension *Accessibility* betrachtet. Hier sind zwei Datenqualitätsmetriken von signifikanter Ähnlichkeit identifiziert worden. Der erste Ansatz wird in Formel 1 von Elouataoui et al. (2022) beschrieben.

$$Accessibility\ (\%)\ = \frac{Anzahl\ der\ zugänglichen\ Werte}{Gesamtanzahl\ der\ Werte}$$

Formel 1: Metrik zur Berechnung der Dimension Accessibility (Elouataoui et al. 2022)

Dieser Ansatz bezieht sich auf die Anzahl der zugänglichen Werte im Verhältnis zu der Gesamtanzahl der Werte. Der zweite Ansatz ist dem vorherigen ähnlich, jedoch wird in der Berechnung anstatt der Anzahl zugänglicher Daten, die Anzahl Daten, die nicht verfügbar sind, betrachtet. Formel 2 repräsentiert den Ansatz, wie er in der Veröffentlichung von Azeroual et al. (2018) beschrieben wird.

$$Accessibility\ =\ 1 - \frac{Anzahl\ der\ Daten, die\ nicht\ verfügbar\ sind}{Gesamtanzahl\ der\ Daten}$$

Formel 2: Metrik zur Berechnung der Dimension Accessibility (Azeroual et al. 2018)

Über beide Datenqualitätsmetriken kann sichergestellt werden, dass Daten verfügbar und einfach abrufbar sind. Die Gewährleistung der Zugänglichkeit von Daten hat eine hohe Priorität, da nicht zugängliche Daten keinen Nutzen bringen. Daten sollten leicht zugänglich sein und effizient aus ihrem lokalen Speicherort abgerufen werden können, wenn sie für externe Verwendung verteilt werden (Azeroual et al. 2018).

Die nächste untersuchte Dimension behandelt die *Accuracy*. In der Literatur zur Datenqualität kann Genauigkeit als die Nähe zwischen einem Informationssystem und dem Teil der realen Welt beschrieben werden, den es modellieren soll (Batini und Scannapieco 2006).Aus der Perspektive der Naturwissenschaften wird Genauigkeit in der Regel als die *Größe eines Fehlers* definiert (Haegemans et al. 2016). Aufgrund dessen gibt es viele verschiedene Definition zur Berechnung der Datengenauigkeit. In Formel 3 wird der am häufigsten identifizierter Ansatz repräsentiert. Er setzt dabei die Anzahl richtiger Daten mit der Gesamtanzahl der Daten in Verhältnis.

$$Accuracy = \frac{Anzahl\ richtiger\ Daten}{Gesamtanzahl\ der\ Daten}$$

Formel 3: Metrik zur Berechnung der Accuracy in (Taleb et al. 2016; El Alaoui et al.; Heinrich und Klier 2015)

Neben diesem sehr allgemeinen Ansatz werden auch Datenqualitätsmetriken diskutiert, die die Genauigkeit auf Feld- (s. Formel 4) und Datensatzebene (s. Formel 5) definieren

$$Accuracy\ auf\ Feld - Ebene = \frac{Anzahl\ der\ als\ korrekt\ bewerteten\ Felder}{Anzahl\ der\ getesteten\ Felder}$$

Formel 4:Metrik zur Berechnung der Accuracy auf Feld-Ebene (Ehrlinger und Wöß 2022)

$$Accuracy\ auf\ Datensatz - Ebene = \frac{Anzahl\ der\ als\ vollständig\ korrekt\ bewerteten\ Datensätze}{Anzahl\ der\ getesteten\ Datensätze}$$

Formel 5:Metrik zur Berechnung der Accuracy auf Datensatz-Ebene (Ehrlinger und Wöß 2022)

Eine Verallgemeinerung durch den Austausch von *Felder* und *Datensätze* auf *Objekte* wird in der Literatur vorgeschlagen. Dieser Ansatz wird in Formel 6 dargestellt.

$$Accuracy\ auf\ Objekt - Ebene = \frac{Anzahl\ der\ als\ korrekt\ bewerteten\ Objekte}{Anzahl\ der\ getesteten\ Objekte}$$

Formel 6: Metrik zur Berechnung der Accuracy auf Objekt-Ebene (Ehrlinger und Wöß 2022)

Auch eine Datenqualitätsmetrik zur Messung der Accuracy, welche die inverse Nutzung, also das Verhältnis zwischen *Anzahl der Datenobjekte mit Fehlern* und der *Gesamtanzahl der Datenobjekte,* vorschlägt wurde in der Literatur identifiziert (s. Formel 7).

$$Accuracy\ auf\ Objekt - Ebene = 1 - \left(\frac{Anzahl\ der\ Datenobjekte\ mit\ Fehlern}{Gesamtanzahl\ der\ Datenobjekte}\right)$$

Formel 7: Metrik zur Berechnung der Accuracy auf Objekt-Ebene (Ehrlinger und Wöß 2022)

Ein alternativer Ansatz zur Messung der Genauigkeit sieht wie folgt aus: Auf der Ebene von Attribut-Wert wird die Metrik $Q_{Gen.}$durch das Verhältnis zwischen der Arity eines Werts und seiner optimalen Arity für numerische Werte definiert. Für ein numerisches Attribut A repräsentiert $s_{opt.}(A)$ die optimale Anzahl von Ziffern und Dezimalstellen für A. Dabei steht ω für einen Wert von A, und $s(\omega)$ gibt die tatsächliche Anzahl von Ziffern und Dezimalstellen für ω

im Attribut A an. Da $s_{opt.}(A)$ nicht zwangsläufig maximal ist, muss die Metrik auf den Bereich]0, 1] normalisiert werden.

$$Q_{Gen.}(\omega, \mathrm{A}) = min\left(\frac{s(\omega)}{s_{opt.}(A)}, 1\right)$$

Formel 8: Metrik zur Berechnung der Accuracy (Ehrlinger und Wöß 2022)

Die letzte identifizierte Metrik für die Dimension *Accuracy* behandelt nicht numerische Attribute. Dabei stehen $\mathrm{t}.A_1, \ldots, t.A_n$ für die Attributwerte der Attribute $A_1, \ldots, A_n$, die das beobachtete Tupel t spezifizieren. Der Faktor $g_j$gibt die relative Bedeutung von A_j im Verhältnis zum Gesamttupel an und wird vom Experten gewichtet. Die Genauigkeit auf Tabellenebene wird dann als arithmetisches Mittel der Genauigkeitsmessungen der Tupel berechnet, und die Genauigkeit auf Datenbankebene entspricht dem arithmetischen Mittel der Genauigkeitsmessungen auf Tabellenebene.

$$Q_{Gen.}(\mathrm{t}) = \frac{\sum_{j=1}^{n} Q_{Gen.}(\mathrm{t}.A_j, A_j) g_j}{\sum_{j=1}^{n} g_j}$$

Formel 9: Metrik zur Berechnung der Accuracy für nicht numerische Werte (Ehrlinger und Wöß 2022)

Für die Dimension *Appropriate amount of data* wurden in der Literatur ein Ansatz identifiziert. Ein Ansatz besteht darin, das Minimum zwischen dem Verhältnis der erforderlichen Daten zur verfügbaren Datenmenge und seinem Kehrwert zu bestimmen. Dieser Ansatz wird in Formel 10 beschrieben.

$$Appropriate\ amount\ of\ data = \min\left(\frac{erforderliche\ Daten}{verfügbare\ Daten}, \frac{verfügbare\ Daten}{erforderliche\ Daten}\right)$$

Formel 10: Metrik zur Berechnung der Appropriate amount of data (Makhoul 2022; Bonney et al. 2014)

Diese Metrik ermöglicht es, festzustellen, ob die vorhandenen Daten in ausreichender Menge vorliegen, um die gestellten Anforderungen zu erfüllen. Durch die Bestimmung des Minimums wird betont, dass die Datenmenge sowohl in Bezug auf ihre Notwendigkeit als auch ihre Verfügbarkeit gleichermaßen berücksichtigt wird.

Für die Dimension *Believability* wurde in der Literatur eine Metrik identifiziert (s. Formel 11).

$$Believability = \frac{\sum_{i=1}^{N} rate(i)}{Anzahl\ an\ Daten}$$

Formel 11: Metrik zur Berechnung der Believability (Frehe et al. 2016)

Die Glaubwürdigkeit eines Datenfeldes i wird durch die Funktion $rate(i)$ gemessen. Diese Funktion hat den Wert *1*, wenn das Datenfeld aus einer absolut glaubwürdigen Quelle stammt, und den Wert *0*, wenn es komplett unglaubwürdig ist (Frehe et al. 2016). Es können auch Zwischenwerte angenommen werden, abhängig von einer subjektiven Bewertung der Glaubwürdigkeit.

Die nächste untersuchte Datenqualitätsdimension befasst sich mit *Completeness.* In Folge der systematischen Literaturrecherche wurden verschiedene Berechnungsansätze gefunden und diese werden nun vorgestellt. Zum einen gibt es die sehr allgemein gehaltene Metrik zur Berechnung des Anteils der verfügbaren Datensätze (s. Formel 12). Diese Metrik repräsentiert dabei den Prozentsatz der verfügbaren Datensätze im Verhältnis zur Gesamtanzahl der Datensätze.

$$Anteil\ der\ verfügbaren\ Datensätze = \frac{Anzahl\ der\ verfügbaren\ Datensätze}{Gesamtzahl\ der\ Datensätze}$$

Formel 12: Metrik zur Berechnung der Completeness (Vetrò et al. 2016; Behkamal et al. 2014)

Ein weiterer Ansatz misst die Vollständigkeit einer Datenstruktur als Verhältnis der Anzahl der nicht leeren Eigenschaften zur Gesamtanzahl der Eigenschaften. Dieser Wert gibt Aufschluss darüber, wie gut die Datenstruktur mit Informationen gefüllt ist, wobei eine höhere Vollständigkeit Werte näher an *1* bedeutet (Behkamal et al. 2014). Niedrige Vollständigkeitswerte deuten darauf hin, dass viele Eigenschaften leer oder unvollständig sind (s. Formel 13).

$$Completeness = \frac{Anzahl\ der\ nicht\ leeren\ Werten}{Gesamtzahl\ der\ Werte}$$

Formel 13: Metrik zur Berechnung der Completeness (Elouataoui et al. 2022; Gitzel et al. 2016)

Die letzte Metrik für die Dimension *Completeness* ist ein inverser Ansatz (s. Formel 14).

$$Completeness = 1 - \frac{T_R}{N_R}$$

Formel 14: Metrik zur Berechnung der Completeness (Heinrich et al. 2018; Azeroual 2022; Yang et al. 2019; Frehe et al. 2016)

Hierbei steht T_R für die Anzahl der leeren Eigenschaften und N_R repräsentiert die Gesamtanzahl der Eigenschaften. Ein höherer Wert für $1 - \frac{T_R}{N_R}$ deutet auf eine höhere Vollständigkeit hin, während niedrigere Werte darauf hindeuten, dass viele Eigenschaften fehlende Daten enthalten.

In Bezug auf die Datenqualitätsdimension *Conciseness* werden verschiedene Metriken verwendet, um die Effizienz der Informationsrepräsentation in einem Datensatz zu bewerten. Die erste Metrik betrachtet das Verhältnis der Anzahl einzigartiger Objekte zu allen Objektrepräsentationen im Datensatz (s. Formel 15).

$$Conciseness = \frac{Anzahl\ der\ einzigartigen\ Eigenschaften}{Gesamtanzahl\ der\ Eigenschaften}$$

Formel 15: Metrik zur Berechnung der Conciseness (Moaawad et al. 2017; Zaveri et al. 2015)

Ein höherer Wert in dieser Metrik weist auf eine effiziente und prägnante Darstellung hin, da die Anzahl der einzigartigen Objekte im Verhältnis zur Gesamtanzahl der Objektrepräsentationen maximiert wird (Zaveri et al. 2015). Auch für diese Dimension wurde ein inverser Ansatz identifiziert. Dabei betrachtet diese Metrik die Klarheit der Informationsrepräsentation durch die Einhaltung der Eindeutigkeitsregel. Sie ermittelt den Anteil der Instanzen, die gegen die Regel verstoßen. Ein höherer Wert in dieser Metrik weist darauf hin, dass die Mehrheit der Instanzen eindeutig ist, was zu einer klareren Darstellung der Informationen führt (s. Formel 16) (Zaveri et al. 2015).

$$Conciseness = 1 - \left(\frac{Anzahl\ der\ einzigartigen\ Eigenschaften}{Gesamtanzahl\ der\ Eigenschaften}\right)$$

Formel 16: Metrik zur Berechnung der Conciseness (Zaveri et al. 2015)

Die dritte und letzte identifizierte Metrik zur Messung der Dimension *Conciseness* berücksichtigt die Handhabung von mehrdeutigen Instanzen. Sie misst den Anteil der Instanzen, die mehrdeutig sind, im Vergleich zur Gesamtanzahl der Instanzen im semantischen Metadaten-Set. Die Metrik wird in Formel 17 repräsentiert.

$$Conciseness = 1 - \left(\frac{Anzahl\ mehrdeutiger\ Instanzen}{Anzahl\ der\ Instanzen\ im\ semantischen\ Metadaten - Set}\right)$$

Formel 17: Metrik zur Berechnung der Conciseness (Zaveri et al. 2015)

Ein höherer Wert im Ergebnis dieser Metrik deutet darauf hin, dass der Datensatz weniger Unsicherheiten oder Mehrdeutigkeiten aufweist, was zu einer prägnanteren Darstellung von Informationen führt (Zaveri et al. 2015). Zusammenfassend ermöglichen diese Metriken eine umfassende Bewertung der Conciseness in einem Datensatz, wobei Aspekte wie Vielfalt der Objekte, Einhaltung der Eindeutigkeitsregel und der Umgang mit Mehrdeutigkeiten berücksichtigt werden.

Die nächste untersuchte Metrik befasst sich mit der Datenqualitätsdimension *Consistency*. Unter dem Konzept der Konsistenz wird die Eigenschaft der Widerspruchsfreiheit des Datenbestandes verstanden (Hildebrand et al. 2011). Auch hier werden in der Literatur verschiedene Ansätze zur Messung dieser Metrik vorgeschlagen. Der erste Ansatz (s. Formel 18) verfolgt die Berechnung über die *Anzahl der konsistenten Datensätze* im Verhältnis zur *Gesamtanzahl der Datensätze.* Der zweite Ansatz beschriebene Ansatz wiederum betrachtet die inverse Nutzung. Hier wird die Datenqualitätsmetrik über die *Anzahl Datensätze, die gegen die Konsistenz verstoßen* zur G*esamtanzahl der Datensätze* berechnet (s. Formel 19).

$$Anteil\ der\ konsistenten\ Datensätze = \frac{Anzahl\ der\ konsistenten\ Datensätze}{Gesamtanzahl\ der\ Datensätze}$$

Formel 18: Metrik zur Berechnung der Consistency (Heinrich und Klier 2015; Freudiger et al.; Taleb et al. 2016)

$$Anteil\ der\ konsistenten\ Datensätze = 1 - \frac{Anzahl\ Datensätze, die\ gegen\ die\ Konsistenz\ verstoßen}{Gesamtanzahl\ der\ Datensätze}$$

Formel 19: Metrik zur Berechnung der Consistency (Azeroual et al. 2018; Elouataoui et al. 2022)

Neben diesen zwei sehr einfach zu berechnenden Ansätzen werden in der Literatur auch komplexere Ansätze diskutiert. In einer identifizierten Konsistenzmetrik wird davon ausgegangen, dass Fachwissen in Regeln codiert ist und Widersprüche innerhalb der Regeln sowie unscharfe oder probabilistische Annahmen ausgeschlossen sind. Folglich wird die Konsistenz, eines Attributwerts w definiert (s. Formel 20).

$$Q_{Kon}(\omega) = \frac{1}{\sum_{j=1}^{n} r_j(\omega) g_j + 1}$$

Formel 20: Metrik zur Berechnung der Consistency (Ehrlinger und Wöß 2022)

Dabei steht g_j für den Schweregrad von $r_j(\omega)$, wobei $r_j(\omega)$ die Verletzung der Konsistenzregel r_j auf den Attributwert ω innerhalb eines Satzes von n Konsistenzregeln darstellt.

$$r_j(\omega) = \begin{cases} 0, falls\ \omega\ der\ Konsistenzregel\ r_s\ genügt \\ 1, \quad sonst \end{cases}$$

Formel 21: Berechnung der Konsistenzregel r_j (Ehrlinger und Wöß 2022)

Ein weiterer Ansatz, der sich auf die Konsistenzregel bezieht, ist in der Literaturrecherche identifiziert worden. Dieser wird durch folgende Formel 22 beschrieben:

$$Q_{Kons.}(\omega, \mathbb{R}) = \prod_{j=1}^{\mathbb{R}} (1 - r_j(\omega))$$

Formel 22: Metrik zur Berechnung der Consistency (Heinrich und Klier 2015)

Zur Berechnung des Ergebnisses wird noch die Bestimmung des Wertes $r_j(\omega)$ benötigt. Hierbei kann auf Formel 21 zurückgegriffen werden. Beide Ansätze nutzen dieselbe Berechnung der Konsistenzregel. Das Ergebnis der Metrik aus Formel 22 nimmt den Wert eins an, wenn der Attributwert alle in der Regelmenge $\mathbb{R}$ spezifizierten Konsistenzregeln erfüllt (das heißt $r_j(\omega)$ = 0) für alle $r_j(\omega) \in \mathbb{R}$). Im Gegensatz dazu ist der resultierende Wert der Metrik auf Attributwertebene null, wenn mindestens eine der spezifizierten Regeln verletzt ist (das heißt, es existiert ein $r_j(\omega) \in \mathbb{R}$, für das $r_j(\omega)$ = 1). Als Konsistenzregeln könnten dabei unter anderem formalisierte Geschäftsregeln oder domänenspezifische Funktionen dienen.

Ein letzter identifizierter Ansatz bezieht sich auf die Konsistenzregel in Formel 23. Aufbauend darauf wurde folgende Metrik identifiziert.

$$Consistency\ (t) = \sum_{r \in \mathbb{R}} \begin{cases} w^+(r), & falls\ t\ r\ erfüllt \\ w^-, & Wenn\ t\ r\ verletzt \\ w^0, & Wenn\ r\ nicht\ zutrifft, \end{cases}$$

Formel 23: Metrik zur Berechnung der Consistency (Heinrich et al. 2018; Alpar und Winkelsträter 2014)

Die Datenqualitätsdimension *Consistent Representation* adressiert die Heterogenität und Komplexität von Daten (s. Tabelle 1). Die in Formel 24 präsentierte Metrik für die Dimension *Consistent Representation* quantifiziert die Konsistenz der Daten anhand zuvor definierter Regeln. Dabei wird jedes Datenfeld j untersucht, und die Funktion $brokeRule(i)$ gibt an, ob die vordefinierten Regeln eingehalten werden (*Funktionswert 1*) oder nicht (*Funktionswert 0*)

(Frehe et al. 2016). Hierbei steht M für die Gesamtheit der Regeln für ein Datenfeld, und N repräsentiert die Anzahl der Datenfelder in einem Datensatz. Es ist jedoch zu beachten, dass diese Metrik die Definition von Regeln für die Überprüfung der Datensätze erfordert, weshalb sie kritisch betrachtet werden sollte.

$$Consistent\ Representation = \frac{\left(\frac{\sum_{j=1}^{N} Datenfeld_j (\sum_{i=1}^{M} brokeRule(i)}{M}\right)}{N}$$

Formel 24: Metrik zur Berechnung der Consistent Representation (Frehe et al. 2016)

Für die Dimension *Currency* sind in der Literatur verschiedene Ansätze beschrieben. Zum einen wird die *Curreny* über den Anteil der neuesten Datensätze berechnet, zum anderen über den Verfall des Datenwertes.

$$Anteil\ der\ neuen\ Datensätze = \frac{Anzahl\ der\ neusten\ Datensätze}{Gesamtanzahl\ der\ Datensätze}$$

Formel 25: Metrik zur Berechnung der Currency (Elouataoui et al. 2022)

Die in Formel 25 abgebildeter Darstellung der Berechnung gibt Aufschluss darüber, welcher Anteil der Datensätze als die neusten betrachtet werden. Dabei steht die *Anzahl der neuesten Datensätze* für die Menge der Datensätze, die als die aktuellen oder neuesten betrachtet werden. Hierzu muss ein Wert definiert werden, wie lange der Datensatz als neu betrachtet wird. Der zweite Ansatz berücksichtig das *Alter (ω, A)* des Attributwertes sowie der Verfallsrate *Verfall(A)*, welche den Anteil der Datenwerte des Attributs angibt, der durchschnittlich innerhalb einer Zeiteinheit inaktuell wird (s. Formel 26).

$$Q_{Akt.}(\omega, A) = \mathrm{e}^{(-Verfall(A)\ x\ Alter\ (\omega,A)}$$

Formel 26: Metrik zur Berechnung der Currency (Frehe et al. 2016; Hildebrand et al. 2021; Azeroual 2022)

Hierbei repräsentiert $Q_{Akt.}$ (ω, *A*) die Wahrscheinlichkeit, dass der vorliegende Attributwert ω noch den aktuellen Gegebenheiten entspricht. Die Annahme einer exponentialverteilten Gültigkeitsdauer der zugrunde liegenden Datenwerte mit dem Parameter Verfall(*A*) ist in diesem Kontext charakteristisch für die Lebensdauer eines Datenwerts und hat sich besonders im

Bereich der Qualitätssicherung bewährt (Hildebrand et al. 2021). Ein weiterer Ansatz, der sich mit der Dimension *Currency befasst, w*ird in der Abhandlung von Lidiansa (2014) beschrieben (s. Formel 27).

$$Currency = Zeit, in\ der\ Daten\ im\ System\ gespeichert\ werden - Zeit, in\ der\ Daten\ in\ der\ realen\ Welt\ aktualisiert\ werden$$

Formel 27: Metrik zur Berechnung der Currency (Lidiansa 2014)

Sie wird berechnet, indem die Zeit, in der Daten im System gespeichert werden, von der Zeit subtrahiert wird, in der die Daten in der realen Welt aktualisiert werden. Die resultierende Differenz gibt an, wie aktuell die im System gespeicherten Daten im Vergleich zur letzten Aktualisierung in der realen Welt sind. Weiterhin wird in der Abhandlung ein zusätzlicher Ansatz behandelt. Dies ist zugleich die letzte identifizierte Metrik zur Messung der *Currency.* Dieser Ansatz zur Messung der *Currency* wird durch die Formel 28 repräsentiert.

$$Currency = Alter + (Lieferzeit - Eingabezeit)$$

Formel 28: Metrik zur Berechnung der Currency (Lidiansa 2014)

Hierbei repräsentiert *Alte*r die vergangene Zeit seit einem bestimmten Ereignis, *Lieferzeit* die Dauer für die Bereitstellung von Informationen oder Ressourcen, und *Eingabezeit* den Zeitpunkt der Dateneingabe.

Die identifizierte Metrik für die Dimension *Data Coverage* dient dazu, den Umfang der Datenerfassung während eines definierten Bezugszeitraums zu quantifizieren (s. Formel 29).

$$Data\ Coverage = \frac{Anzahl\ der\ gültigen\ Messungen\ im\ relevanten\ Bezugszeitraum}{Gesamtanzahl\ der\ potenziellen\ Messungen\ im\ relevanten\ Bezugszeitraum}$$

Formel 29: Metrik zur Berechnung der Data Coverage (Brown und Woods 2014)

Eine weitere Metrik die durch Brown und Woods (2014) beschrieben ist wird in Formel 30 dargestellt.

$$Data\ coverage = t_v\ x\ d_c$$

Formel 30: Metrik zur Berechnung der Data Coverage (Brown und Woods 2014)

Hierbei steht $t_v$für die zeitliche Abdeckung und $d_c$für die Datenerfassung. Die Datenabdeckung wird also als Produkt der zeitlichen Abdeckung und der Datenerfassung berechnet. Diese Metrik gibt Aufschluss darüber, wie gut die verfügbaren Daten den zeitlichen Bereich abdecken und wie umfassend die Datenerfassung während dieses Zeitraums erfolgt. Ein höherer Wert in dieser Metrik deutet auf eine umfassendere und zuverlässigere Datenerfassung im Vergleich zur zeitlichen Abdeckung hin (Brown und Woods 2014).

Für die Datenqualitätsdimension *Data Decay* wird in der Literatur eine Metrik beschrieben, die das Verhältnis zwischen *Anzahl der Datensätze mit negativer Veränderung* und *Gesamtanzahl der Datensätze* aufweist. Diese *Negativitätsrate* gibt an welchem Prozentsatz der Datensätze eine negative Veränderung oder ein Rückgang im Vergleich zu vorherigen Werten aufweist.

$$Data\ Decay = \frac{Anzahl\ der\ Datensätze\ mit\ negativer\ Veränderung}{Gesamtanzahl\ der\ Datensätze}$$

Formel 31: Metrik zur Berechnung der Data Decay (Chernov 2022)

Die nächste, von Sidi et al. (2012), beschriebene Datenqualitätsdimension befasst sich mit der *Data specification.* Für diese Dimension wurde eine Metrik identifiziert, welche in Formel 32 beschrieben wird

$$Anteil\ der\ Datensätze, die\ der\ Spezifikation\ entsprechen = \frac{Anzahl\ der\ Datensätze, die\ der\ Spezifikation\ entsprechen}{Gesamtanzahl\ der\ Datensätze}$$

Formel 32: Metrik zur Berechnung der Data specification (Caballero et al. 2022; Pradhan und Tungal 2021)

Die folgende, in der Literatur beschriebene, Metrik fokussiert sich auf die Datenqualitätsdimension *Duplication*. Sie ist ein Maß für die Duplikationshäufigkeit in einem Datensatz und wird in Formel 33 beschrieben.

$$\text{Anteil der duplizierten Werte} = \frac{\text{Anzahl der duplizierten Datensätze}}{Gesamtanzahl\ der\ Datensätze}$$

Formel 33: Metrik zur Berechnung der Duplication (Chung et al. 2017; Taggart et al. 2015; Efimova et al. 2021; Saroja 2016)

Ein höherer Wert weist dabei darauf hin, dass es eine größere Anzahl von Duplikaten im Datensatz gibt, während ein niedrigerer Wert auf eine geringere Häufigkeit von Duplikaten hindeutet. Diese Metrik kann nützlich sein, um die Qualität und Konsistenz eines Datensatzes zu bewerten, insbesondere wenn Duplikate unerwünscht oder problematisch sind.

Eine weitere Datenqualitätsdimension, die durch Sidi et al. (2012) beschrieben wird, befasst sich mit der *Ease of Manipulation.* Die Messung dieser Metrik steht im Zusammenhang mit dem investierten Aufwand zur Vorbereitung von Daten für die Manipulation. Um diesen Aufwand zu quantifizieren, werden die Daten in ihren ursprünglichen Schemata mit den Daten nach der Vorverarbeitung verglichen. Somit wird die *Ease of Manipulation* als das Verhältnis der Unterschiede zwischen den Rohdaten und den vorverarbeiteten Daten zur Gesamtmenge der Daten definiert.

$$Ease\ of\ Manipulation = \frac{Anzahl\ der\ Unterschiede\ zwischen\ der\ Original - und\ der\ bereinigten\ Tabelle}{Gesamtdaten}$$

Formel 34: Metrik zur Berechnung der Ease of Manipulation (Elouataoui et al. 2022; Makhoul 2022)

Eine Metrik im Kontext der Datenqualitätsdimension wird zur Quantifizierung der Genauigkeit von Daten verwendet. Diese Metrik wird durch die Formel 35 ausgedrückt.

$$Free\ of\ error = 1 - \left(\frac{Anzahl\ fehlerhafter\ Daten}{Gesamtanzahl\ an\ Daten}\right)$$

Formel 35: Metrik zur Berechnung der Free of error (Gitzel et al. 2016; Makhoul 2022; Gitzel et al.; Xu et al. 2022; Kim und Lee 2015; Gitzel et al. 2018; Günther et al. 2019; Laranjeiro et al. 2015)

Diese Metrik hat das Hauptziel, den Grad der Fehlerfreiheit in den vorliegenden Daten zu erfassen. Die Metrik erfasst, in welchem Maße die vorhandenen Daten frei von Fehlern sind. Ein höherer Wert in der *Free of error*-Metrik deutet darauf hin, dass ein geringerer Anteil der Daten fehlerhaft ist im Vergleich zur Gesamtanzahl der Daten. Die Formel 35 ermöglicht eine quantitative Berechnung dieser Metrik. Sie kann dazu verwendet werden, die Qualität der Daten im Hinblick auf ihre Genauigkeit zu bewerten.

Für die Dimension *Freshness* wurden durch die systematischen Literaturrecherche verschieden Ansätze identifiziert. Die Metrik wird teilweise als identisch betrachtet mit der Dimension *Currency* sowie der später beschriebenen Dimension *Timeliness*. Daher wird in der Literatur

auch die Metrik aus der Formel 25 für die Dimension *Freshness* verwendet. Neben diesem bereits bekannten Ansatz ist noch ein anderer Ansatz identifiziert worden (s. Formel 36).

$$Freshness = ts_{cur} - ts_t$$

Formel 36: Metrik zur Berechnung der Freshness (Ehrlinger und Wöß 2022)

In dieser Metrik repräsentiert ts_{cur} den aktuellen Zeitstempel. Die Variable ts_t repräsentiert den Zeitstempel der letzten Aktualisierung aus der Zielanfrage. Diese Metrik quantifiziert den zeitlichen Unterschied zwischen dem aktuellen Zeitpunkt und dem Zeitpunkt der letzten Aktualisierung der Daten im Zielsystem.

Die Messung der Datenqualitätsdimension *Relevancy* ist äußerst kontextanhängig und hängt vom beabsichtigten Verwendungszweck der Daten ab. In der Literatur wurden daher verschiedene Messgrößen definiert. In der folgenden Formel wird die Datenrelevanz mit der Anzahl der Datenzugriffe verknüpft, wodurch die am häufigsten abgerufenen Daten als die relevantesten betrachtet werden (s. Formel 37).

$$Relevantestes\ Datenfeld\ f = \frac{Anzahl\ der\ Zugriffe\ auf\ f}{Gesamtzugriffe\ auf\ die\ Tabelle, die\ F\ enthält}$$

Formel 37: Metrik zur Berechnung der Relevancy (Elouataoui et al. 2022)

Neben diesem Ansatz ist durch die systematische Literaturrecherche ein weiterer Ansatz zur Berechnung der *Relevancy* identifiziert worden. Dieser allgemein gehaltene Ansatz beruht auf der subjektiven Bewertung der *Relevancy* der Daten (s. Formel 38).

$$Relevancy = \frac{Anzahl\ der\ relevanten\ Daten}{Gesamtanzahl\ der\ Daten}$$

Formel 38: Metrik zur Berechnung der Relevancy (Hassine und Clément 2020; Hejazi et al. 2017; Makhoul 2022; Raghavendra 2017; Fadlallah et al. 2023)

Um dies umzusetzen, müssen klare Kriterien festgelegt werden, um zu bestimmen, welche Daten als relevant für den spezifischen Kontext oder das Ziel gelten. Diese Kriterien können beispielsweise durch Fachexperten, Stakeholder oder vordefinierte Standards definiert werden.

Eine weitere Dimension, die im Rahmen der Literaturrecherche betrachtet wurde und für die eine Metrik identifiziert wurde, ist die *Reliability*. In der untersuchten Literatur wird eine Metrik,

basierend auf der Beantwortung von n gleich wichtigen Fragen, die die Zuverlässigkeit eines Datensatzes bewerten, vorgeschlagen. Jede Antwort wird durch eine spezielle Art von unscharfer Zahl repräsentiert, die drei Werte enthält $a_{1i} = s_i c_i$, $a_{2i} = s_i$ und $a_{1i} = s_i c_i + 1$, wobei s_i der Zufriedenheitsgrad und c_i der Gewissheitsgrad sind (Heinrich et al. 2018). Die Gesamtzuverlässigkeit des Datensatzes wird durch die Summe dieser unscharfen Zahlen für alle Fragen gebildet (s. Formel 39).

$$Reliability = \sum_{i=1}^{n} Q_i$$

Formel 39: Metrik zur Berechnung der Reliability (Heinrich et al. 2018)

Um diese Gesamtzuverlässigkeit zu bewerten, wird die Zentroidenmethode verwendet, um eine defuzzifizierte Zuverlässigkeit zu berechnen (Heinrich et al. 2018):

$$Defuzzifzierte\ Reliability = \sum_{i=1}^{n} \frac{a_{1i} + a_{2i} + a_{3i}}{3}$$

Formel 40: Metrik zu Berechnung der defuzzifzierte Reliability (Heinrich et al. 2018)

Diese Bewertung ermöglicht die Zuordnung zu unterschiedlichen Zuverlässigkeitsniveaus.

Im Rahmen der systematischen Literaturrecherche wird die Datenqualitätsdimension *Security* betrachtet. Für diese Dimension sind zwei Ansätze identifizier worden. Die Sicherheit von Daten bezieht sich auf die angemessene Einschränkung des Datenzugriffs. Angesichts zunehmender Datenschutzverletzungen und Sicherheitsangriffe hat die Gewährleistung von Datenvertraulichkeit und -sicherheit erhebliche Priorität erlangt (Elouataoui et al. 2022). In der Literatur werden zur Bewertung dieser Dimension gewichtete Leitfragen aufgestellt. Über die Beantwortung der Fragen kann die Datenqualitätsmetrik errechnet werden (s. Formel 41).

$$Security = \sum_{i=1}^{5} 0{,}2\ x\ Frage\ i$$

Formel 41: Metrik zur Berechnung der Security (Elouataoui et al. 2022)

Folgende Leitfragen müssen beantwortet werden mit einem Wert zwischen 0 und 1.

Frage 1: Existiert eine Sicherheitsrichtlinie, die die Nutzung der Daten einschränkt?

Frage 2: Werden Sicherheitsprotokolle für die Datenübertragung verwendet?

Frage 3: Gibt es effektive Maßnahmen zur Bedrohungserkennung?

Frage 4: Sind die Daten angemessen verschlüsselt?

Frage 5: Liegt den Daten eine Sicherheitsdokumentation bei?

Eine weitere identifizierte Metrik zur Bewertung der *Security* umfasst ebenfalls ein Punktebewertungssystem. Diese Sicherheitsmetrik integriert verschiedene Funktionen, die im Rahmen der Analysephase durchgeführt werden, um das Sicherheitsniveau zu beurteilen. Die Punktevergabe erstreckt sich über verschiedene Sicherheitsstufen, einschließlich Authentifizierung, Vertraulichkeit, Integrität und Datenschutz. Beispielsweise erhält eine digitale Signatur die maximale Punktzahl von 1, während bei einer NULL-Authentifizierungszeichenkette die Punktzahl bei 0 liegt. Schwächere Authentifizierungsmethoden wie kurze Passwörter werden mit einer Punktzahl von 0,2 bewertet. Diese Metrik gewährt einen ganzheitlichen Überblick über verschiedene Sicherheitsaspekte und ordnet den Datenpunkten entsprechende Bewertungen zu, um ein umfassendes Verständnis für das Sicherheitsniveau zu erlangen (Sicari et al. 2016; Goknil et al. 2023).

Für die Dimension *Timeliness* werden in der Literatur verschiedene Ansätze für die Berechnung des Grades der Aktualität angegeben. Die zeitlich begrenzte Herangehensweise geht davon aus, dass die Werte eines Attributes nach einer festgelegten Zeitspanne zwangsläufig ungültig werden und den Verfall der Daten bis zu diesem Zeitpunkt modelliert. Die Dimension der Aktualität wird durch dem $\mathrm{Alterswert\ einer\ Datenkomponente}$ und die $\mathrm{Haltbarkeit}$, in der die erfassten Daten gültig bleiben, quantifiziert. Die Quantifizierung erfolgt durch die Formel 42:

$$\mathrm{Timeliness} = max\left[\left(1 - \frac{\mathrm{Alterswert\ einer\ Datenkomponente}}{\mathrm{Haltbarkeit}}\right), 0\right]^{s}$$

Formel 42: Metrik zur Berechnung der Timeliness (Makhoul 2022; Rapp 2020)

Der Parameter *s* passt die Sensibilität an, mit der auf das Verhältnis $\frac{\mathrm{Alterswert\ einer\ Datenkomponente}}{\mathrm{Haltbarkeit}}$ reagiert wird, an den Kontext an. Die Variable $\mathrm{Haltbarkeit}$ muss ebenfalls an den Kontext angepasst werden, wobei ein niedriger Wert für Attribute mit häufigen Änderungen und ein hoher Wert für Attribute mit seltenen Änderungen verwendet wird. Neben

diesem Ansatz gibt es noch einen weiteren. Bei diesem wird der Anteil der rechtzeitig erhaltenden Daten in ein Verhältnis zu der Gesamtzahl der erhaltenden Daten gesetzt.

Ein weiterer Ansatz der häufig im Zusammenhang mit der Dimension *Timeliness* beschrieben wird, ist identisch mit dem Ansatz aus Formel 26 (Azeroual et al. 2018; Ehrlinger und Wöß 2022; Hildebrand et al. 2021; Baier 2020; Azeroual 2022). Daher wird dieser hier nicht nochmal aufgeführt.

Zur übersichtlichen Darstellung und zur einfachen Auswertung wurden die untersuchten Inhalte der Veröffentlichungen in eine Tabelle übertragen, die im Anhang zu finden ist. In Tabelle 3 ist ein Auszug zu finden.

Tabelle 3: Auszug aus Metriktabelle der systematischen Literaturrecherche

Dimension	Metrik
Accessibility	$Accessibility\ (\%) = \frac{Anzahl\ der\ zugänglichen\ Werte}{Gesamtanzahl\ der\ Werte}$
Accessibility	$Accessibility = 1 - \frac{Anzahl\ der\ Daten, die\ nicht\ verfügbar\ sind}{Gesamtanzahl\ der\ Daten}$
Accuracy	$Accuracy = \frac{Anzahl\ richtiger\ Daten}{Gesamtanzahl\ der\ Daten}$
Accuracy	$Accuracy\ auf\ Feld - Ebene = \frac{Anzahl\ der\ als\ korrekt\ bewerteten\ Felder}{Anzahl\ der\ getesteten\ Felder}$

Nachdem nun alle identifizieren Datenqualitätsmetriken vorgestellt worden sind, werden im Folgenden alle Datenqualitätsdimensionen aufgezählt, für die keine Datenqualitätsmetriken mit der systematischen Literaturrecherche gefunden worden sind:

1. Amount of data
2. Availability

3. Consistency and Synchronization
4. Data integrity fundamentals
5. Ease of Use and mantainability
6. Effectiveness
7. Efficiency
8. Interpretability
9. Lernability
10. Navigation
11. Objectively / Objectivtiy
12. Presentation Quality
13. Reputation
14. Safety
15. Timeliness and Availability
16. Transactability
17. Understandability
18. Useability
19. Useful
20. Value Added

Alle identifizierten Datenqualitätsmetriken aus der systematischen Literaturrecherche wurden vorgestellt. Im weiteren Verlauf erfolgt die Bewertung dieser Metriken im Kontext produzierender Unternehmen.

3.2 Bewertung von Datenqualitätsmetriken im Kontext von produzierenden Unternehmen

Im Folgenden werden die in Abschnitt 0 identifizierten Datenqualitätsmetriken für ihre Nutzung in produzierenden Unternehmen bewertet. Hierfür wird das, in Abschnitt 2.6, beschrieben Modell, repräsentier in Abbildung 7 genutzt. Es wird sich auf die dort beschriebene Datenqualitätsdimensionen konzentriert. Zuerst werden die Datenqualitätsmetriken für die Charakteristika *Velocity* bewertet. Die hierfür beschriebenen Dimensionen sind folgende: *Timeliness, Currency, Editability* und *Understandability.*

Die beiden vorgestellten Metriken für die Dimension *Timeliness* weisen unterschiedliche Formeln auf und sollten entsprechend ihrer Komplexität, Interpretation, Anwendungsbereiche und Datenanforderungen bewertet werden. Die Formel 42 könnte in produzierenden Unternehmen, insbesondere im Kontext von Lagerverwaltung, Produktionsprozessen oder Qualitätssicherung, von Bedeutung sein, wie von Hildebrand et al. (2021b) beschrieben. Diese Formel ermöglicht die Bewertung der Aktualität von Daten, wobei der Parameter *s* die Sensitivität gegenüber Änderungen steuert. Jedoch bringt diese Formel einige Herausforderungen mit sich, wie von Frehe et al. (2016) erwähnt wird. Die klare und konsistente Definition von Alterswert und Haltbarkeit ist entscheidend, was für produzierende Unternehmen möglicherweise eine zusätzliche Komplexität bedeutet. Die Festlegung der erwarteten Haltbarkeit kann nur subjektiv bestimmt werden und je nach Datenart und Prozess variieren. Daher ist es wichtig, die Formel an die spezifischen Anforderungen und Definitionen des Unternehmens anzupassen. Die Echtzeitanforderungen und die Dynamik der Produktionsumgebung müssen ebenfalls berücksichtigt werden. Im Gegensatz dazu bietet die Formel 26 eine einfachere Alternative. Durch den Exponentialwert nimmt der Datenwert automatisch ab, je älter die Daten sind, ohne dass ein zusätzlicher Parameter wie *s* bestimmt werden muss. Dies deutete darauf hin, dass der Aktualitätsgrad exponentiell mit dem Produkt aus Verfall und Alter abnimmt. Dies ist konsistent mit dem Konzept, dass ältere Daten weniger Einfluss auf die Aktualität haben sollten (Zaveri et al. 2015). Hierdurch könnte die Implementierung in produzierenden Unternehmen erleichtern werden, insbesondere wenn die klare Definition von Alterswert und Haltbarkeit eine Herausforderung darstellt. Insgesamt sollte die Auswahl zwischen den beiden Formeln auf den spezifischen Kontext und die Anforderungen des produzierenden Unternehmens abgestimmt werden. Während Formel 42 detailliertere Anpassungsmöglichkeiten bietet, kann Formel 26 aufgrund ihrer Einfachheit und Automatisierungsvorteile in Bezug auf die Abnahme des Datenwerts im Alter eine praktikablere Lösung darstellen. Daher wird empfohlen die Datenqualitätsmetrik aus Formel 26Formel 25 zu verwenden.

Neben der Dimension *Timeliness* gibt es noch die Dimension der *Currency* zur Bewertung der *Velocity*. Die Betrachtung der Dimension Currency (Aktualität) für produzierende Unternehmen schließt die bereits in der vorherigen Diskussion zur Dimension Timeliness empfohlen Formel 26Formel 25, aus.Die Fokussierung liegt nun auf den verbleibenden Formeln. Die Formel 25 bietet eine einfache Metrik, indem sie den prozentualen Anteil neuer Datensätze betrachtet.

Diese Metrik könnte in der Produktion nützlich sein, um eine grobe Einschätzung der Aktualität des Gesamtdatensatzes zu erhalten. Insbesondere in schnelllebigen Produktionsumgebungen, in denen Echtzeitinformationen entscheidend sind (Hoffmann 2017), kann der Anteil der neuen Datensätze als Indikator für die Frische der vorliegenden Daten dienen. Durch ihre Einfachheit ist diese Formel leicht verständlich und kann schnell in Entscheidungsprozessen integriert werden, ohne komplexe Berechnungen oder umfangreiche Datenanalysen durchführen zu müssen. Formel 27 konzentriert sich auf die zeitliche Differenz zwischen der Speicherung von Daten im System und dem Zeitpunkt ihrer Aktualisierung in der realen Welt. Diese Herangehensweise bietet einen Einblick in die Verzögerungen, die zwischen der Erfassung von Daten im System und ihrer tatsächlichen Veränderung in der physischen Welt auftreten können. Diese Perspektive auf die Datenaktualität könnte insbesondere in Produktionsprozessen von Bedeutung sein, in denen präzise und zeitnahe Informationen entscheidend sind (Hoffmann 2017). Die Relevanz dieser Formel in der Produktion liegt darin, dass sie aufzeigt, wie lange es dauert, bis Änderungen, die in der realen Welt auftreten, im System reflektiert werden. Dies ist wichtig, um sicherzustellen, dass die im System vorhandenen Daten den aktuellen Zustand der physischen Umgebung widerspiegeln. Formel 28 dressiert die Aktualität von Daten durch die Integration von Liefer- und Eingabezeiten als zentrale Faktoren. In produktionsbezogenen Umgebungen kann dies von erheblicher Relevanz sein, insbesondere wenn Daten aus externen Quellen stammen, um sicherzustellen, dass die vorliegenden Informationen zeitnah und korrekt sind (Hoffmann 2017). Die Berücksichtigung der Lieferzeit ist insbesondere in Branchen mit komplexen Lieferketten wichtig (Windelband et al. 2011). Hierbei geht es darum, wie schnell Daten nach ihrer Lieferung in das System integriert und verfügbar gemacht werden. Dies könnte Einfluss auf verschiedene Aspekte der Produktion haben, wie beispielsweise die Planung von Ressourcen, Lagerbeständen oder Produktionsprozessen. Die Eingabezeit hingegen betrifft den Zeitpunkt, zu dem Daten in das System eingegeben oder aktualisiert werden. In Produktionsumgebungen kann dies verschiedene Bereiche abdecken, von der Erfassung von Sensorwerten bis hin zur manuellen Eingabe von Produktionsdaten. Die zeitnahe und präzise Erfassung dieser Eingaben ist entscheidend, um sicherzustellen, dass die Daten im System aktuell und verlässlich sind (Kletti et al. 2015). In der abschließenden Betrachtung wird deutlich, dass die Wahl der geeigneten Formel für die Bewertung der Aktualität von Daten in produzierenden Unternehmen von verschiedenen Faktoren abhängt. Insbesondere spielen die spezifischen Anforderungen der Produktionsumgebung sowie

die praktische Umsetzbarkeit der gewählten Metrik eine entscheidende Rolle. Die Formel 27 und Formel 28, die einen detaillierteren Ansatz verfolgen, indem sie Liefer- und Eingabezeiten bzw. die Differenz zwischen der System- und Realzeit berücksichtigen, bieten eine feinere Granularität bei der Beurteilung der Datenaktualität. Diese detaillierten Informationen können in komplexen Produktionsprozessen von entscheidender Bedeutung sein, um Verzögerungen genau zu erfassen und daraus resultierende Auswirkungen auf die Entscheidungsfindung zu verstehen. Allerdings gehen mit dieser Detailliertheit auch Herausforderungen einher, insbesondere hinsichtlich der präzisen Erfassung von Zeitangaben (Kletti et al. 2015). Die Komplexität der Formeln könnte zudem die Implementierung erschweren, da klare Definitionen und Standards für Parameter wie Lieferzeit und Eingabezeit erforderlich sind. Dies könnte zusätzlichen Aufwand bei der Integration unterschiedlicher Datenquellen und -systeme bedeuten. Formel 25, die den Anteil der neuen Datensätze betrachtet, zeichnet sich durch ihre Einfachheit aus. Sie bietet eine schnelle und leicht verständliche Metrik für die Aktualität von Daten, indem sie den prozentualen Anteil neuer Datensätze im Gesamtdatensatz berechnet. Dies könnte in der Praxis besonders nützlich sein, wenn eine schnelle Einschätzung der Gesamtdatenaktualität erforderlich ist, ohne sich auf komplexe zeitliche Aspekte einzulassen. Neben den Dimensionen *Currency* und *Timeliness* wird im Modell von Frehe et al. (2016) die Dimension *Editability* und *Understandability* benannt. Für diese Dimensionen wurden in der systematischen Literaturrecherche keine Datenqualitätsmetriken identifiziert. Daher ist hier auch keine Bewertung möglich.

Ein weiteres Charakteristikum der 5Vs ist *Volume.* In der Abhandlung von Frehe et al. (2016) werden die Dimensionen *Appropiate amount of data, Clarity* und *Security* genannt. Die in Formel 10 beschriebene Datenqualitätsmetrik für die Dimension *Appropiate amount of data* betrachtet das Verhältnis zwischen den benötigten und den verfügbaren Daten und wählt das Minimum zwischen den beiden Quotienten. Dieser Ansatz zielt darauf ab, den Engpass oder das kritischste Verhältnis zwischen den erforderlichen und verfügbaren Daten zu identifizieren (Lidiansa 2014). In Bezug auf produzierende Daten zeigt diese Metrik auf, ob ausreichend Daten vorhanden sind, um den reibungslosen Ablauf von Produktionsprozessen zu gewährleisten. Diese könnte beispielweise die Überwachung der Anlagenleistung, der Qualitätssicherung oder andere prozessrelevante Aspekte umfassen. Allerdings ist die Anwendbarkeit dieser Metrik stark von der Anzahl Daten abhängig. Die Metrik kann sehr empfindlich auf kleine

Datenmengen reagieren. In Situationen, in denen die verfügbaren Daten begrenzt sind, könnte die Anwendung der Metrik zu unrealistischen oder unzuverlässigen Ergebnissen führen. Daher ist es wichtig, die Metrik in Umgebungen mit ausreichender Datenbasis anzuwenden, um eine angemessene Aussagekraft zu gewährleisten. Insgesamt bietet die identifizierte Metrik für die Dimension *Appropriate Amount of Data* einen einfachen, dennoch wirkungsvollen Ansatz, um die Datenlage in der Produktion zu bewerten. Durch ihre Anwendung kann sie dazu beitragen, Engpässe und kritische Punkte in Bezug auf Datenverfügbarkeit und -qualität zu identifizieren, was wiederum die Grundlage für gezielte Verbesserungsmaßnahmen in produktionsbezogenen Prozessen legt.

Für die Datenqualitätsdimension *Security* sind zwei Ansätze identifiziert worden, die mit einem Punktesystem arbeiten. Die erste identifizierte Metrik, Formel 41, basiert auf einem klaren Punktesystem, das durch die Beantwortung von fünf spezifischen Leitfragen generiert wird. Diese Fragen decken wichtige Sicherheitsaspekte ab, von der Existenz einer Sicherheitsrichtlinie bis zu dem Vorhandensein von Sicherheitsprotokollen und Datenverschlüsselung (Elouataoui et al. 2022). Die Gesamtpunktzahl, errechnet durch die Summe der gewichteten Antworten auf diese Fragen, bietet eine schnelle und transparente Einschätzung der Datensicherheit. Eine der Stärken der in Formel 41 beschriebenen Datenqualitätsmetrik liegt in ihrer Einfachheit und der klaren Struktur. Die gleichmäßige Gewichtung jeder Frage mit 0,2 unterstreicht die Gleichberechtigung verschiedener Sicherheitsaspekte. Dies macht die Metrik besonders zugänglich und anwendbar, insbesondere in produzierenden Unternehmen, die eine rasche und klare Beurteilung der Sicherheit ihrer Daten benötigen. Weiterhin bietet die Metrik eine Grundlage für die Priorisierung von Sicherheitsmaßnahmen. Durch die individuelle Gewichtung und Bewertung jeder der fünf Fragen können Unternehmen identifizieren, welche Sicherheitsaspekte dringend verbessert werden müssen. Dies ermöglicht eine gezielte Ressourcenallokation, um die Schwachstellen zu adressieren und die Sicherheitsmaßnahmen entsprechend zu priorisieren. Die Metrik dient somit nicht nur als Bewertungsinstrument, sondern auch als strategische Orientierungshilfe für die effektive Stärkung der Datensicherheit in produktiven Umgebungen. Die zweite identifizierte Metrik geht einen detaillierteren Weg und integriert ein Punktebewertungssystem für verschiedene Sicherheitsaspekte. Diese Metrik ermöglicht eine umfassendere Analyse von Authentifizierung, Vertraulichkeit, Integrität und Da-

tenschutz (Sicari et al. 2016). Die differenzierte Punktevergabe gestattet eine präzisere Identifikation von Schwachstellen in der Sicherheitslandschaft. Insbesondere in produzierenden Unternehmen, wo die Sicherheit sensibler Daten von entscheidender Bedeutung ist (Windthorst 2020), könnte diese Datenqualitätsmetrik durch ihre detailliertere Bewertung wertvoll sein. Die Berücksichtigung verschiedener Sicherheitsstufen erlaubt eine gezielte Priorisierung und Investition in spezifische Sicherheitsaspekte (Sicari et al. 2016). Die Wahl zwischen diesen beiden Metriken hängt von den spezifischen Anforderungen und dem gewünschten Detailgrad der Sicherheitsbewertung ab. Der erste Ansatz bietet eine schnelle und einfache Möglichkeit für eine grundlegende Einschätzung, während der zweite Ansatz eine tiefere Analyse ermöglicht. Eine kombinierte Anwendung beider Metriken könnte für produzierende Unternehmen eine sinnvolle Strategie darstellen, um sowohl eine schnelle Übersicht als auch eine detaillierte Bewertung der Datensicherheit zu gewährleisten.

Die Datenqualitätsdimension *Clarity* wurde in der systematischen Literaturrecherche nicht betrachtet. Daher ist für diese Dimension keine Metrik identifiziert worden.

Für die Charakteristika *Veracity* werden im Modell von Frehe et al. (2016) die Datenqualitätsdimensionen *Free of error, Believability, Completeness, Reputation und Uniqueness* identifiziert.

Die identifizierte Metrik für die Dimension *Free of error,* repräsentiert in Formel 35 bietet eine prägnante und gut nachvollziehbare Methode, die Fehlerfreiheit von Daten zu bewerten. Ein positiver Aspekt dieser Metrik liegt in ihrer Einfachheit und Transparenz. Die Formel nutzt eine bekannte und leicht interpretierbare Struktur, die es ermöglicht, schnell eine Einschätzung der Fehlerfreiheit zu erhalten. Durch den Ausdruck als Prozentsatz wird die Metrik intuitiv verständlich, wodurch sie in verschiedenen Kontexten leicht anwendbar ist. Im produzierenden Umfeld, wo die Genauigkeit und Qualität von Daten eine entscheidende Rolle spielt (Hudasch 1996), bietet diese Metrik die Möglichkeit, systematische Muster und Trends in den auftretenden Fehlern zu identifizieren. Durch eine detaillierte Analyse der Fehlerquote in verschiedenen Teilen der Daten können Unternehmen potenzielle Schwachstellen oder wiederkehrende Fehlerquellen besser verstehen. Dies ermöglicht eine gezielte Fehlerprävention und Qualitätsverbesserung in den relevanten Produktionsprozessen. Allerdings gibt es auch Limitationen. Die Metrik berücksichtigt nicht die Schwere der Fehler oder potenzielle Auswirkungen auf Ge-

schäftsprozesse. Ein einzelner schwerwiegender Fehler wird in dieser Metrik genauso gewichtet wie mehrere geringfügige Fehler. Daher eignet sich diese Metrik besonders für Szenarien, in denen eine schnelle, oberflächliche Bewertung der allgemeinen Fehlerfreiheit ausreichend ist. Zusammenfassend bietet die identifizierte Datenqualitätsmetrik für die Dimension *Free of Error* eine einfache und zugängliche Möglichkeit, die Fehlerfreiheit von Daten zu beurteilen. Ihre Anwendung könnte in produzierenden Unternehmen als schnelle, erste Einschätzung nützlich sein, insbesondere wenn die Gesamtanzahl der Daten und die Gesamtfehleranzahl gut definiert und verständlich sind.

Die identifizierte Datenqualitätsmetrik für die Dimension *Believability,* durch Formel 11 beschrieben, bietet einen Ansatz, um die Glaubwürdigkeit von Daten zu bewerten. Ein zentraler Aspekt dieser Metrik liegt in der individuellen Bewertung der Glaubwürdigkeit einzelner Datenfelder durch die Funktion $rate(i)$. Diese Funktion ermöglicht eine feingranulare Analyse, indem sie Werte zwischen 0 und 1 annimmt, abhängig von der subjektiven Einschätzung der Glaubwürdigkeit eines jeden Datenfeldes (Frehe et al. 2016). Es ist jedoch zu beachten, dass die Einschätzung aufgrund ihrer Subjektivität zu unterschiedlichen Interpretationen führen kann. Es besteht die Möglichkeit, dass verschiedene Personen die Glaubwürdigkeit eines Datenfeldes unterschiedlich beurteilen (Teuteberg und Freundlieb 2009). Im Produktionskontext kann diese Metrik wertvoll sein, indem sie einen detaillierten Einblick in die Glaubwürdigkeit von Daten gewährt. Dies beruht auf der Möglichkeit jedes Datenfeld individuell zu bewerten. Dadurch ist es Unternehmen möglich, gezielt Schwachstellen in Bezug auf Glaubwürdigkeit zu erkennen und Maßnahmen zur Verbesserung der Datenqualität zu ergreifen (Teuteberg und Freundlieb 2009). Zusammenfassend ermöglicht die identifizierte Datenqualitätsmetrik für die Dimension *Believability* eine präzise Bewertung der Glaubwürdigkeit von Daten auf individueller Feldbasis. Diese Granularität erlaubt eine gezielte Identifizierung von Schwachstellen in Bezug auf die Glaubwürdigkeit, was wiederum Unternehmen befähigt, spezifische Maßnahmen zur Verbesserung der Datenqualität zu ergreifen. Somit fungiert die Metrik nicht nur als Bewertungsinstrument, sondern auch als praktisches Werkzeug zur strategischen Optimierung der Datenqualität.

Für die Dimension *Completeness* sind drei verschiedene Datenqualitätsmetriken identifiziert worden. Im Kontext produzierender Unternehmen spielt die Vollständigkeit, eine entscheidende Rolle bei der Gewährleistung effizienter und präziser Geschäftsprozesse (Balzer et al. 2020). Formel 12 präsentiert eine anwenderfreundliche Metrik, die den prozentualen Anteil der

vorhandenen Datensätze angibt. In produktiven Umgebungen kann diese Formel als schnelle Indikation für die Datenverfügbarkeit dienen, doch sie gibt keine detaillierte Information über die Vollständigkeit einzelner Datensätze. Formel 13 fokussiert sich auf die Anwesenheit von Werten in einem Datensatz und kann in produzierenden Unternehmen nützlich sein, um leere oder fehlende Werte zu identifizieren. Formel 14 bietet die umgekehrte Perspektive auf die Vollständigkeit. Sie zeigt auf, wie viele Datensätze fehlen. In industriellen Umgebungen weist diese Metrik darauf hin, in welchen Bereichen Verbesserungen erforderlich sind. Jedoch berücksichtigen weder Formel 13 noch Formel 14 die Bedeutung einzelner Werte. Dies bedeutet, dass ein einzelner schwerwiegender Fehler genauso stark in die Bewertung einfließt wie mehrere geringfügige Fehler, da die Formel keine Unterscheidung zwischen der Bedeutung verschiedener Werte vornimmt. Dies könnte zu einer verzerrten Einschätzung führen, insbesondere wenn einige Werte wesentlich für den Produktionsprozess sind und deren Fehlen schwerwiegender ist als das Fehlen anderer, weniger kritischer Informationen. Der Vergleich zwischen diesen Metriken verdeutlicht, dass die Auswahl davon abhängt, welche Aspekte der Vollständigkeit im Kontext eines produzierenden Unternehmens priorisiert werden sollen. Formel 12 ist einfach und bietet eine zügige, dabei jedoch grundlegende Einschätzung der Datenverfügbarkeit, was eine effiziente Ressourcenzuweisung und eine rasche Identifikation von Engpässen ermöglicht. Darüber hinaus fördert die Unkompliziertheit von Formel 12 eine zeitnahe Datenauswertung, was zu einer Zeitersparnis führt. Diese zeitliche Effizienz ermöglicht es produzierenden Unternehmen, proaktiv auf Datenengpässe zu reagieren und gezielte Maßnahmen zur Verbesserung der Datenverfügbarkeit zu ergreifen, wodurch die Grundlage für eine effektive und proaktive Datenverwaltung geschaffen wird. Im Gegensatz dazu konzentrieren sich Formel 13 und Formel 14 primär auf die Anwesenheit von Werten ohne Berücksichtigung ihrer individuellen Bedeutung. Daher könnten sie weniger geeignet sein, wenn es darum geht, die Bedeutung einzelner Werte für produktive Prozesse zu berücksichtigen. Insbesondere wird Formel 14 empfohlen, da sie eine inverse Perspektive auf die *Completeness* bietet, indem sie aufzeigt, wie viele Datensätze fehlen. In produktiven Umgebungen ist dies hilfreich, um gezielt diejenigen Bereiche zu identifizieren, in denen Daten fehlen und somit Verbesserungsbedarf besteht. Durch diese umgekehrte Vollständigkeitsperspektive können Unternehmen ihre Bemühungen zur Datenverbesserung auf besonders fehlerbehaftete Bereiche konzentrieren.

Für die Dimensionen *Reputation* wurde in der systematischen Literaturrecherche keine Datenqualitätsmetrik identifiziert.

Die Dimension *Uniqueness* wurde von Sidi et al. (2012) nicht beschrieben und ist daher kein Bestandteil dieser Literaturrecherche. Um jedoch eine weitere Metrik zur Messung der Datequalität im Hinblick auf die *Veracity* zu erhalten, wir die Integration der Dimension *Duplication* in das Modell wird empfohlen. Diese Erweiterung könnte dazu beitragen, die Gesamtheit der Datenqualität genauer zu erfassen, insbesondere im Kontext von Datenredundanz und mehrfachen Dateneinträgen, die relevante Faktoren für produzierende Unternehmen darstellen können. *Duplication* adressiert genau diese Problematik und bezieht sich auf das Vorhandensein von doppelten oder redundanten Daten in einem Datensatz (Sidi et al. 2012). Diese Empfehlung beruht auf der essenziellen Anforderung, die Zuverlässigkeit und Qualität der Daten sicherzustellen, da Duplikate potenziell zu Unstimmigkeiten, Inkonsistenzen und fehlerhaften Schlussfolgerungen führen können (Helmis und Hollmann 2009). Im Rahmen der Charakteristik *Veracity* wird betont, dass die Qualität der verarbeiteten Daten gewährleistet sein muss (Baars und Kemper 2021). In diesem Kontext ist es besonders bedeutsam sicherzustellen, dass die Daten frei von unnötigen Duplikaten sind, um mögliche Inkonsistenzen und Verzerrungen in Analysen oder Berichten zu vermeiden. Die gezielte Reduktion von Duplikaten spielt somit eine entscheidende Rolle bei der Gewährleistung der Veracity und trägt dazu bei, die Datenqualität in Bezug auf Zuverlässigkeit und Genauigkeit zu steigern. Diese systematische Vorgehensweise stellt sicher, dass die in analytischen Prozessen verwendeten Daten zuverlässig und frei von überflüssigen Wiederholungen sind.

Die identifizierte Datenqualitätsmetrik für die Dimension *Duplication,* beschrieben in Formel 33, bewertet inwieweit Duplikate in einem Datensatz vorhanden sind. Ein niedriger Anteil duplizierter Werte deutet darauf hin, dass die Datenqualität in Bezug auf die Redundanz gering ist, was für produzierende Unternehmen von Vorteil ist. Dies kann zuverlässige Analysen und fundierte Entscheidungsprozesse unterstützen (Schmitt et al. 2020). Ein hoher Anteil duplizierter Werte kann hingegen auf Probleme mit der Datenintegrität hinweisen, was die Zuverlässigkeit der Daten beeinträchtigen kann (Schmitt et al. 2020).

Für die vierte Charakteristika *Variety* werden im Modell die Dimensionen *Consistent Representation*, *Objectively/Objectivity* und *Variety of Data Sources* vorgeschlagen.

Die Dimension *Variety of Data Sources* wurde in der systematischen Literaturrecherche nicht berücksichtigt, da sie nicht in der Beschreibung von Sidi et al. (2012) aufgeführt war (s. Tabelle

1). Daher konnte für diese Dimension keine spezifische Datenqualitätsmetrik identifiziert werden.

Ebenso wurde für die Dimension *Objectively/Objectivity* trotz der systematischen Literaturrecherche keine entsprechende Metrik identifiziert werden. Daher liegt der Fokus ausschließlich auf der Dimension *Consistent Representation*.

Die in Formel 24 dargestellte Datenqualitätsmetrik bietet eine quantitative Bewertung der Konsistenz von Daten anhand zuvor definierter Regeln. Die Metrik bietet eine strukturierte Möglichkeit, die Konsistenz der Daten zu quantifizieren, indem sie vordefinierte Regeln für jedes Datenfeld überprüft. Für produzierende Unternehmen, die oft mit umfangreichen und vielschichtigen Datensätzen arbeiten, ist es von essenzieller Bedeutung, sicherzustellen, dass Daten konsistent und einheitlich repräsentiert werden (Witzenleiter 2023). Diese Konsistenz ist notwendig, um fundierte Entscheidungen treffen zu können, sei es in Bezug auf Produktionsprozesse, Qualitätsmanagement oder Lagerverwaltung (Witzenleiter 2023). Ein entscheidender Aspekt bei der Anwendung dieser Metrik in produzierenden Unternehmen besteht darin, dass die Konsistenz der Daten stark von der Qualität und Gültigkeit der festgelegten Regeln abhängt. Daher erfordert die effektive Nutzung der Metrik eine kontinuierliche Überwachung und Aktualisierung der Regeln, um sicherzustellen, dass sie weiterhin den sich ändernden Anforderungen der Produktionsprozesse entsprechen. Dieser Anpassungsprozess kann jedoch zeitaufwändig sein und erfordert eine fortlaufende Abstimmung mit den sich entwickelnden Geschäftsanforderungen. Die enge Abstimmung mit den Fachexperten zur Definition der Regeln und den individuellen Anforderungen des Unternehmens ist dabei entscheidend, um den maximalen Nutzen aus dieser Metrik zu ziehen.

Zusammenfassend ergeben sich für die Bewertung der Datenqualität in produzierenden Unternehmen die folgenden praxisorientierten Empfehlungen. Die empfohlene Datenqualitätsmetriken werden in Abbildung 9 dargestellt und im Anschluss kurz zusammengefasst.

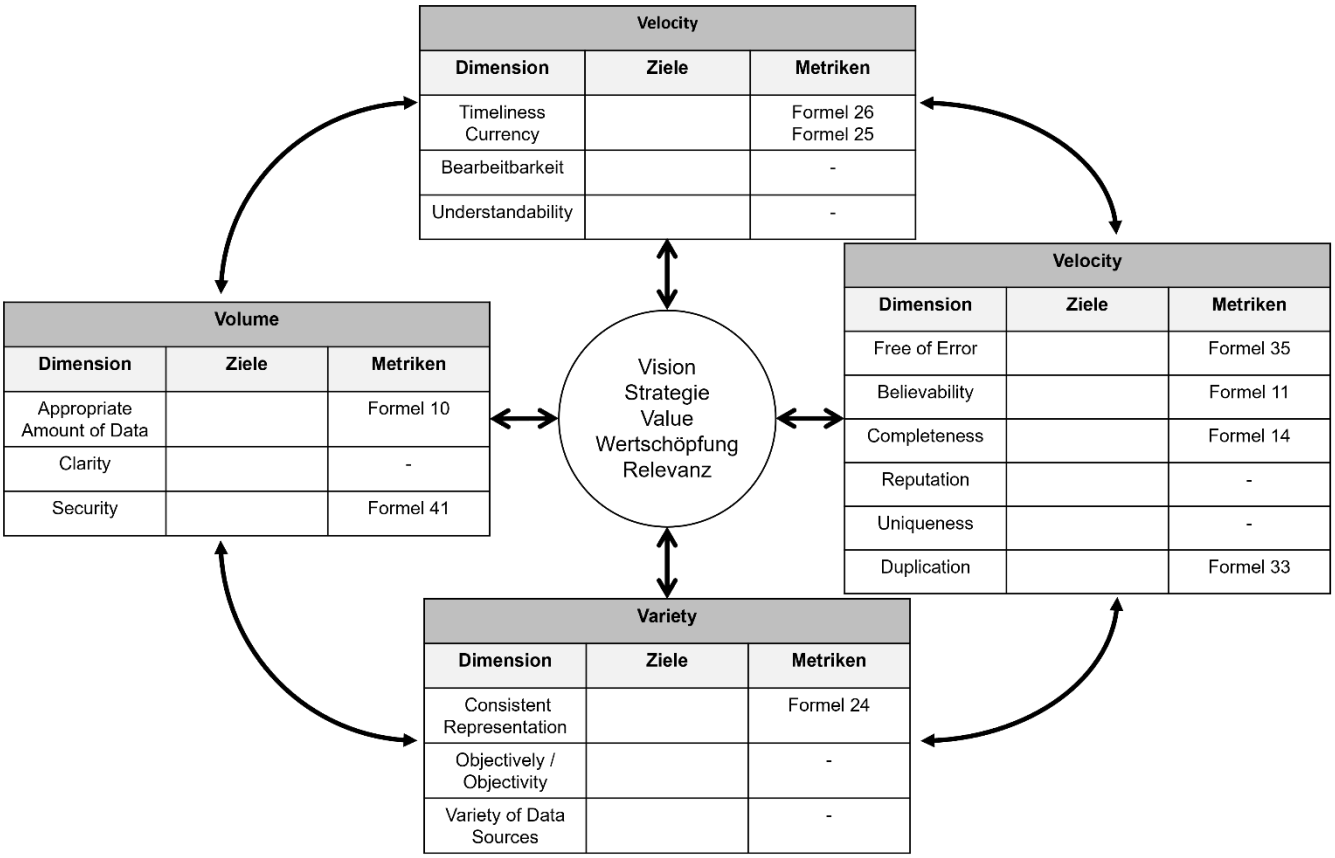

Abbildung 9: Handlungsempfehlung zur Messung der Datenqualität in produzierende Unternehmen

Die Dimension *Timeliness*, die die Aktualität von Daten betrachtet, sollte mittels Formel 26 evaluiert werden. Diese Formel ermöglicht eine differenzierte Beurteilung der Datenaktualität, insbesondere im Hinblick auf Produktionsprozesse, Lagerverwaltung und Qualitätssicherung. Für die Dimension *Currency*, die auf die Aktualität von Daten abzielt wird die Anwendung von Formel 25 empfohlen. Diese einfache Metrik liefert eine schnelle Einschätzung der Gesamtdatenaktualität, insbesondere in schnelllebigen Produktionsumgebungen. Die Datenqualitätsdimension *Appropriate Amount of Data*, die das Verhältnis zwischen benötigten und verfügbaren Daten betrachtet, wird durch die Nutzung von Formel 10 bewertet. Diese Metrik identifiziert den Engpass oder das kritischste Verhältnis zwischen erforderlichen und verfügbaren Daten, was insbesondere in produktiven Umgebungen für einen reibungslosen Ablauf von Produktionsprozessen von Bedeutung ist. Die Dimension *Security*, die die Sicherheit von Daten betrachtet, sollte mit Formel 41 bewertet werden. Diese Metrik basiert auf einem klaren Punktesystem und berücksichtigt wichtige Sicherheitsaspekte von der Existenz einer Sicherheitsrichtlinie bis zur Datenverschlüsselung. Für die Dimension *Free of Error* wird Formel 35 als emp-

fohlene Metrik vorgeschlagen. Diese Metrik ermöglicht eine klare und leicht verständliche Bewertung der Fehlerfreiheit von Daten. Es ist jedoch wichtig zu beachten, dass diese Metrik die Schwere der Fehler nicht berücksichtigt und somit für eine oberflächliche Einschätzung der Fehlerfreiheit geeignet ist. Die Dimension *Believability* bezieht sich auf die Glaubwürdigkeit von Daten. Hier wird Formel 11 als empfohlene Metrik verwendet. Diese Metrik bietet einen anspruchsvolleren Ansatz, indem sie die Glaubwürdigkeit jedes einzelnen Datenfeldes individuell bewertet. Es ist jedoch zu beachten, dass die subjektive Natur dieser Bewertung zu unterschiedlichen Interpretationen führen kann, da verschiedene Personen die Glaubwürdigkeit eines Datenfeldes unterschiedlich beurteilen könnten. Die Dimension *Completeness* betrachtet die Vollständigkeit von Daten. Für diese Dimension wird Formel 14 als empfohlene Metrik präsentiert. In produktiven Umgebungen ist dies hilfreich, um gezielt diejenigen Bereiche zu identifizieren, in denen Daten fehlen und somit Verbesserungsbedarf besteht. Die Dimension *Duplication*, die auf das Vorhandensein von doppelten oder redundanten Daten abzielt, wird durch die Anwendung von Formel 33 evaluiert. Diese Metrik quantifiziert den Anteil duplizierter Werte in einem Datensatz und gibt Hinweise auf die Redundanz von Daten. Die Dimension *Consistent Representation* fokussiert auf die konsistente Darstellung von Daten. Die empfohlene Metrik Formel 24 bietet eine quantitative Bewertung der Konsistenz anhand vordefinierter Regeln. Bei der Implementierung dieser Metrik ist eine sorgfältige Definition und Anpassung der Regeln erforderlich, um relevante und aussagekräftige Ergebnisse zu gewährleisten. Für die Dimensionen *Editability*, *Understandability, Clarity, Reputation, Uniqueness*, *Objectively/Objectivity* und *Variety of Data Sources* konnten keine spezifischen Metriken identifiziert werden. Daher wird empfohlen, die Bewertung dieser Dimensionen individuell auf die spezifischen Anforderungen und Gegebenheiten des produzierenden Unternehmens abzustimmen.

Nachdem die Evaluierung der Datenqualitätsmetriken abgeschlossen ist, ist es von entscheidender Bedeutung zu betonen, dass die Anwendung dieser Metriken stark vom spezifischen Einsatzgebiet und den damit verbundenen Zielen abhängt (Hildebrand et al. 2021b). Unternehmen sollten die Auswahl und Relevanz der Metriken individuell prüfen und gegebenenfalls Anpassungen an den Formeln vornehmen. Die maßgeschneiderte Anpassung von Metriken an die spezifischen Anforderungen produzierender Unternehmen ist unerlässlich, um sicherzustellen, dass die Daten den spezifischen Bedürfnissen der Produktion und der gesamten Wertschöpfungskette gerecht werden. Dies gewährleistet nicht nur die Effizienz, sondern auch

die Qualität und Wettbewerbsfähigkeit des Unternehmens (Hildebrand et al. 2021b). Insgesamt verdeutlicht dies die Notwendigkeit einer kontextbezogenen Implementierung von Datenqualitätsmetriken, um maximalen Nutzen in industriellen Produktionsumgebungen zu erzielen. Im folgenden Kapitel wird nun der Einfluss von exemplarisch ausgewählten Datenvorverarbeitungsverfahren auf ausgewählte Datenqualitätsmetriken evaluiert und beschrieben.

4 Exemplarische Anwendung von Datenqualitätsmetriken unter Einsatz unterschiedlicher Datenvorverarbeitungsverfahren

Nach dem im vorherigen Kapitel eine umfassende Betrachtung von Datenqualitätsmetriken im Kontext der fünf Charakteristika der Big Data im Zusammenhang mit produzierenden Unternehmen erfolgt ist, wird im Folgenden eine exemplarische Anwendung der empfohlenen Datenqualitätsmetriken unter Einsatz unterschiedlicher Datenvorverarbeitungsverfahren durchgeführt. Die Datenvorverarbeitung, ein integraler Bestandteil des im Abschnitt 2.4 eingeführten CRISP-DM, wird in Abschnitt 2.5 durch die Darlegung grundlegender Verfahren näher erläutert. Die in diesem Abschnitt präsentierten Verfahren finden Anwendung im vorliegenden Kapitel. Hierzu wird zuerst der Datensatz beschrieben und anschließend die Auswahl der Datenqualitätsmetriken begründet. Im Anschluss werden die ausgewählten Metriken errechnet und der Einfluss der Datenvorverarbeitungsverfahren auf das Ergebnis der Metriken eingeordnet.

4.1 Datensatzcharakterisierung

Der vorliegende Datensatz mit dem Titel "Find a defect in the production" wurde von der Datenplattform Kaggle bezogen und fokussiert sich auf eine Extrusionsanlage. Diese Anlage besteht aus drei separaten Einheiten, die in einem komplexen Prozess granulares Rohmaterial aufnehmen und es durch Erhitzen und Ziehen in drei Schichten zu einem Kunststofffilm verarbeiten. Die Datenerhebung erstreckte sich über einen Zeitraum von einem Jahr, vom 25. Juni 2018 bis zum 25. Juni 2019. Dieser Zeitrahmen ermöglicht eine umfassende Analyse der Produktionsdaten und bietet Einblicke in langfristige Trends sowie saisonale Schwankungen.

Der Datensatz setzt sich aus insgesamt drei Dateien zusammen. Zwei dieser Dateien sind im CSV-Format vorliegend, während die dritte als XLSX-Datei strukturiert ist. Die Hauptdatei im CSV-Format umfasst 481 Merkmale. Diese Merkmale gliedern sich in 472 numerische und 9 textbasierte Kategorien. Diese Fülle von Informationen ermöglicht eine detaillierte Analyse verschiedener Aspekte der Produktionsprozesse in der Extrusionsanlage. Mit 237.862 Zeilen bietet die Hauptdatei eine umfangreiche Grundlage für statistische Analysen, maschinelles Lernen und tiefergehende Erkenntnisse in die Produktionsabläufe. Die Struktur der Daten ist tabellarisch angelegt, was die Handhabung und Analyse durch verschiedene statistische Methoden erleichtert. Die Informationen in der zweiten Datei im XLSX-Format werden präzise erklärt

und fungieren als Legende der Spalten, wobei jede Spalte eine detaillierte Beschreibung und die zugehörigen Einheiten enthält. Diese Dokumentation ist entscheidend für Forscher und Analysten, um die Daten korrekt zu verstehen und sinnvolle Schlussfolgerungen zu ziehen. Die dritte Datei im CSV-Format spielt eine unterstützende Rolle, indem sie Beschriftungen für die Spalten bereitstellt. Dieser zusätzliche Layer von Metadaten erleichtert die Interpretation der Daten weiter und stellt sicher, dass keine Unklarheiten bezüglich der Bedeutung einzelner Spalten bestehen. Solche Beschriftungen sind sinvoll, um eine klare Kommunikation zwischen den an der Datenauswertung beteiligten Parteien zu gewährleisten. Die Datenerhebung erfolgte durch den Einsatz von Sensoren an verschiedenen kritischen Stellen der Extrusionsanlage. Diese Sensoren ermöglichen eine kontinuierliche Erfassung von relevanten Produktionsparametern und schaffen eine umfassende Grundlage für die spätere Analyse. Der Einsatz von Sensoren ist in der modernen Fertigungsindustrie weit verbreitet und ermöglicht eine präzise Überwachung und Steuerung der Produktionsprozesse.

Bevor nun auf die Anpassungen der Datensätze zur weiteren Nutzung beschrieben werden, ist es wichtig, die genutzte Softwareumgebung zu beleuchten, um eine transparente und reproduzierbare Forschungsgrundlage zu gewährleisten. Die folgende Version der Programmiersprache sowie Bibliotheken wurden verwendet, jeweils mit ihren spezifischen Versionen:

- Python-Version: 3.10.12 (main, Nov 20 2023, 15:14:05) [GCC 11.4.0]
- pandas-Version: 1.5.3
- openpyxl-Version: 3.1.2

Im Verlauf der Arbeit wurden die ursprüngliche Datensätze gezielt bearbeitet, um ihn optimal an das Forschungsziel anzupassen. Die Hauptdatei wurde von einer CSV-Datei in eine XLSX-Datei konvertiert, um die Handhabung der Daten zu erleichtern. Ursprünglich mit 481 Merkmalen ausgestattet, wurden die Merkmale auf drei reduziert. Diese umfassen das *Datum* als Primärschlüssel mit Informationen über Tag und Uhrzeit, die IST-Werte aus der Spalte *ST110_VAREx_0_Dos_0_IstLMGewicht* (tatsächliches Laufmetergewicht, Gerät: Extruder, Einheit: g/m) und die Soll-Werte in der Spalte *ST110_VAREx_0_SollLM* in derselben Struktur wie die IST-Werte. Diese gezielte Anpassung ermöglicht eine vereinfachte Analyse der ausgewählten Metriken. Weiterhin wurde der Primärschlüssel *Datum* aufgeteilt in die Merkmal *Dat.* und *Uhrzeit,* wobei der Primärschlüssel *Datum* bestehen bleibt. Auch der hierzu genutzte Code ist in der angehängten Datei zu finden. Nach dem nun der Datensatz charakterisiert worden

und auf die Bedürfnisse dieser Arbeit angepasst worden ist, wird im folgenden Kapitel die Auswahl der genutzten Metriken anhand von Merkmalen des Datensatzes beschrieben.

4.2 Selektion geeigneter Datenqualitätsmetriken

Die initiale Charakterisierung des vorliegenden Datensatzes offenbarte bestimmte Herausforderungen und potenzielle Schwachstellen. Durch die Identifikation von Leerstellen, 0-Werten, Abweichungen zwischen IST- und Soll-Werten sowie Duplikationen wurden erste Anomalien erkannt, die die Datenqualität beeinträchtigen könnten. Um diese Unregelmäßigkeiten systematisch zu erfassen und zu bewerten, wurden gezielt bestimmte Datenqualitätsdimensionen ausgewählt. Diese ausgewählten Dimensionen konzentrieren sich auf die identifizierten Mängel im Datenbestand. In einem weiteren Schritt wurden Datenqualitätsmetriken ausgewählt, die speziell auf die identifizierten Dimensionen abzielen. Durch die Anwendung dieser Metriken wird eine quantitative Analyse der Dimensionen *Completeness*, *Free of Error* und *Duplication* ermöglicht. Die Metriken, die ausgewählt wurden, entsprechen den im Abschnitt 3.1 identifizierten und in Abschnitt 3.2 bewerteten und empfohlenen Metriken für die entsprechende Dimension. Im Folgenden werden die gewählten Metriken im Detail erläutert und deren Anwendbarkeit auf die spezifischen Herausforderungen des Datensatzes diskutiert. Die ausgewählten Datenqualitätsmetriken werden in Tabelle 4 dargestellt.

Tabelle 4: Übersicht über ausgewählte Datenqualitätsmetriken

Datenqualitätsmetrik	Formel
Completeness	$Completeness = 1 - \frac{T_R}{N_R}$
Free of Error	$Fehlerfreiheit = 1 - (\frac{Anzahl\ fehlerhafter\ Daten}{Gesamtanzahl\ an\ Daten})$
Duplication	$Anteil\ der\ duplizierten\ Werte = \frac{Anzahl\ der\ duplizierten\ Datensätze}{Gesamtanzahl\ der\ Datensätze}$

Eine initiale Prüfung identifizierte Leerstellen und Zeilen mit IST-Werten von 0 als potenzielle Unregelmäßigkeiten. Die Identifikation von Leerstellen und 0-Werten weist auf mögliche Lücken in den Daten hin. Um das Ausmaß der Vollständigkeit zu erfassen, wurde die Metrik *Completeness* ausgewählt. Diese Metrik ermöglicht eine quantitative Bewertung, indem leere und fehlende Zeilen ins Verhältnis zur Gesamtanzahl der Zeilen im Datensatz gesetzt werden. So lässt sich genau bestimmen, inwieweit die Daten vollständig sind. Weiterhin wurde bei der Prüfung der Daten, einige Datensätze gefunden, die eine Abweichung von 90% aufweisen. Ein Beispiel hier wäre der Datensatz am *22.03.2018 um 17:39 Uhr.* Hier entspricht der IST-Wert *25.40 und der Soll-Wert 265.56.* Datensätze, die eine solche Charakteristika aufweisen, wurden als fehlerhaft eingestuft. Um nun die Fehlerfreiheit der Daten zu bewerten und im Anschluss die Verbesserung zu überprüfen wird die Datenqualitätsmetrik *Free of error* ausgewählt. Diese Metrik ermöglicht die quantitative Analyse von Datenfehlern, indem Datensätze mit erheblichen Abweichungen von den Soll-Werten erfasst und quantifiziert werden. Dies dient der Identifikation und Bewertung von fehlerhaften Daten im Datensatz. Eine weitere Unregelmäßigkeit, die bei der Prüfung der Daten aufgefallen ist, betrifft duplizierte Werte. Doppelte Werte wurden über den Primärschlüssel erkannt und als Duplikate betrachtet und werden durch die Datenqualitätsmetrik *Duplication* berechnet.

Die gezielte Auswahl dieser Metriken spiegelt die Bestrebungen wider, die spezifischen Qualitätsprobleme des Datensatzes präzise zu erfassen und zu quantifizieren. Jede Metrik fokussiert sich auf einen spezifischen Aspekt der Datenqualität, von Vollständigkeit über Fehlerfreiheit bis hin zur Duplikation und ermöglicht somit eine umfassende Bewertung und gezielte Verbesserung der Datenqualität. Die Neuberechnung der Datenqualitätsmetriken nach der Durchführung von Datenvorbereitungsschritten ermöglicht eine objektive Bewertung der Effektivität der durchgeführten Maßnahmen zur Steigerung der Datenqualität. Durch den Vergleich der Datenqualitätsmetrikwerte vor und nach der Datenbearbeitung wird deutlich, inwiefern die angewandten Maßnahmen einen positiven Einfluss auf die Vollständigkeit, Fehlerfreiheit und Duplikation der Daten hatten.

4.3 Anwendung des Fallbeispiels zur Berechnung der Metriken

Nachdem eine Auswahl der Datenqualitätsmetriken stattgefunden hat und die Daten bereits in einer verkleinerten xlsx-Datei vorliegen, werden zunächst die ausgewählten Metriken ohne

weitere Datenvorverarbeitungsschritte auf die Daten angewendet. Die Vollständigkeitsmetrik wird durch das Einlesen der Excel-Datei und das Zählen sowohl der leeren Zeilen als auch der Zeilen mit dem Wert 0 ermittelt. Die Summe dieser beiden Kategorien wird als Variable "Anzahl_leere_werte" festgelegt, wobei parallel dazu die Gesamtanzahl aller Zeilen erfasst wird. Insgesamt weist der Datensatz 64 leere Zeilen, 31.811 Zeilen mit dem Wert 0 und eine Gesamtanzahl von 237.862 Zeilen auf. Die *Completeness* wird durch Anwendung der in Tabelle 4 vorgestellten Formel berechnet und beträgt 86,6%. Im Kontext der Fehlerfreiheitsmetrik erfolgt die Analyse von Datensätzen, die die in Abschnitt 4.2 beschriebenen Voraussetzungen erfüllen. Hierbei werden die Abweichungen zwischen den IST- und SOLL-Werten untersucht. Zeilen mit Abweichungen im Bereich von 9% bis 11% werden als fehlerhaft eingestuft, wie im dazugehörigen Code in Anhang dargestellt. Die Wahl dieser Abweichungsgrenze erfolgte aufgrund der Annahme, dass bei den identifizierten Datensätzen die Möglichkeit besteht, dass die letzte Ziffer fehlt oder nicht korrekt übernommen wurde. Im betrachteten Datensatz werden 2 fehlerhafte Zeilen identifiziert, was eine Fehlerfreiheit von 99,9% ergibt. Die Duplikationsmetrik wiederum überprüft die Übereinstimmung des Primärschlüssels und zählt die Duplikate mithilfe eines programmierten Codes, der in der beigefügten Datei abgebildet ist. Insgesamt werden 11.326 Duplikate identifiziert, was einer Duplikationsrate von 4,76% entspricht. Diese detaillierten Analysen bieten Einblicke in die Qualität der vorliegenden Daten und legen den Grundstein für weiterführende Untersuchungen im Rahmen dieser Arbeit. Nach dem nun die grundsätzlichen Datenlage analysiert wurde, wird nun eine Datenbereinigung durchgeführt, um zu prüfen, welchen Einfluss diesen auf die Datenqualitätsmetriken besitzen.

Im Rahmen der Datenbereinigung wird die Listwise Deletion angewendet. Bei dieser werden alle Beobachtungen, für die mindestens bei einer Variablen ein Wert fehlt, ausgeschlossenen (Stoetzer 2020). Um diese Vorgehensweise zur realisieren, wird ein Python-Code verwendet, der mithilfe der Pandas-Bibliothek implementiert ist. Zunächst werden die Daten aus der Eingabe-XLSX-Datei in ein DataFrame namens "data" eingelesen. Anschließend wird eine spezifische Bereinigung durchgeführt, um Zeilen zu entfernen, die leere Werte oder den Wert 0 in der Spalten ST110_VAREx_0_Dos_0_IstLMGewicht enthalten. Der Code beginnt damit, die Anzahl aller Zeilen vor dem Löschen zu erfassen und speichert diese Information in der Variable "anzahl_vor_loeschen". Die eigentliche Bereinigung erfolgt durch die Anwendung einer Lambda-Funktion auf jede Zelle des DataFrames. Hierbei werden Zellen mit dem Wert 0 durch

das Pandas-Objekt "pd.NA" ersetzt. Die Methode "dropna" wird daraufhin verwendet, um Zeilen zu entfernen, in denen in den Spalten 'ST110_VAREx_0_Dos_0_IstLMGewicht' oder 'ST110_VAREx_0_SollLM' leere Werte vorhanden sind. Die Anzahl der gelöschten Zeilen wird dann durch Subtraktion von der ursprünglichen Anzahl aller Zeilen ermittelt und in der Variable "anzahl_geloeschter_zeilen" gespeichert. Die bereinigten Daten werden schließlich in einer neuen XLSX-Datei unter Verwendung des Dateipfads "output_xlsx_path" gespeichert. Abschließend gibt der Code deskriptive Informationen aus, darunter die ursprüngliche Anzahl aller Zeilen vor dem Löschen, die Anzahl der gelöschten Zeilen und die Anzahl der verbleibenden Zeilen nach der Bereinigung. Der Dateipfad, unter dem die bereinigten Daten gespeichert wurden, wird ebenfalls angezeigt. Zusammenfassend automatisiert dieser Code den Prozess der Datenbereinigung, indem er leere Werte und Nullen in bestimmten Spalten entfernt und gleichzeitig statistische Informationen über den Bereinigungsprozess ausgibt. Die automatisierte Bereinigung hat dabei 11.326 Datenzeilen entdeckt und gelöscht, die den vorher definierten Charakteristiken entsprochen haben. So bleiben weiterhin 205.985 Datenzeilen im gesamten Datensatz bestehen.

Nachdem nun der erste Schritt der Datenbereinigung 0 durchgeführt worden ist, werden nun die ausgewählten Datenqualitätsmetriken neu berechnet, um zu prüfen, ob es zu Veränderungen in der Bewertung durch die Datenqualitätsmetriken kommt. Durch die Bereinigung der Daten mit der Listwise Deletion ist die Gesamtanzahl der Daten reduziert worden und führt dabei gleichzeitig zu einer *Completeness* von 100%. Die Anzahl falscher Daten blieb auf dem vorherigen Niveau von zwei, was zu einer Fehlerfreiheit von 99,9% führte. Ein Unterschied zeigt sich auch bei der Anzahl der Duplikate in der Kombination von 'Uhrzeit' und 'Datum', die nach der Bereinigung 9,841 betrug, während es im ursprünglichen Datensatz 11,326 Duplikate waren. Dies entspricht einem Prozentsatz von 4.78%.

Im Anschluss soll nun der Einfluss eines weiteren Vorverarbeitungsverfahren geprüft werden. Hierzu wurde die Eliminierung von doppelten Datensätzen ausgewählt. Diese Praxis wurde bereits im Jahr 2000 von Rahm als empfehlenswert erachtet, um die Qualität von Daten zu optimieren und Redundanzen zu eliminieren. Durch die Umsetzung dieses bewährten Ansatzes wird angestrebt, die Zuverlässigkeit des Datensatzes zu erhöhen. Der zugrunde liegende Python-Code implementiert diese Datenaufbereitung mithilfe der Pandas-Bibliothek. Zunächst wird der Datensatz aus einer Excel-Datei geladen und in ein Pandas DataFrame transformiert.

Im Anschluss erfolgten die Identifikation und Entfernung von Duplikaten, wobei die Kombination aus den Spalten 'Dat.' und 'Uhrzeit' als eindeutiges Merkmal herangezogen wird. Dieser Schritt dient dazu, sicherzustellen, dass nur das erste Auftreten eines Datensatzes beibehalten wird, während nachfolgende Duplikate entfernt werden. Durch diese Methoden wurde insgesamt 11.326 Zeilen identifiziert und gelöscht, womit 226.536 Datensätze weiterhin betrachtet werden. Nach der Bereinigung werden die aktualisierten Daten in eine neue Excel-Datei gespeichert, wobei der Index aus Gründen der Datensauberkeit ausgelassen wird. Abschließend gibt der Code Auskunft über die Anzahl der gelöschten und verbleibenden Zeilen, um Transparenz über den Bereinigungsprozess zu schaffen. Rahms (2000) Empfehlungen bilden somit die Grundlage für eine zeitgemäße und effektive Datenbereinigung, die auf bewährten Prinzipien basiert und dazu beiträgt, die Integrität der Daten zu gewährleisten.

Bei der erneuten Betrachtung der Datenqualitätsmetriken für den neu erstellten Datensatz sind folgende Ergebnisse errechnet worden. Die *Completeness* des Datensatzes erreicht 86.58%, wobei 62 leere Zeilen und 30,328 Zeilen mit dem Wert 0 verbleiben. Trotz dieser leichten Abnahme bleibt die Gesamtanzahl der Zeilen mit 226,536 hoch. In Bezug auf die Dimension *Free of error* gab es zum anderen Vorverarbeitungsverfahren keine Änderungen. Dies bedeutet, dass von insgesamt 226,536 Daten sind, nur noch 2 falsche Daten vorhanden, was zu einer Fehlerfreiheit von 99.99% führt. Besonders auffallend ist die Duplikationsrate, die nach der gezielten Entfernung von doppelten Zeilen in der Kombination von 'Uhrzeit' und 'Datum' nun bei 0 liegt. Dies bedeutet, dass keine Duplikate mehr in diesem Schlüsselbereich vorhanden sind, was auf eine erfolgreiche Bereinigung und Optimierung des Datensatzes hinweist.

Im Anschluss werden, wie in der Literatur empfohlen, verschiedene Datenvorverarbeitungsverfahren genutzt (Maydanchik 2007). Ein ganzheitlicher Ansatz, der mehrere Verfahren integriert, ermöglicht eine zielgerichtete und umfassende Verbesserung der Datenqualität, um den branchenspezifischen Anforderungen gerecht zu werden (Maydanchik 2007). Es werden drei verschiedene Datenvorverarbeitungsverfahren auf die Datensätze angewandt. Zum einen werden die Zeilen mit Nullen oder leeren Werten entfernt. Dies trägt zur Verbesserung der Datenqualität bei, indem irrelevante oder fehlerhafte Einträge beseitigt werden. Anschließend erfolgt die Entfernung der duplizierten Werte analog dem vorherigen Code. Als letzter Schritt der Datenvorverarbeitung werden die falsch eingetragenen Werte anhand der definierten Toleranzwerte identifiziert und durch die Soll-Werte ausgetauscht mit Hilfe eines Masken-Arrays.

Abschließend gibt der Code detaillierte Informationen über die Anzahl der gelöschten Zeilen, die Anzahl der verbleibenden Zeilen nach der Vorverarbeitung, die Anzahl der gelöschten Duplikate und die Anzahl der ausgetauschten Werte aus. Diese Ausgaben bieten Transparenz und Einblick in den durchgeführten Bereinigungsprozess. Nachdem die ausgewählten Metriken anhand exemplarisch durchgeführter Datenvorverarbeitungsverfahren errechnet worden sind, werden im Folgenden die Ergebnisse diskutiert und interpretiert.

4.4 Diskussion der Ergebnisse

Im kommenden Abschnitt werden die Ergebnisse der Metriken zur Messung der Datenqualität erörtert und Implikationen aufgezeigt. Verschiedene Vorverarbeitungstechniken wurden auf den ausgewählten Datensatz angewendet, um die Auswirkungen auf die in Abschnitt 4.2 ausgewählten Datenqualitätsmetriken zu überprüfen. Das Ziel der Datenvorverarbeitung bestand darin, die Datenqualität zu optimieren und eine zuverlässige Grundlage für weiterführende Analysen zu schaffen.

Zur Schaffung der Grundlagen wurden zunächst die Datenqualitätsmetriken anhand des unbehandelten Datensatzes berechnet. Dies diente dazu, einen Referenzwert zu erhalten, mit dem die berechneten Werte der behandelten Datensätze verglichen werden konnten. Der unbehandelte Datensatz wies eine *Completeness* von 86,6% auf, was auf das Vorhandensein von leeren Zeilen, Nullwerten und falschen Daten hinweist. Darüber hinaus betrug die Metrik der Dimension *Free of Error* im unbehandelten Datensatz 99,9%, und die Duplikationsrate lag bei 4,76%.

Darauf folgte die Listwise Deletion, bei der leere Zeilen und Nullwerte gelöscht wurden. In diesem Fall zeigte der Datensatz veränderte Metrikergebnisse. Die *Completeness* stieg auf den erwarteten Wert von 100% an. Das Ergebnis der Fehlerfreiheit blieb jedoch unverändert, was darauf hindeutet, dass keine entsprechenden Daten gelöscht wurden. Überraschenderweise verschlechterte sich die Duplikationsrate leicht auf 4,78%, obwohl offensichtlich 1.485 Duplikate aus dem Datensatz entfernt wurden. Dies lässt sich darauf zurückführen, dass Duplikate in den Charakteristika der gelöschten Datensätze vorhanden waren. Dies ist ein Hinweis darauf, dass sowohl doppelte als auch nicht-doppelte Werte, insbesondere diejenigen mit dem Wert 0 oder ohne Wert, in etwa gleicher Häufigkeit auftreten.

Als zweite Form der Datenvorbereitung wurde eine Duplikationserkennung und -löschung auf den unbehandelten Datensatz angewandt, wodurch insgesamt 11.326 doppelte Datenzeilen entfernt wurden. Dies hatte wiederum Auswirkungen auf die Ergebnisse der ausgewählten Datenqualitätsmetriken. Die Metrik der *Completeness* verringerte sich leicht im Vergleich zum unbehandelten Datensatz und betrug nun 86,58%. Auch hier wurde deutlich, dass einige doppelte Werte in die Charakteristiken der Metrik Vollständigkeit fielen. Das Ergebnis der *Free of error* Berechnung blieb unverändert, was darauf schließen lässt, dass diese bestimmte Fehlerart entweder nicht beeinträchtigt wurde oder keine zugehörigen Daten gelöscht wurden. Die Duplikation sank auf den erwarteten Wert von 0%, was darauf hindeutet, dass alle doppelten Werte gelöscht wurden. Dies deutet darauf hin, dass der Code ordnungsgemäß funktioniert.

Zuletzt wurden die einzelnen Datenvorverarbeitungsverfahren in der Gesamtheit auf den unbehandelten Datensatz angewendet. Insgesamt wurden 41.716 Datensätze gelöscht und weitere zwei Datensätze ausgetauscht, wodurch der Datensatz nur noch 196.146 Datensätze enthielt. Durch diese Vorverarbeitungsschritte konnte die Qualität der ausgewählten Datenqualitätsmetriken weiter verbessert werden. Die Berechnung der Metriken *Completeness* und *Duplication* ergab einen Wert von 100%, was darauf hinweist, dass es weder leere Zeilen noch Duplikate im Datensatz gibt. Die Berechnung der Fehlerfreiheit zeigte, dass der Austausch der als falsch deklarierten Datensätze erfolgreich war. Durch die Anwendung mehrerer Datenvorverarbeitungsverfahren hat sich die Datenqualität deutlich verbessert, wie durch den Vergleich der berechneten Datenqualitätsmetriken ersichtlich wird. Auch im Vergleich zu den einzelnen Verfahren konnte die Qualität der Daten durch die Nutzung mehrerer Verfahren gesteigert werden. Daraus lässt sich ableiten, dass die Anwendung vieler verschiedener Datenvorverarbeitungsschritte die Datenqualität stärker verbessert als die Anwendung einzelner Verfahren. Jedoch muss der Datensatz im Vorfeld analysiert und geprüft werden, um sinnvolle Verfahren zu identifizieren. Weiterhin müssen, um die Veränderungen am Datensatz zu quantifizieren, Datenqualitätsmetriken sinnvoll gewählt werden. Die Auswahl der Metriken steht dabei im direkten Zusammenhang mit den ausgewählten Datenvorverarbeitungsverfahren, da verschiedene Vorverarbeitungsverfahren unterschiedliche Effekte auf die Datenqualität ausüben. Daher ist es wichtig, Metriken zu wählen, die gut zu den spezifischen Vorverarbeitungsschritten passen. Diese Metriken helfen sicherzustellen, dass die angewendeten Vorverarbeitungsverfahren die gewünschten Ergebnisse liefern und die Daten für die geplante Analyse geeignet

sind. Nachdem der Einfluss von Datenvorverarbeitungen auf ausgewählte Metriken diskutiert worden ist, wird im folgenden Kapitel ein Gesamtfazit der Arbeit gezogen. Dabei wird auf die in Kapitel 1 definierten Ziele eingegangen und die Ergebnisse zusammengefasst und kritisch eingeordnet.

5 Schlussbetrachtung

In diesem Kapitel werden die zuvor vorgestellten Ergebnisse diskutiert und kritisch eingeordnet. Dabei wird Bezug auf die zentralen Ziele der Arbeit aus Kapitel 1 genommen und mit Hilfe der Ergebnisse evaluiert. Anschließend werden die Implikationen und ein Ausblick gegeben.

5.1 Diskussion der Ergebnisse

Das erste Teilziel, das Grundverständnis für Datenqualität und der Wissensentdeckung zu schaffen, wurde in Kapitel 2 durch eine tiefgehende Auseinandersetzung mit Begrifflichkeiten wie der Hierarchie des Wissens und dem Prozess der Wissensentdeckung in Datenbanken erfüllt. Besondere Aufmerksamkeit galt dabei dem CRISP-DM, einem bewährten Modell für den Datenanalyseprozess. Die Einführung in Datenvorverarbeitungsverfahren, ihre Bedeutung für produzierende Unternehmen und die Herausforderungen im Kontext von Industrie 4.0 und Big Data bildeten die Grundlage für das Verständnis der Datenqualität.

Die Liste von Datenqualitätsdimensionen, die im Zuge des ersten Teilziels erstellt wurde, erwies sich als wertvolles Werkzeug für die systematische Literaturrecherche in Kapitel 3 – dem zweiten Teilziel. Durch diese Recherche wurden 42 verschiedene Datenqualitätsmetriken für 20 Datenqualitätsdimensionen identifiziert. Die Suche nach passenden Metriken für weitere 20 Dimensionen verdeutlichte gleichzeitig die Herausforderungen und Lücken in der Literatur hinsichtlich bestimmter Datenqualitätsaspekte. Nach der systematischen Literaturrecherche wurden die identifizierten Datenqualitätsmetriken eingehend miteinander verglichen und bewertet, insbesondere im Hinblick auf ihre Anwendbarkeit in produzierenden Unternehmen. Dieser Bewertungsprozess orientierte sich am Rahmenkonzept von Frehe et al. (2016), welches um die zusätzliche Dimension der Duplikation erweitert wurde. Die eingehende Bewertung verdeutlichte, dass nicht alle Datenqualitätsmetriken gleichermaßen für die Implementierung in produzierenden Unternehmen geeignet sind. Diese Differenzierung lässt sich teilweise auf die komplexere Struktur und den Aufbau einiger Metriken zurückführen, jedoch auch auf die komplizierte Definition von bestimmten Faktoren innerhalb der Metrik. Ein tiefgreifendes Verständnis der spezifischen Anforderungen und Herausforderungen in produzierenden Unternehmen spielten dabei eine entscheidende Rolle. Im Ergebnis der Bewertung konnten konkrete Handlungsempfehlungen für produzierende Unternehmen abgeleitet werden. Diese

Empfehlungen beinhalten die gezielte Auswahl bestimmter Datenqualitätsmetriken innerhalb des Rahmenkonzepts, um die adressierten Dimensionen effektiv zu bemessen. Durch diese maßgeschneiderte Auswahl wird gewährleistet, dass die gewählten Metriken nicht nur vielfältig sind, sondern auch präzise auf die spezifischen Bedürfnisse und Charakteristiken der produzierenden Industrie zugeschnitten sind.

Im Rahmen des letzten Teilziels sollte der Einfluss von verschiedenen Datenvorverarbeitungsverfahren evaluiert werden. Hierfür wurde ein Datensatz mit verschiedenen Datenvorverarbeitungsverfahren behandelt. Dabei konnte festgestellt werden, dass die Anwendung verschiedener Datenvorverarbeitungsverfahren einen Beitrag zur Verbesserung der Datenqualität leistet, wie durch den Vergleich der berechneten Datenqualitätsmetriken deutlich wird. Die erzielte Qualitätssteigerung durch die Kombination mehrerer Verfahren übertrifft dabei die Effekte der einzelnen Vorverarbeitungsschritte. Dies unterstreicht die Wirksamkeit eines umfassenden Ansatzes bei der Bearbeitung von Datensätzen. Es ist jedoch von essenzieller Bedeutung, den zu analysierenden Datensatz im Vorfeld eingehend zu prüfen, um die am besten geeigneten Verfahren zu identifizieren. Die sorgfältige Auswahl von Datenqualitätsmetriken spielt eine entscheidende Rolle, um Veränderungen am Datensatz zu quantifizieren. Die Wahl der Metriken sollte dabei eng mit den angewendeten Datenvorverarbeitungsverfahren verknüpft sein, da unterschiedliche Vorverarbeitungsschritte variierende Auswirkungen auf die Datenqualität haben. Eine präzise Abstimmung zwischen den gewählten Metriken und den spezifischen Vorverarbeitungsschritten ist daher unerlässlich. Insgesamt zeigt die Arbeit, dass ein systematischer und ganzheitlicher Ansatz zur Datenvorverarbeitung in Verbindung mit einer geeigneten Auswahl von Datenqualitätsmetriken entscheidend dazu beiträgt, die Qualität von Daten zu verbessern und sie für weiterführende Analysen in einem produzierenden Umfeld nutzbar zu machen.

5.2 Implikationen

Die vorherige Auseinandersetzung mit Datenqualitätsmetriken und ihrer zielgerichteten Anwendung in produzierenden Unternehmen zieht verschiedene Implikationen nach sich. Diese Arbeit leistet einen Beitrag im Hinblick auf die bisherige Forschungslücke im Bereich der Datenqualität, indem sie sich umfassend mit einer breiten Palette von Datenqualitätsdimensionen auseinandersetzt. Dabei wurden 42 Datenqualitätsmetriken für 20 Datenqualitätsdimensionen

identifiziert. Dies zeigt die Vielfalt, aber auch die Herausforderungen im Bereich der Datenqualitätsmessung auf. Die Feststellung, dass es bisher kaum Beiträge gibt, die sich ausführlich mit Metriken für eine derart umfassende Anzahl von Dimensionen beschäftigen, hebt die Arbeit in der Datenqualitätsforschung hervor. Dieser Mangel an Forschung betont die Relevanz der vorliegenden Arbeit für die theoretische Weiterentwicklung des Verständnisses von Datenqualität in komplexen industriellen Umgebungen.

Weiterhin führt die Evaluierung der Datenqualitätsmetriken in dieser Arbeit zu praxisrelevanten Implikationen, die direkt auf spezifische Anwendungsbereiche in produzierenden Unternehmen abzielen. Die Nutzung der empfohlenen Metriken aus Abbildung 9 sollten von produzierenden Unternehmen in Erwägung gezogen werden, da Datenqualität eine fundamentale Grundlage für die unternehmerische Entscheidungsfindung darstellt. Die klare Anwendbarkeit der vorgestellten Formeln für verschiedene Datenqualitätsdimensionen ermöglicht es Unternehmen weiterhin, gezielt die Qualität ihrer Daten zu bewerten und zu verbessern. Angesichts der sich ständig wandelnden Produktionsumgebungen und der zunehmenden Bedeutung datengesteuerter Prozesse wird dringend empfohlen, die vorgestellten Metriken in der Praxis zu nutzen. Die Anpassbarkeit dieser Metriken ermöglicht produzierenden Unternehmen eine maßgeschneiderte Nutzung entsprechend ihren individuellen Anforderungen. Durch die Integration dieser Metriken in Datenmanagementstrategien können Entscheidungsträger die Datenqualität verbessern und nachhaltig datengetriebene Entscheidungsfindung fördern. Es wird daher empfohlen, diese Metriken aktiv in bestehenden Modellen zu implementieren.

5.3 Eruierung von Limitationen und Forschungsausblick

Im folgenden Abschnitt werden sowohl die Limitationen dieser Arbeit als auch Perspektiven für zukünftige Forschungsbemühungen beleuchtet. Ein zentraler Aspekt ist die begrenzte Berücksichtigung von Datenqualitätsdimensionen. Die Ausrichtung dieser Studie konzentriert sich hauptsächlich auf die in der Arbeit von Sidi et al. (2012) beschriebene Dimension. Es ist jedoch wichtig zu beachten, dass es weitere relevante Datenqualitätsdimensionen gibt, die in dieser speziellen Auflistung nicht explizit berücksichtigt wurden. Zukünftige Forschungsarbeiten könnten daher von Nutzen sein, indem sie den Fokus auf eine breitere Palette von Datenqualitätsdimensionen erweitern. Dimensionen wie *Editability*, *Clarity* und *Variety of Data Sources* sind Beispiele für solch vernachlässigten Bereiche, die eine vertiefte Betrachtung und

Evaluierung in Bezug auf Datenqualität verdienen. Eine umfassendere Analyse dieser zusätzlichen Dimensionen würde einen detaillierteren Einblick in die Vielschichtigkeit der Datenqualitätsmetriken in produzierenden Umgebungen ermöglichen und somit die Basis für präzisere und umfassendere Empfehlungen schaffen.

Ein weiterer wichtiger Aspekt ist die begrenzte Generalisierbarkeit der vorgestellten Metriken auf verschiedene industrielle Kontexte. Die spezifischen Anforderungen und Prozesse können zwischen Unternehmen variieren, was die Notwendigkeit weiterer Untersuchungen zur Validierung und Anpassung der Metriken an unterschiedliche Branchen unterstreicht.

Des Weiteren sollte in kommenden Arbeiten der Einfluss technologischer Entwicklungen, wie der verstärkten Nutzung von KI und maschinellem Lernen, auf die Datenqualität in produzierenden Unternehmen genauer erforscht werden. Dies ist besonders relevant, da solche Technologien zunehmend in datenintensiven Produktionsprozessen eingesetzt werden.

Ein vielversprechender Forschungsausblick liegt auch in der Integration von Echtzeit-Datenqualitätsüberwachungssystemen. Solche Systeme könnten dazu beitragen, nicht nur retrospektiv, sondern auch proaktiv auf Datenqualitätsprobleme zu reagieren und somit die Kontinuität der datenbasierten Entscheidungsfindung zu gewährleisten.

Insgesamt bieten die identifizierten Limitationen und Forschungsperspektiven einen klaren Handlungsrahmen für zukünftige Arbeiten im Bereich der Datenqualität in produzierenden Unternehmen. Durch die Adressierung dieser Aspekte können zukünftige Forschungen dazu beitragen, die Effektivität und Anwendbarkeit von Datenqualitätsmetriken zu optimieren und die sich ständig wandelnden Anforderungen der industriellen Landschaft zu berücksichtigen.

Literaturverzeichnis

Aggarwal, C. C. (2015): Data Mining. The Textbook. 1. Aufl. Cham: Springer Cham.

Alpar, P.; Winkelsträter, S. (2014): Assessment of data quality in accounting data with association rules. In: *Expert Systems with Applications* 41 (5), S. 2259–2268.

Apel, D.; Behme, W.; Eberlein, R.; Merighi, C. (2015): Datenqualität erfolgreich steuern. Praxislösungen für Business-Intelligence-Projekte. 3. Aufl. Heidelberg: Carl Hanser Verlag GmbH & Co. KG.

Azeroual, O. (2022): Untersuchung der Datenqualität in FIS. In: O. Azeroual (Hg.): Untersuchungen zur Datenqualität und Nutzerakzeptanz von Forschungsinformationssystemen: Framework zur Überwachung und Verbesserung der Qualität von Forschungsinformationen. Wiesbaden: Springer Fachmedien Wiesbaden, S. 49–148.

Azeroual, O.; Saake, G.; Wastl, J. (2018): Data measurement in research information systems: metrics for the evaluation of data quality. In: *Scientometrics* 115, S. 1271–1290.

Azevedo, A.; Santos, M. (2008): KDD, SEMMA and CRISP-DM: a parallel overview. In: *IS-CAP - Sistemas de Informação - Comunicações em eventos científicos*, S. 182–185.

Baars, H.; Kemper, H. G. (2021): Business Intelligence & Analytics. Grundlagen und praktische Anwendungen: Ansätze der IT-basierten Entscheidungsunterstützung. 4. Aufl. Wiesbaden, Heidelberg: Springer Vieweg.

Baier, A. (2020): Entwicklung eines Kennzahlencockpits für Supply Chain Management am Beispiel eines Unternehmens der Elektronikbranche. Montanuniversität Leoben, Leoben.

Ballou, D. P.; Pazer, Harold L. (1985): Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems. In: *Management Science* 31 (2), S. 150–162.

Balzer, V.; Schrodi, T.; Wiendahl, H. H. (2020): Strategien und Struktur produzierender Unternehmen. In: T. Bauernhansl (Hg.): Fabrikbetriebslehre 1: Management in der Produktion. Berlin, Heidelberg: Springer Berlin Heidelberg, S. 35–66.

Bange, C.; Janoschek, N. (2014): Big Data Analytics-Auf dem Weg zur datengetriebenen Wirtschaft. Würzburg: BARC Institut.

Batini, C.; Cappiello, C.; Francalanci, C.; Maurino, A. (2009): Methodologies for Data Quality Assessment and Improvement. In: *ACM Comput. Surv.* 41 (3).

Batini, C.; Scannapieco, M. (2006): Data Quality: Concepts, Methodologies and Techniques. Heidelberg: Springer Berlin.

Bauer, D; Maurer, T; Henkel, C.; Bildstein, A. (2017): Big-Data-Analytik: Datenbasierte Optimierung produzierender Unternehmen. Stuttgart: Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA.

Behkamal, B.; Kahani, M.; Bagheri, E.; Jeremic, Z. (2014): A Metrics-Driven Approach for Quality Assessment of Linked Open Data. In: *Journal of Theoretical and Applied Electronic Commerce Research* 9 (2), S. 64–79.

Bode, J. (1997): Der Informationsbegriff in der Betriebswirtschaftslehre. In: *zfbf (Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung)* 49 (5), S. 449–468.

Bonney, W.; Scobbie, D.; Nind T. (2014): Profiling Clinical Datasets for Data Quality Assessment and Improvement. In: *EventBCS Health Informatics Scotland*, S. 1–8.

Bramer, M. (2016): Principles of Data Mining. 3. Aufl. London: Springer London (Undergraduate Topics in Computer Science).

Brown, R. J. C.; Woods, P. T. (2014): Proposals for new data quality objectives to underpin ambient air quality monitoring networks. In: *Accreditation and Quality Assurance* 19 (6), S. 465–471.

Budach, L.; Feuerpfeil, M.; Ihde, N.; Nathansen, A.; Noack, N.; Patzlaff, H. et al. (2022): The Effects of Data Quality on Machine Learning Performance.

Caballero, I.; Gualo, F.; Rodríguez, M.; Piattini, M. (2022): BR4DQ: A methodology for grouping business rules for data quality evaluation. In: *Information Systems* 109, S. 102058.

Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Daimlerchrysler, C. et al. (2000): CRISP-DM 1.0: Step-by-step data mining guide. In: SPSS inc, 9(13), S. 1–72.

Chernov, Y. (2022): Data Quality Measurement Based on Domain-Specific Information. In: B. Santhosh Kumar (Hg.): Data Integrity and Data Governance. Rijeka: IntechOpen, 20-40.

Chung, Y.; Krishnan, S.; Kraska, T. (2017): A Data Quality Metric (DQM): How to Estimate The Number of Undetected Errors in Data Sets. In: *Proceedings of the VLDB Endowment* (10), S. 1094–1105.

Deuse, J.; Erohin, O.; Lieber, D. (2014): Wissensentdeckung in vernetzten, industriellen Datenbeständen. In: *Industrie4* (4), S. 373–395.

Dilda, V.; Lapo Mori, M.; Noterdaeme, O.; Schmitz, C.; van Niel, J. (2017): Manufacturing: Analytics unleashes productivity and profitability. In: *McKinsey & Company*.

Efimova, O. V.; Igolnikov, B. V.; Isakov, M. P.; Dmitrieva, E. I. (2021): Data Quality and Standardization for Effective Use of Digital Platforms. In: 2021 International Conference on Quality Management, Transport and Information Security, Information Technologies (IT&QM&IS), S. 282–285.

Ehrlinger, L.; Wöß, W. (2022): A Survey of Data Quality Measurement and Monitoring Tools. In: *Frontiers in Big Data* 5, Artikel 850611.

Eickelmann, M.; Wiegand, M.; Konrad, B.; Deuse, J. (2015): Die Bedeutung von Data-Mining im Kontext von Industrie 4.0 (11), S. 738–743.

Eigner, M.; August, U.; Schmichm M. (2016): Smarte Produkte erfordern ein Umdenken bei Produktstrukturen und Prozessen. In: *Siemens PLM Software*.

El Alaoui, I.; Youssef, G.; Messoussi, R.: Big Data Quality Metrics for Sentiment Analysis Approaches. In: Proceedings of the 2019 International Conference on Big Data Engineering.

Elouataoui, W.; El Alaoui, I.; el Mendili, S.; Youssef, G. (2022): An Advanced Big Data Quality Framework Based on Weighted Metrics. In: *Big Data and Cognitive Computing* 13, S. 36–43.

Fadlallah, H.; Kilany, R.; Dhayne, H.; El Haddad, R.; Haque, R.; Taher, Y.; Jaber, A. (2023): Context-Aware Big Data Quality Assessment: A Scoping Review. In: *Journal of Data and Information Quality* 15 (3), Artikel 25.

Fink, A. (2014): Conducting research literature reviews: From the Internet to paper, 3. 4. Aufl. Thousand Oaks, CA, US: Sage Publications, Inc.

Frehe, V.; Adelmeyer, T.; Teuteberg, F. (2016): Eine Balanced Scorecard für das systematische Datenqualitätsmanagement im Kontext von Big Data. In: *Multikonferenz Wirtschaftsinformatik*, S. 143–154.

Freudiger, J.; Rane, S.; Brito, A.; Uzun, E.: Privacy Preserving Data Quality Assessment for High-Fidelity Data Sharing. In: Proceedings of the 2014 ACM Workshop on Information Sharing & Collaborative Security, S. 21–29.

Frick, D.; Gadatsch, A.; Kaufmann, J.; Lankes, B.; Quix, C.; Schmidt, A.; Schmitz, U. (Hg.) (2021): Data Science: Konzepte, Erfahrungen, Fallstudien und Praxis. Wiesbaden: Springer Fachmedien Wiesbaden.

Gabriel, R.; Gluchowski, P.; Pastwa, A. (2009): Datawarehouse & Data Mining. Herdecke: W3L GmbH.

Galvagno, M.; Dalli, D.; Gummesson, E.; Mele, C.; Polese, F. (2014): Theory of value co-creation: a systematic literature review. In: *Managing Service Quality* 24 (6), S. 643–683.

García-Gil, D.; Luengo, J.; García, S.; Herrera, F. (2019): Enabling Smart Data: Noise filtering in Big Data classification. In: *Information Sciences* 479, S. 135–152.

Gebauer, M.; Windheuser, U. (2011): Strukturierte Datenanalyse, Profiling und Geschäftsregeln. In: K. Hildebrand, M. Gebauer, H. Hinrichs und M. Mielke (Hg.): Daten- und Informationsqualität: Auf dem Weg zur Information Excellence. Wiesbaden: Vieweg+Teubner, S. 88–101.

Gitzel, R.; Subbiah, S.; Ganz, C. (2018): A Data Quality Dashboard for CMMS Data. In: International Conference on Operations Research and Enterprise Systems.

Gitzel, R.; Turrin, S.; Maczey, S.; Wu, S.; Schmitz, B. (2016): A data quality metrics hierarchy for reliability data. In:. Online verfügbar unter https://api.semanticscholar.org/CorpusID:53553663.

Gitzel, R.; Turring, S.; Maczey, S.: A Data Quality Dashboard for Reliability Data. In: *Proceedings of the 9th IMA International Conference on Modelling in Industrial Maintenance and Reliability*, S. 90–97.

Goknil, A.; Phu, N.; Sagar, S.; Politaki, D.; Niavis, H.; Pedersen, K. J. et al. (2023): A Systematic Review of Data Quality in CPS and IoT for Industry 4.0. In: *ACM Comput. Surv.* 55 (14s).

Gröger, C. (2015): Advanced Manufacturing Analytics - Datengetriebene Optimierung von Fertigungsprozessen. Dissertation. Universität Stuttgart, Stuttgart. Institut für Parallele und Verteilte Systeme.

Günther, L. C.; Colangelo, E.; Wiendahl, H. H.; Bauer, C. (2019): Data quality assessment for improved decision-making: a methodology for small and medium-sized enterprises. In: *Procedia Manufacturing* 29, S. 583–591.

Haegemans, T.; Snoeck, M.; Lemahieu, W. (2016): Towards a Precise Definition of Data Accuracy and a Justification for its Measure. In: *MIT International Conference on Information Qualiity*.

Harrington, H. J. (1991): Business process improvement: The breakthrough strategy for total quality, productivity, and competitiveness. New York: McGraw-Hill.

Hassine, S. B.; Clément, D. (2020): Open Data Quality Dimensions and Metrics: State of the Art and Applied Use Cases. In: W. Abramowicz und G. Klein (Hg.): Business Information Systems Workshops. Cham, 2020. Cham: Springer International Publishing, S. 311–323.

Heinrich, B.; Hristova, D.; Klier, M.; Schiller, A.; Szubartowicz, M. (2018): Requirements for Data Quality Metrics. In: *J. Data and Information Quality* 9 (2), S. 1–32.

Heinrich, B.; Klier, M. (2008): Datenqualitätsmetriken für ein ökonomisch orientiertes Qualitätsmanagement. In: *Daten- und Informationsqualität*, S. 47–64.

Heinrich, B.; Klier, M. (2009): Die Messung der Datenqualität im Controlling. In: *Controlling & Management* 53, S. 34–42.

Heinrich, B.; Klier, M. (2015): Metric-Based Data Quality Assessment - Developing and Evaluating a Probability-Based Currency Metric. In: *Decis. Support Syst.* 72 (C), S. 82–96.

Hejazi, A.; Abdolvand, N.; Rajaei Harandi, S. (2017): Assessing the Importance of Data Factors of Data Quality Model in the Business Intelligence Area. In: *International Journal of Trade, Economics and Finance* 8, S. 102–108.

Helmis, S.; Hollmann, R. (Hg.) (2009): Webbasierte Datenintegration: Ansätze zur Messung und Sicherung der Informationsqualität in heterogenen Datenbeständen unter Verwendung eines vollständig webbasierten Werkzeuges. Wiesbaden: Vieweg Teubner.

Heravizadeh, M.; Mendling, J.; Rosemann, M. (2008): Dimensions of Business Processes Quality (QoBP). In: *Lecture Notes in Business Information Processing* 17, S. 80–91.

Hildebrand, K.; Gebauer, M.; Hinrichs, H.; Mielke, M. (Hg.) (2011): Daten- und Informationsqualität: Auf dem Weg zur Information Excellence. Wiesbaden: Vieweg+Teubner.

Hildebrand, K.; Gebauer, M.; Mielke, M. (2021): Daten- und Informationsqualität. Die Grundlage der Digitalisierung. 5. Aufl. Wiesbaden: Springer Vieweg Wiesbaden.

Hinrichs, H. (2001): Datenqualitätsmanagement in Data Warehouse-Umgebungen. In: A. Heuer, F. Leymann und D. Priebe (Hg.): Datenbanksysteme in Büro, Technik und Wissenschaft. Berlin, Heidelberg, 2001. Berlin, Heidelberg: Springer Berlin Heidelberg, S. 187–206.

Hoffmann, M. (2017): Adaptive and Scalable Information Modeling to Enable Autonomous Decision Making for Real-Time Interoperable Factories. Dissertation. RWTH Aachen University, Aachen.

Holland, H. (2020): Big Data. In: H. Holland (Hg.): Digitales Dialogmarketing: Grundlagen, Strategien, Instrumente. Wiesbaden: Springer Fachmedien Wiesbaden, S. 1–23.

Hudasch, M. (1996): Experience in the measurement of short-circuit impedances in high voltage networks; Erfahrungen bei der Messung von Kurzschlussimpedanzen in Hochspannungsnetzen. In: *Elektrizitaetswirtschaft* (95), S. 201–203.

Jarke, M.; Lenzerini, M.; Vassiliou, Y.; Vassiliadis, P. (2003): Fundamentals of Data Warehouses. Heidelberg: Springer Berlin.

Kim, Y. K.; Lee, K. M. (2015): Saliency Score-Based Visualization for Data Quality Evaluation. In: *IJFIS* 15 (4), S. 289–294.

King, W. R.; Epstein, B. J. (1983): Assesing in information system value: An experimental study. In: *Decision Sciences* 14 (1), S. 34–45.

Kletti, J.; Deisenroth, R.; Diesner, M.; Kletti, W.; Lübbert, J. P.; Schumacher, J.; Strebel, T. (2015): MES als Werkzeug für die perfekte Produktion. In: J. Kletti (Hg.): MES - Manufacturing Execution System: Moderne Informationstechnologie unterstützt die Wertschöpfung. Berlin, Heidelberg: Springer Berlin Heidelberg, S. 19–29.

Knight, S.; Burn, J. (2005): Developing a Framework for Assessing Information Quality on the World Wide Web. In: *Informing Science Journal* 8, S. 159–172.

Krechting, Denis Phillip (2021): Methodennavigator für Business-Analytics produzierender Unternehmen. Dissertation. RWTH Aachen University, Aachen.

Kureljusic, M.; Karger, E. (2022): Data Preprocessing as a Service – Outsourcing der Datenvorverarbeitung für KI-Modelle mithilfe einer digitalen Plattform. In: *Informatik Spektrum* 45 (1), S. 13–19.

Laranjeiro, N.; Soydemir, S. N.; Bernardino, J. (2015): A Survey on Data Quality: Classifying Poor Data. In: 2015 IEEE 21st Pacific Rim International Symposium on Dependable Computing (PRDC), S. 179–188.

Lidiansa, M. (2014): Developing Data Quality Metrics for a Product Master Data Model. Delft: Elsevier B.V.

Lieber, D.; Erohin, O.; Deuse, J. (2013): Wissensentdeckung im industriellen Kontext. In: *Herausforderungen und Anwendungsbeispiele* 108 (6), S. 388–393.

Makhoul, N. (2022): Review of data quality indicators and metrics, and suggestions for indicators and metrics for structural health monitoring. In: *Advances in Bridge Engineering* 3 (1), S. 17.

Man, Y.; Wei, L.; Gang, H.; Juntao, G. (2010): A Noval Data Quality Controlling and Assessing Model Based on Rules. In: 2010 Third International Symposium on Electronic Commerce and Security. 2010 Third International Symposium on Electronic Commerce and Security, S. 29–32.

Maydanchik, A. (2007): Data quality assessment. New York: Technics publications.

McAfee, A.; Brynjolfsson, E. (2012): Big data: the management revolution. In: *Harvard business review* 90 (10), 60-6, 68, 128.

McGilvray, D. (Hg.) (2021): Executing Data Quality Projects. 2. Aufl. London: Academic Press.

Miller, H. (1996): The multiple dimensions of information quality. In: *Information Systems Management* 13 (2), S. 79–82.

Moaawad, M. R.; Mokhtar, H. M.; Al Feel, H. T. (2017): On-The-Fly Academic Linked Data Integration. In: Proceedings of the International Conference on Compute and Data Analysis. New York, NY, USA: Association for Computing Machinery (ICCDA '17), S. 114–122.

Müller, J. (2000): Transformation operativer Daten zur Nutzung im Data Warehouse. Wiesbaden: Deutscher Universitätsverlag Wiesbaden.

North, K. (Hg.) (2021): Wissensorientierte Unternehmensführung: Wissensmanagement im digitalen Wandel. Wiesbaden: Springer Fachmedien Wiesbaden.

Okoli, C. (2015): A Guide to Conducting a Systematic Literature Review of Information Systems Research. In: *SSRN Electronic Journal* 10.

Peralta, C. (2006): Data quality evaluation in data integration systems. Dissertation. Universität Versailles, Versailles. Computer Science.

Pipino, L.; Lee, Y.; Wang, R. (2003): Data Quality Assessment. In: *Communications of the ACM* 45, S. 211–218.

Plaue, M. (2021): Data Science: Grundlagen, Statistik und maschinelles Lernen. Heidelberg: Springer Spektrum Berlin.

Porter, M.; Heppelmann, J. (2014): Wie smarte Produkte den Wettbewerb verändern. In: *Harvard Business Manager*.

Pradhan, S.; Tungal, S. (2021): Quality Attributes of Data in Distributed Deep Learning Architectures. Gothenburg: University of Gothenburg.

Probst, G.; Raub, S.; Romhardt, K. (2012): Wissen managen. Wissen managen. 7. Aufl. Wiesbaden: Gabler Verlag.

Quix, C. (2021): Big-Data-Technologien. In: D. Frick, A. Gadatsch, J. Kaufmann, B. Lankes, C. Quix, A. Schmidt und U. Schmitz (Hg.): Data Science: Konzepte, Erfahrungen, Fallstudien und Praxis. Wiesbaden: Springer Fachmedien Wiesbaden, S. 133–148.

Raghavendra, S. (2017): Relevance of the two adjusting screws in data analytics: data quality and optimization of algorithms. In:

Rahm, R.; Do, H. (2000): Data Cleaning: Problems and Current Approaches. In: *IEEE Data Eng. Bull.* 23, S. 3–13.

Rapp, J. (2020): Datenqualitätsmetriken zur Unterstützung von Domänenexperten bei interaktiven Analysen. Universität Stuttgart, Stuttgart.

Saleh, H. (2018): Machine learning fundamentals. 1. Aufl. Birmingham: Packt Publishing.

Saroja, S. (2016): Measurement of the quality of structured and unstructured data accumulating in the product life cycle in a data quality dashboard. Dissertation, Stuttgart.

Schmitt, R.; Kurzhals, R.; Ellerich, M.; Nilgen, G.; Schlegel, P.; Dietrich, E. et al. (2020): Predictive Quality - Data Analytics in produzierenden Unternehmen. In: *Internet of Production - Turning Data into Value*, S. 226–253.

Schröer, C.; Kruse, F.; Gómez, J. M. (2021): A Systematic Literature Review on Applying CRISP-DM Process Model. In: *Procedia Computer Science* 181, S. 526–534.

Schuh, G.; Prote, J. P.; Sauermann, F.; Schmitz, S. (2019): Production Analytics. In: *Produktionsdaten anwendergerecht auswerten* 114 (9), S. 588–591.

Schuh, G.; Schacht, M.; Holst, L. (2023): Digitaler Schatten der Kundeninteraktionen produzierender Unternehmen. In: M. Bruhn und K. Hadwich (Hg.): Gestaltung des Wandels im Dienstleistungsmanagement. Kundenperspektive - Anbieterperspektive - Mitarbeiterperspektive. 2. Aufl. Wiesbaden: Springer Fachmedien Wiesbaden, S. 311–335.

Schuh, G.; Stich, V.; Basse, F.; Franzkoch, B.; Harzenetter, F.; Luckert, M. et al. (2017): Change Request im Produktionsbetrieb. 1. Aufl. Aachen: Apprimus.

Sebastian-Coleman, L. (2012): Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework. San Franzisco: Morgan Kaufmann Publishers Inc.

Seufert, A. (2016): Die Digitalisierung als Herausforderung für Unternehmen: Status Quo, Chancen und Herausforderungen im Umfeld BI & Big Data. In: D. Fasel und A. Meier (Hg.): Big Data: Grundlagen, Systeme und Nutzungspotenziale. Wiesbaden: Springer Fachmedien Wiesbaden, S. 39–57.

Sicari, S.; Cappiello, C.; Pellegrini, F.; Miorandi, D.; Coen-Porisini, A. (2016): A security-and quality-aware system architecture for Internet of Things. In: *Information Systems Frontiers* 18 (4), S. 665–677.

Sidi, F.; Shariat Panahy, P. H.; Affendey, L. S.; Jabar, M. A.; Ibrahim, H.; Mustapha A. (2012): Data quality: A survey of data quality dimensions. In: 2012 International Conference on Information Retrieval & Knowledge Management. 2012 International Conference on Information Retrieval & Knowledge Management, S. 300–304.

Smith, F. (2006): Data science as an academic discipline. In: *Data Science Journal* 5, S. 163–164.

Steiner, R. (2017): Grundkurs Relationale Datenbanken. Einführung in die Praxis der Datenbankentwicklung für Ausbildung, Studium und IT-Beruf. 9. Aufl. Wiesbaden: Springer Vieweg Wiesbaden.

Stoetzer, M. W. (2020): Fehlende Datenwerte/Missing Values. In: M. W. Stoetzer (Hg.): Regressionsanalyse in der empirischen Wirtschafts- und Sozialforschung: Komplexe Verfahren. 2. Aufl. Berlin, Heidelberg: Springer Berlin Heidelberg, S. 297–357.

Taggart, J.; Liaw, S. T.; Yu, Hairong (2015): Structured data quality reports to improve EHR data quality. In: *International journal of medical informatics* 84 (12), S. 1094–1098.

Taleb, I.; Kassabi, H. T. E.; Serhani, M. A.; Dssouli, R.; Bouhaddioui, C (2016): Big Data Quality: A Quality Dimensions Evaluation. In: 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress, S. 759–765.

Teuteberg, F.; Freundlieb, M. (2009): Systematisches Datenqualitätsmanagement. In: *Das Wirtschaftsstudium: wisu* 38 (8), S. 1140–1147.

Trunzer, E.; Weiß, I.; Pötter, T.; Vermum, C.; Odenweller, M.; Schütz, D.; Vogel-Heuser, B. (2019): Big Data trifft Produktion. Neun Pfeiler der industriellen Smart-Data-Analyse. In: *atp magazin* (61).

Vetrò, Antonio; Canova, Lorenzo; Torchiano, Marco; Minotas, Camilo Orozco; Iemma, Raimondo; Morando, Federico (2016): Open data quality measurement framework: Definition

and application to Open Government Data. In: *Government Information Quarterly* 33 (2), S. 325–337.

Vollmuth, J. H.; Zwettler, R. (2021): Kennzahlen. Freiburg: Haufe.

Wand, Y.; Wang, R. Y. (1996): Anchoring Data Quality Dimensions in Ontological Foundations. In: *Commun. ACM* 39 (11), S. 86–95.

Wang, R. Y.; Strong, D. M. (1996): Beyond Accuracy: What Data Quality Means to Data Consumers. In: *Journal of Management Information Systems* 12 (4), S. 5–33.

Wenzel, S.; Jessen, U.; Bernhard, J. (2005): Classifications and conventions structure the handling of models within the Digital Factory. In: *Computers in Industry* 56 (4), S. 334–346.

Windelband, L.; Fenzl, C.; Hunecker, F.; Riehle, T.; Spöttl, G.; Staedtler, H. et al. (2011): Qualifikationsanforderungen durch das Internet der Dinge in der Logistik. Bremen: FreQueNz.

Windthorst, K. (Hg.) (2020): Herausforderungen für Familienunternehmen: Digitalisierung, Internationalisierung, Governance. Baden-Baden. 1. Aufl.: Nomos Verlagsgesellschaft mbH & Co. KG.

Witte, F. (2018): Metriken für das Testreporting. 1. Aufl. Wiesbaden: Springer Vieweg Wiesbaden.

Witzenleiter, M. (2023): Datenstrategie und die Bedeutung von Daten in Unternehmen. In: M. Witzenleiter (Hg.): Quick Guide Product Analytics: Wie Sie mit Systemen wie Google Analytics 4 und Co. mehr über Ihre Nutzer und deren Produktakzeptanz lernen können. Wiesbaden: Springer Fachmedien Wiesbaden, S. 69–79.

Würthele, V. G. (2003): Datenqualitätsmetrik für Informationsprozesse. Datenqualitätsmanagement mittels ganzheitlicher Messung der Datenqualität. Dissertation. Eidgenössische Technische Hochschule ETH, Zürich.

Xu, D.; Zhang, Z.; Shi, J. (2022): A Data Quality Assessment and Control Method in Multiple Products Manufacturing Process. In: 2022 5th International Conference on Data Science and Information Technology (DSIT), S. 1–5.

Yang, J.; Zhao, C.; Xing, C. (2019): Big Data Market Optimization Pricing Model Based on Data Quality. In: *Complexity* 2019, S. 5964068.

Zaveri, A.; Rula, A.; Maurino, A.; Pietrobon, R.; Lehmann, J.; Auer, S. (2015): Quality assessment for Linked Data: A Survey. In: *Semantic Web* 7, S. 63–93.

Anhang

Anhang A: Datenqualitätsdimensionen

Dimensionen	Definitionen
Accessibility (Zugänglichkeit)	Das Ausmaß, in dem Informationen verfügbar sind oder leicht und schnell abgerufen werden können (Wang und Strong 1996).
Accuracy (Genauigkeit)	Daten sind genau, wenn die in der Datenbank gespeicherten Datenwerte den realen Werten entsprechen (Batini et al. 2009; Ballou und Pazer 1985). Es bezieht sich auf den Grad, in dem Daten korrekt, zuverlässig und zertifiziert sind (Wang und Strong 1996).
Amount of data (Datenmenge)	Das Ausmaß, in dem die Menge oder das Volumen der verfügbaren Daten für die aktuelle Aufgabe angemessen ist (Wang und Strong 1996).
Appropriate amount of data (Angemessene Datenmenge)	Das Ausmaß, in dem das Datenvolumen für die aktuelle Aufgabe angemessen ist (Pipino et al. 2003)
Availability (Verfügbarkeit)	Das Ausmaß, in dem Informationen physisch zugänglich sind (Knight und Burn 2005).
Believability (Glaubwürdigkeit)	Das Ausmaß, in dem Informationen als wahr und glaubwürdig angesehen werden (Wang und Strong 1996).
Completeness (Vollständigkeit)	Die Fähigkeit eines Informationssystems, jeden sinnvollen Zustand des dargestellten realen Weltsystems zu repräsentieren (Batini et al. 2009; Wand und Wang 1996). Es bezieht sich darauf, inwieweit Daten ausreichende Breite, Tiefe und

	Reichweite für die aktuelle Aufgabe haben (Wang und Strong 1996).
Concise (Kompaktheit)	Das Ausmaß, in dem Informationen kompakt dargestellt werden, ohne überwältigend zu sein (d.h., prägnant in der Darstellung, aber dennoch vollständig und auf den Punkt gebracht) (Wang und Strong 1996).
Consistency (Konsistenz)	Das Ausmaß, in dem Daten im gleichen Format präsentiert werden und mit früheren Daten kompatibel sind (Wang und Strong 1996). Es bezieht sich auf die Verletzung semantischer Regeln, die für die Menge der Daten definiert sind (Batini et al. 2009).
Consistency and Synchronization (Konsistenz und Synchronisation)	Ein Maß für die Äquivalenz von Informationen, die in verschiedenen Datenspeichern, Anwendungen und Systemen verwendet werden, sowie für die Prozesse zur Herstellung von äquivalenten Daten (McGilvray 2021).
Consistent Representation (Konsistente Darstellung)	**D**as Ausmaß, in dem Daten im gleichen Format präsentiert werden (Pipino et al. 2003).
Currency (Aktualität)	Die Aktualität ist der Grad der Aktualität eines Datums. Ein Datenwert ist dann aktuell, wenn er trotz möglicher Diskrepanzen durch zeitbedingte Änderungen des richtigen Wertes korrekt ist (Batini et al. 2009; Jarke et al. 2003).
Data Coverage (Datenabdeckung)	Ein Maß für die Verfügbarkeit und Vollständigkeit von Daten im Vergleich zum Gesamtdatenuniversum oder der Interessengruppe (McGilvray 2021).
Data Decay (Datenverfall)	Ein Maß für die Rate negativer Veränderungen von Daten (McGilvray 2021).

Data integrity fundamentals (Datenintegrität Grundlagen)	Ein Maß für das Vorhandensein, die Gültigkeit, die Struktur, den Inhalt und andere grundlegende Eigenschaften der Daten (McGilvray 2021).
Data specification (Datenangaben)	Ein Maß für das Vorhandensein, die Vollständigkeit, die Qualität und die Dokumentation von Datenstandards, Datenmodellen, Geschäftsregeln, Metadaten und Referenzdaten (McGilvray 2021).
Duplication (Duplikation)	Ein Maß für unerwünschte Duplikate, die in oder zwischen Systemen für ein bestimmtes Feld, einen Datensatz oder eine Datensammlung vorhanden sind (McGilvray 2021).
Ease of Manipulation (Leichte Manipulierbarkeit)	Das Ausmaß, in dem Daten leicht manipuliert und auf verschiedene Arten angewendet werden können (Pipino et al. 2003).
Ease of Use and maintainability (Benutzerfreundlichkeit und Wartbarkeit)	Ein Maß für den Grad, in dem Daten abgerufen und verwendet werden können und der Grad, in dem Daten aktualisiert, gewartet und verwaltet werden können (McGilvray 2021).
Effectiveness (Effektivität)	Die Fähigkeit der Funktion, Benutzern zu ermöglichen, in einem bestimmten Nutzungskontext festgelegte Ziele mit Genauigkeit und Vollständigkeit zu erreichen (Batini et al. 2009).
Efficiency (Effizienz)	Das Ausmaß, in dem Daten schnell die Informationsbedürfnisse für die aktuelle Aufgabe erfüllen können (Wang und Strong 1996).
Free of error (Fehlerfreiheit)	Das Ausmaß, in dem Daten korrekt und zuverlässig sind (Pipino et al. 2003).

Freshness (Frische)	Frische repräsentiert eine Familie von Qualitätsfaktoren, wobei jeder einen bestimmten Aspekt der Aktualität darstellt und Metriken dafür aufweist (Peralta 2006)].
Interpretability (Interpretierbarkeit)	Das Ausmaß, in dem Daten geeignete Sprachen, Symbole und Einheiten verwenden und die Definitionen klar sind (Pipino et al. 2003).
Learnability (Lernfähigkeit)	Es bedeutet die Fähigkeit der Funktion, es dem Benutzer zu ermöglichen, sie zu erlernen (Heravizadeh et al. 2008).
Navigation (Navigation)	Das Ausmaß, in dem Daten leicht gefunden und verknüpft werden können (Knight und Burn 2005).
Objectively / Objectivity (Objektivität)	Das Ausmaß, in dem Informationen unvoreingenommen und unparteiisch sind (Wang und Strong 1996).
Presentation Quality (Qualität der Präsentation)	Ein Maß dafür, wie Informationen präsentiert, gesammelt und genutzt werden. Format und Erscheinungsbild unterstützen die angemessene Verwendung von Informationen (McGilvray 2021).
Relevancy (Relevanz)	Das Ausmaß, in dem Informationen für die aktuelle Aufgabe anwendbar und hilfreich sind (Wang und Strong 1996).
Reliability (Zuverlässigkeit)	Das Ausmaß, in dem Informationen korrekt und zuverlässig sind (Wang und Strong 1996). Es ist die Fähigkeit der Funktion, ein bestimmtes Leistungsniveau aufrechtzuerhalten, wenn sie unter bestimmten Bedingungen verwendet wird (Heravizadeh et al. 2008).
Reputation (Reputation)	Das Ausmaß, in dem Informationen in Bezug auf Herkunft oder Inhalt hoch angesehen werden (Wang und Strong 1996).

Safety (Sicherheit)	Fähigkeit der Funktion, akzeptable Risikolevel hinsichtlich Schäden für Menschen, Prozesse, Eigentum oder die Umwelt zu erreichen (Heravizadeh et al. 2008)
Security (Sicherheit)	Das Ausmaß, in dem der Zugriff auf Informationen angemessen eingeschränkt ist, um deren Sicherheit zu gewährleisten (Wang und Strong 1996).
Timeliness (Aktualität)	Das Ausmaß, in dem das Alter der Daten für die aktuelle Aufgabe angemessen ist (Wang und Strong 1996). Aktualität bezieht sich ausschließlich auf die Verzögerung zwischen einer Änderung des realen Weltzustands und der daraus resultierenden Änderung des Informationszustands des Systems (Batini et al. 2009; Wand und Wang 1996).
Timeliness and Availability (Aktualität und Verfügbarkeit)	Ein Maß für den Grad, in dem Daten aktuell und verfügbar sind und wie spezifiziert und im erwarteten Zeitrahmen verwendet werden können (McGilvray 2021).
Transactability (Transaktionsfähigkeit)	Ein Maß für den Grad, in dem Daten die gewünschte Geschäftstransaktion oder das gewünschte Ergebnis erzeugen (McGilvray 2021).
Understandability (Verständlichkeit)	Das Ausmaß, in dem Daten ohne Mehrdeutigkeiten klar und leicht verständlich sind (Wang und Strong 1996).
Useability (Benutzbarkeit)	Das Ausmaß, in dem Informationen klar und einfach verwendet werden können (Knight und Burn 2005).
Useful (Nützlichkeit)	Das Ausmaß, in dem Informationen für die aktuelle Aufgabe anwendbar und hilfreich sind (Wang und Strong 1996).
Value added (Mehrwert)	Das Ausmaß, in dem Informationen von Nutzen sind und Vorteile durch ihre Verwendung bieten (Wang und Strong 1996).

Anhang B: Literaturtabelle der Systematischen Literaturrecherche

ID	Dimension	Titel	Sprache	Jahr
48	Accessibility	An Advanced Big Data Quality Framework Based on Weighted Metrics	Englisch	2022
111	Accessibility	Data measurement in research information systems: metrics for the evaluation of data quality	Englisch	2018
205	Accessibility	Open Data Quality Dimensions and Metrics: State of the Art and Applied Use Cases	Englisch	2020
20	Accessibility	A Novel Data Quality Metric for Minimality	Englisch	2019
196	Accessibility	Methodology for linked enterprise data quality assessment through information visualizations	Englisch	2019
236	Accessibility	Review of data quality indicators and metrics, and suggestions for indicators and metrics for structural health monitoring	Englisch	2022
74	Accessibility	Big Data Quality Metrics for Sentiment Analysis Approaches	Englisch	2019
127	Accessibility	Daten- und Informationsqualität	Deutsch	2021
243	Accessibility	SemQuire - Assessing the Data Quality of Linked Open Data Sources Based on DQV	Englisch	2018
235	Accuracy	Requirements for Data Quality Metrics	Englisch	3018
25	Accuracy	A Survey of Data Quality Measurement and Monitoring Tools	Englisch	2022
111	Accuracy	Data measurement in research information systems: metrics for the evaluation of data quality	Englisch	2018

48	Accuracy	An Advanced Big Data Quality Framework Based on Weighted Metrics	Englisch	2022
74	Accuracy	Big Data Quality Metrics for Sentiment Analysis Approaches	Englisch	2019
198	Accuracy	Metric-based data quality assessment — Developing and evaluating a probability-based currency metric	Englisch	2015
47	Accuracy	Akzeptanz und Hemmnisse bei der Nutzung und Bewertung von Daten	Deutsch	2023
127	Accuracy	Daten- und Informationsqualität	Deutsch	2020
16	Accuracy	A Method for Developing Data Quality Measures and Metrics for Primary Health Care	Englisch	2016
77	Accuracy	Big Data Quality: A Quality Dimensions Evaluation	Englisch	2016
254	Amount of data	The Challenges of Data Quality and Data Quality Assessment in the Big Data Era	Englisch	2015
11	Amount of data	A Data Quality in Use model for Big Data	Englisch	2016
55	Amount of data	An Overview of Data Quality Frameworks	Englisch	2019
86	Amount of data	Context-aware data quality assessment for big data	Englisch	2018
78	Amount of data	Big Data Quality: A Survey	Englisch	2018
141	Amount of data	Eine Balanced Scorecard für das systematische Datenqualitätsmanagement im Kontext von Big Data	Deutsch	2015

1	Amount of data	20 Years of Data Quality Research: Themes, Trends and Synergies	Englisch	2011
29	Amount of data	A Taxonomy of Data Quality Challenges in Empirical Software Engineering	Englisch	2013
217	Amount of data	Quality Assessment Methodologies for Linked Open Data	Englisch	2013
14	Amount of data	A Maturity Model for Enterprise Data Quality Management	Englisch	2015
85	Appropriate amount of data	Context-aware Big Data Quality Assessment: A Scoping Review	Englisch	2023
236	Appropriate amount of data	Review of data quality indicators and metrics, and suggestions for indicators and metrics for structural health monitoring	Englisch	2022
89	Appropriate amount of data	Data Quality Assessment	Englisch	2019
273	Appropriate amount of data	Towards a context-dependent numerical data quality evaluation framework	Englisch	2018
157	Appropriate amount of data	Exploring the Impact of Data Quality on Business Performance in CRM Systems for Home Appliance Business	Englisch	2023
212	Appropriate amount of data	Profiling Clinical Datasets for Data Quality Assessment and Improvement	Englisch	2014
96	Appropriate amount of data	Data Quality Management with Semantic Technologies	Englisch	2015
133	Appropriate amount of data	Developing a Data Quality Evaluation Framework for Sewer Inspection Data	Englisch	2023

5	Appropriate amount of data	A Data Quality Assessment and Control Method in Multiple Products Manufacturing Process	Englisch	2022
219	Appropriate amount of data	Quality Assessment of Smart Grid Data	Englisch	2018
33	Availability	A data quality metrics hierarchy for reliability data	Englisch	2016
48	Availability	An Advanced Big Data Quality Framework Based on Weighted Metrics	Englisch	2022
25	Availability	A Survey of Data Quality Measurement and Monitoring Tools	Englisch	2022
188	Availability	Luzzu—A Methodology and Framework for Linked Data Quality Assessment	Englisch	2016
7	Availability	A Data Quality Dashboard for Reliability Data	Englisch	2015
74	Availability	Big Data Quality Metrics for Sentiment Analysis Approaches	Englisch	2019
236	Availability	Review of data quality indicators and metrics, and suggestions for indicators and metrics for structural health monitoring	Englisch	2022
264	Availability	The data quality analyzer: A quality control program for seismic data	Englisch	2015
230	Availability	Reducing defects in the datasets of clinical research studies: conformance with data quality metrics	Englisch	2019
127	Availability	Daten- und Informationsqualität	Deutsch	2021

236	Believability	Review of data quality indicators and metrics, and suggestions for indicators and metrics for structural health monitoring	Englisch	2022
53	Believability	An Explainable Machine Learning Approach for Automated Data Quality Assessment and Improvement Processes	Englisch	2022
100	Believability	Data Quality and Trust: Review of Challenges and Opportunities for Data Sharing in IoT	Englisch	2020
143	Believability	Enhancing data quality to mine credible patterns	Englisch	2023
135	Believability	Development of Data Quality Dimensions from User's Perspective Framework	Englisch	2018
278	Believability	Trace It Like You Believe It: Time-Continuous Believability Prediction	Englisch	2021
274	Believability	Towards altruistic data quality assessment for mobile sensing	Englisch	2017
194	Believability	Measuring the Effect of Fraud on Data-Quality Dimensions	Englisch	2023
184	Believability	Linked Data Quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO	Englisch	2018
140	Believability	Eine Balanced Scorecard für das systematische Datenqualitätsmanagement im Kontext von Big Data	Deutsch	2016
283	Completeness	Untersuchung der Datenqualität in FIS	Deutsch	2022

206	Completeness	Open data quality measurement framework: Definition and application to Open Government Data	Englisch	2016
223	Completeness	Quality assessment for Linked Data: A Survey	Englisch	2015
185	Completeness	Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO	Englisch	2018
140	Completeness	Eine Balanced Scorecard für das systematische Datenqualitätsmanagement im Kontext von Big Data	Deutsch	2016
199	Completeness	Metrics-Driven Approach for Quality Assessment of Linked Open Data	Englisch	2014
73	Completeness	Big Data Market Optimization Pricing Model Based on Data Quality,"	Englisch	2019
235	Completeness	Requirements for Data Quality Metrics	Englisch	2022
10	Completeness	A Data Quality Metrics Hierarchy for Reliability Data	Englisch	2016
48	Completeness	An Advanced Big Data Quality Framework Based on Weighted Metrics	Englisch	2022
20	Concise	A Novel Data Quality Metric for Minimality	Englisch	2019
182	Concise	Linked Data Quality	Englisch	2019
144	Concise	Enhancing the Conciseness of Linked Data by Discovering Synonym Predicates	Englisch	2019
188	Concise	Luzzu—A Methodology and Framework for Linked Data Quality Assessment	Englisch	2016

243	Concise	SemQuire - Assessing the Data Quality of Linked Open Data Sources Based on DQV	Englisch	2018
223	Concise	Quality assessment for Linked Data: A Survey	Englisch	2015
3	Concise	A Data Driven Approach for Discovering Data Quality Requirements	Englisch	2014
151	Concise	Evaluating the quality of the LOD cloud: An empirical investigation	Englisch	2018
204	Concise	On-The-Fly Academic Linked Data Integration	Englisch	2017
64	Concise	Assessing the Importance of Data Factors of Data Quality Model in the Business Intelligence Area	Englisch	2017
63	Consistency	Assessing data quality – A probability-based metric for semantic consistency	Englisch	2018
235	Consistency	Requirements for Data Quality Metrics	Englisch	2018
67	Consistency	Assessment of data quality in accounting data with association rules	Englisch	2014
111	Consistency	Data measurement in research information systems: metrics for the evaluation of data quality	Englisch	2018
48	Consistency	An Advanced Big Data Quality Framework Based on Weighted Metrics	Englisch	2022
211	Consistency	Privacy Preserving Data Quality Assessment for High-Fidelity Data Sharing	Englisch	2014
77	Consistency	Big Data Quality: A Quality Dimensions Evaluation	Englisch	2016

25	Consistency	A Survey of Data Quality Measurement and Monitoring Tools	Englisch	2022
17	Consistency	A Metrics-Driven Approach for Quality Assessment of Linked Open Data	Englisch	2014
134	Consistency	Developing a Data Quality Framework on Azure Cloud: Ensuring Accuracy, Completeness, and Consistency	Englisch	2023
156	Consistency and Synchronization	Executing Data Quality Projects	Englisch	2021
40	Consistency and Synchronization	A survey on dataset quality in machine learning	Englisch	2023
150	Consistency and Synchronization	Evaluating quality of event data within event logs an extensible framework	Englisch	2016
37	Consistency and Synchronization	A pattern based approach for data quality requirements modelling	Englisch	2016
220	Consistency and Synchronization	Quality Attributes of Data in Distributed Deep Learning Architectures	Englisch	2021
255	Consistency and Synchronization	The Curse of Dimensionality in Data Quality	Englisch	2013
49	Consistency and Synchronization	An Analysis of Data Quality Dimensions	Englisch	2013
212	Consistent Representation	Profiling Clinical Datasets for Data Quality Assessment and Improvement	Englisch	2014
228	Consistent Representation	Real-Time Data Quality Analysis	Englisch	2020

273	Consistent Representation	Towards a context-dependent numerical data quality evaluation framework	Englisch	2018
241	Consistent Representation	SQL Extensions for domain agnostic data representational consistency	Englisch	2016
89	Consistent Representation	Data Quality Assessment	Englisch	2019
22	Consistent Representation	A Process Pattern Model for Tackling and Improving Big Data Quality	Englisch	2018
285	Consistent Representation	Using Trust as a Measure to Derive Data Quality in Data Shared IoT Deployments	Englisch	2020
7	Consistent Representation	A Data Quality Dashboard for Reliability Data	Englisch	2015
64	Consistent Representation	Assessing the Importance of Data Factors of Data Quality Model in the Business Intelligence Area	Englisch	2017
88	Consistent Representation	Data Quality	Englisch	2018
198	Currency	Metric-based data quality assessment — Developing and evaluating a probability-based currency metric	Englisch	2015
48	Currency	An Advanced Big Data Quality Framework Based on Weighted Metrics	Englisch	2022
116	Currency	Data quality assessment in credit risk management by customized total data quality management approach	Englisch	2016

132	Currency	Developing Data Quality Metrics for a Product Master Data Model	Englisch	2014
111	Currency	Data measurement in research information systems: metrics for the evaluation of data quality	Englisch	2018
140	Currency	Eine Balanced Scorecard für das systematische Datenqualitätsmanagement im Kontext von Big Data	Deutsch	2016
25	Currency	A Survey of Data Quality Measurement and Monitoring Tools	Englisch	2022
74	Currency	Big Data Quality Metrics for Sentiment Analysis Approaches	Englisch	2019
127	Currency	Daten- und Informationsqualität	Deutsch	2021
283	Currency	Untersuchung der Datenqualität in FIS	Deutsch	2022
214	Data Coverage	Proposals for new data quality objectives to underpin ambient air quality monitoring networks	Englisch	2014
105	Data Coverage	Data Quality of Points of Interest in Selected Mapping and Social Media Platforms	Englisch	2018
12	Data Coverage	A Framework for Digital Data Quality Assessment in Digital Biomarker Research	Englisch	2023
97	Data Coverage	Data Quality Measurement Based on Domain-Specific Information	Englisch	2022
103	Data Coverage	Data Quality in Imitation Learning	Englisch	2023
202	Data Coverage	On the Impact of Data Quality on Image Classification Fairness	Englisch	2023

129	Data Coverage	Decision Support System for Implementing Data Quality Projects	Englisch	2016
40	Data Coverage	A survey on dataset quality in machine learning	Englisch	2023
220	Data Coverage	Quality Attributes of Data in Distributed Deep Learning Architectures	Englisch	2021
46	Data Coverage	Air Quality Sensor Networks for Evidence-Based Policy Making: Best Practices for Actionable Insights	Englisch	2022
104	Data Decay	Data Quality in the Era of Big Data: A Global Review	Englisch	2022
240	Data Decay	SNSQ ontology: A domain ontology for SNSs data quality	Englisch	2017
97	Data Decay	Data Quality Measurement Based on Domain-Specific Information	Englisch	2022
150	Data Decay	Evaluating quality of event data within event logs an extensible framework	Englisch	2016
40	Data Decay	A survey on dataset quality in machine learning	Englisch	2023
220	Data Decay	Quality Attributes of Data in Distributed Deep Learning Architectures	Englisch	2021
277	Data Decay	Towards interactive event log forensics: Detecting and quantifying timestamp imperfections	Englisch	2022
190	Data Decay	Master data management: its importance and reasons for failed implementations	Englisch	2020

255	Data Decay	The Curse of Dimensionality in Data Quality	Englisch	2013
49	Data Decay	An Analysis of Data Quality Dimensions	Englisch	2013
51	Data specification	An Analysis of Data Quality Requirements for Machine Learning Development Pipelines Frameworks	Englisch	2023
71	Data specification	BR4DQ: A methodology for grouping business rules for data quality evaluation	Englisch	2022
168	Data specification	Geometric quality assessment of lidar data based on SWATH overlap	Englisch	2016
40	Data specification	A survey on dataset quality in machine learning	Englisch	2023
220	Data specification	Quality Attributes of Data in Distributed Deep Learning Architectures	Englisch	2021
277	Data specification	Towards interactive event log forensics: Detecting and quantifying timestamp imperfections	Englisch	2022
49	Data specification	An Analysis of Data Quality Dimensions	Englisch	2013
139	Data specification	"Does the EU Insurance MDoes the EU Insurance Mediation Directive Help to improve Data Quality?	Englisch	2008
25	Duplication	A Survey of Data Quality Measurement and Monitoring Tools	Englisch	2022
20	Duplication	A Novel Data Quality Metric for Minimality	Englisch	2018
9	Duplication	A Data Quality Metric (DQM): How to Estimate The Number of Undetected Errors in Data Sets	Englisch	2017

137	Duplication	Discovering Data Quality Problems	Englisch	2019
250	Duplication	Structured data quality reports to improve EHR data quality	Englisch	2015
68	Duplication	Automated Continuous Data Quality Measurement with QualIe	Englisch	2018
125	Duplication	Data quality-aware genomic data integration	Englisch	2021
99	Duplication	Data Quality and Standardization for Effective Use of Digital Platforms	Englisch	2021
106	Duplication	Data Quality – The Role of Empiricism	Englisch	2018
192	Duplication	Measurement of the quality of structured and unstructured data accumulating in the product life cycle in a data quality dashboard	Englisch	2017
48	Ease of Manipulation	An Advanced Big Data Quality Framework Based on Weighted Metrics	Englisch	2022
97	Ease of Manipulation	Data Quality Measurement Based on Domain-Specific Information	Englisch	2022
100	Ease of Manipulation	Data Quality and Trust: Review of Challenges and Opportunities for Data Sharing in IoT	Englisch	2020
89	Ease of Manipulation	Data Quality Assessment	Englisch	2019
236	Ease of Manipulation	Review of data quality indicators and metrics, and suggestions for indicators and metrics for structural health monitoring	Englisch	2022
136	Ease of Manipulation	Dimensionality for Big Data Quality A review	Englisch	2021

228	Ease of Manipulation	Real-Time Data Quality Analysis	Englisch	2020
273	Ease of Manipulation	Towards a context-dependent numerical data quality evaluation framework	Englisch	2018
41	Ease of Manipulation	A systematic literature review on data quality assessment	Englisch	2023
42	Ease of Manipulation	A visual understanding of metadata towards an Open Data reuse and exploitation	Englisch	2017
220	Ease of Use and maintainability	Quality Attributes of Data in Distributed Deep Learning Architectures	Englisch	2021
122	Effectiveness	Data quality monitoring and performance metrics of a prospective, population-based observational study of maternal and newborn health in low resource settings	Englisch	2015
115	Effectiveness	Data quality assessment framework to assess electronic medical record data for use in research	Englisch	2016
256	Effectiveness	The Effectiveness of Incentives on Completion Rates, Data Quality, and Nonresponse Bias in a Probability-based Internet Panel Survey	Englisch	2020
55	Effectiveness	An Overview of Data Quality Frameworks	Englisch	2019
74	Effectiveness	Big Data Quality Metrics for Sentiment Analysis Approaches	Englisch	2019
107	Effectiveness	Data Quality: The Role of Empiricism	Englisch	2017

157	Effectiveness	Exploring the Impact of Data Quality on Business Performance in CRM Systems for Home Appliance Business	Englisch	2023
220	Effectiveness	Quality Attributes of Data in Distributed Deep Learning Architectures	Englisch	2021
171	Effectiveness	Identifying and managing data quality requirements: a design science study in the field of automated driving	Englisch	2023
133	Effectiveness	Developing a Data Quality Evaluation Framework for Sewer Inspection Data	Englisch	2023
10	Free of error	A Data Quality Metrics Hierarchy for Reliability Data	Englisch	2016
228	Free of error	Real-Time Data Quality Analysis	Englisch	2020
236	Free of error	Review of data quality indicators and metrics, and suggestions for indicators and metrics for structural health monitoring	Englisch	2022
7	Free of error	A Data Quality Dashboard for Reliability Data	Englisch	2015
5	Free of error	A Data Quality Assessment and Control Method in Multiple Products Manufacturing Process	Englisch	2022
242	Free of error	Saliency Score-Based Visualization for Data Quality Evaluation	Englisch	2015
6	Free of error	A Data Quality Dashboard for CMMS Data	Englisch	2017
114	Free of error	Data quality assessment for improved decision-making: a methodology for small and medium-sized enterprises	Englisch	2019

27	Free of error	A Survey on Data Quality: Classifying Poor Data	Englisch	2015
88	Free of error	Data Quality	Englisch	2015
48	Freshness	An Advanced Big Data Quality Framework Based on Weighted Metrics	Englisch	2022
198	Freshness	Metric-based data quality assessment — Developing and evaluating a probability-based currency metric	Englisch	2015
279	Freshness	Trust evaluation for stream data services based on data quality and service performance	Englisch	2022
25	Freshness	A Survey of Data Quality Measurement and Monitoring Tools	Englisch	2022
87	Freshness	Context-based Data Quality Metrics in Data Warehouse Systems	Englisch	2017
74	Freshness	Big Data Quality Metrics for Sentiment Analysis Approaches	Englisch	2019
91	Freshness	Data Quality Computation For Obsolescence Detection Within Connected Environments	Englisch	2023
259	Freshness	The Impact of Big Data Quality on Sentiment Analysis Approaches	Englisch	2019
83	Freshness	Connecting Semantic Situation Descriptions with Data Quality Evaluations—Towards a Framework of Automatic Thematic Map Evaluation	Englisch	2020
28	Freshness	A Systematic Review of Data Quality in CPS and IoT for Industry 4.0	Englisch	2023

48	Relevancy	An Advanced Big Data Quality Framework Based on Weighted Metrics	Englisch	2022
205	Relevancy	Open Data Quality Dimensions and Metrics: State of the Art and Applied Use Cases	Englisch	2020
64	Relevancy	Assessing the Importance of Data Factors of Data Quality Model in the Business Intelligence Area	Englisch	2017
236	Relevancy	Review of data quality indicators and metrics, and suggestions for indicators and metrics for structural health monitoring	Englisch	2022
100	Relevancy	Data Quality and Trust: Review of Challenges and Opportunities for Data Sharing in IoT	Englisch	2020
232	Relevancy	Relevance of the two adjusting screws in data analytics: data quality and optimization of algorithms	Englisch	2017
166	Relevancy	Fulmqa: a fuzzy logic-based model for social media data quality assessment	Englisch	2023
166	Relevancy	Fulmqa: a fuzzy logic-based model for social media data quality assessment	Englisch	2023
85	Relevancy	Context-aware Big Data Quality Assessment: A Scoping Review	Englisch	2023
249	Relevancy	Some Thoughts on Testing the Data Quality Metric	Englisch	2022
133	Relevancy	Developing a Data Quality Evaluation Framework for Sewer Inspection Data	Englisch	2020

10	Reliability	A Data Quality Metrics Hierarchy for Reliability Data	Englisch	2016
235	Reliability	Requirements for Data Quality Metrics	Englisch	2018
7	Reliability	A Data Quality Dashboard for Reliability Data	Englisch	2015
111	Reliability	Data measurement in research information systems: metrics for the evaluation of data quality	Englisch	2018
236	Reliability	Review of data quality indicators and metrics, and suggestions for indicators and metrics for structural health monitoring	Englisch	2022
6	Reliability	A Data Quality Dashboard for CMMS Data	Englisch	2018
166	Reliability	Fulmqa: a fuzzy logic-based model for social media data quality assessment	Englisch	2023
133	Reliability	Developing a Data Quality Evaluation Framework for Sewer Inspection Data	Englisch	2023
86	Reliability	Context-aware data quality assessment for big data	Englisch	2018
161	Reliability	Factors affecting data quality of online questionnaires: Issues and metrics for sensory and consumer research	Englisch	2022
234	Reputation	Reputation and Trust Models with Data Quality Metrics for Improving Autonomous Vehicles Traffic Security and Safety	Englisch	2020
64	Reputation	Assessing the Importance of Data Factors of Data Quality Model in the Business Intelligence Area	Englisch	2017

7	Safety	A Data Quality Dashboard for Reliability Data	Englisch	2015
48	Security	An Advanced Big Data Quality Framework Based on Weighted Metrics	Englisch	2022
211	Security	Privacy Preserving Data Quality Assessment for High-Fidelity Data Sharing	Englisch	2014
100	Security	Data Quality and Trust: Review of Challenges and Opportunities for Data Sharing in IoT	Englisch	2020
177	Security	Integrated Framework for Data Quality and Security Evaluation on Mobile Devices	Englisch	2020
38	Security	A security-and quality-aware system architecture for Internet of Things	Englisch	2016
126	Security	Data security and quality evaluation framework: Implementation empirical study on android devices	Englisch	2017
74	Security	Big Data Quality Metrics for Sentiment Analysis Approaches	Englisch	2019
93	Security	Data Quality Indicators Composition and Calculus: Engineering and Information Systems Approaches	Englisch	2015
28	Security	A Systematic Review of Data Quality in CPS and IoT for Industry 4.0	Englisch	2023
111	Timeliness	Data measurement in research information systems: metrics for the evaluation of data quality	Englisch	2018
48	Timeliness	An Advanced Big Data Quality Framework Based on Weighted Metrics	Englisch	2022

16	Timeliness	A Method for Developing Data Quality Measures and Metrics for Primary Health Care	Englisch	2016
230	Timeliness	Reducing defects in the datasets of clinical research studies: conformance with data quality metrics	Englisch	2019
25	Timeliness	A Survey of Data Quality Measurement and Monitoring Tools	Englisch	2022
236	Timeliness	Review of data quality indicators and metrics, and suggestions for indicators and metrics for structural health monitoring	Englisch	2022
127	Timeliness	Daten- und Informationsqualität	Deutsch	2021
146	Timeliness	Entwicklung eines Kennzahlencockpits für Supply Chain Management am Beispiel eines Unternehmens der Elektronikbranche	Deutsch	2020
283	Timeliness	Untersuchung der Datenqualität in FIS	Deutsch	2022
128	Timeliness	Datenqualitätsmetriken zur Unterstützung von Domänenexperten bei interaktiven Analysen	Deutsch	2020
166	Timeliness	Fulmqa: a fuzzy logic-based model for social media data quality assessment	Englisch	2023

Anhang C: Metriktabelle der systematischen Literaturrecherche

Dimension	Metrik
Accessibility	$Accessibility\ (\%)\ = \frac{Anzahl\ der\ zugänglichen\ Werte}{Gesamtanzahl\ der\ Werte}$
Accessibility	$Accessibility\ = 1 - \frac{Anzahl\ der\ Daten, die\ nicht\ verfügbar\ sind}{Gesamtanzahl\ der\ Daten}$
Accuracy	$Accuracy = \frac{Anzahl\ richtiger\ Daten}{Gesamtanzahl\ der\ Daten}$
Accuracy	$Accuracy\ auf\ Feld - Ebene = \frac{Anzahl\ der\ als\ korrekt\ bewerteten\ Felder}{Anzahl\ der\ getesteten\ Felder}$
Accuracy	$Accuracy\ auf\ Datensatz - Ebene = \frac{Anzahl\ der\ als\ vollständig\ korrekt\ bewerteten\ Datensätze}{Anzahl\ der\ getesteten\ Datensätze}$
Accuracy	$Accuracy\ auf\ Objekt - Ebene = \frac{Anzahl\ der\ als\ korrekt\ bewerteten\ Objekte}{Anzahl\ der\ getesteten\ Objekte}$
Accuracy	$Accuracy\ auf\ Objekt - Ebene = 1 - \left(\frac{Anzahl\ der\ Datenobjekte\ mit\ Fehlern}{Gesamtanzahl\ der\ Datenobjekte}\right)$
Accuracy	$Q_{Gen.}\ (\omega, A)\ = min\left(\frac{s(\omega)}{s_{opt.}(A)}, 1\right)$
Accuracy	$Q_{Gen.}\ (t)\ = \frac{\sum_{j=1}^{n} Q_{Gen.}\ (t.A_j, A_j) g_j}{\sum_{j=1}^{n} g_j}$

Appropriate amount of data	$Appropriate\ amount\ of\ data = \min\left(\frac{erforderliche\ Daten}{verfügbare\ Daten}, \frac{verfügbare\ Daten}{erforderliche\ Daten}\right)$
Believability	$Believability = \frac{\sum_{i=1}^{N} rate(i)}{Anzahl\ an\ Daten}$
Completeness	$Anteil\ der\ verfügbaren\ Datensätze = \frac{Anzahl\ der\ verfügbaren\ Datensätze}{Gesamtzahl\ der\ Datensätze}$
Completeness	$Completeness = \frac{Anzahl\ der\ nicht\ leeren\ Werten}{Gesamtzahl\ der\ Werte}$
Completeness	$Completeness = 1 - \frac{T_R}{N_R}$
Conciseness	$Conciseness = \frac{Anzahl\ der\ einzigartigen\ Eigenschaften}{Gesamtanzahl\ der\ Eigenschaften}$
Conciseness	$Conciseness = 1 - \left(\frac{Anzahl\ der\ einzigartigen\ Eigenschaften}{Gesamtanzahl\ der\ Eigenschaften}\right)$
Conciseness	$Conciseness = 1 - \left(\frac{Anzahl\ mehrdeutiger\ Instanzen}{Anzahl\ der\ Instanzen\ im\ semantischen\ Metadaten - Set}\right)$
Consistency	$Anteil\ der\ konsistenten\ Datensätze = \frac{Anzahl\ der\ konsistenten\ Datensätze}{Gesamtanzahl\ der\ Datensätze}$
Consistency	$Anteil\ der\ konsistenten\ Datensätze = 1 - \frac{Anzahl\ Datensätze, die\ gegen\ die\ Konsistenz\ verstoßen}{Gesamtanzahl\ der\ Datensätze}$

Consistency	$$Q_{Kon}(\omega) = \frac{1}{\sum_{j=1}^{n} r_j(\omega) g_j + 1}$$
Consistency	$$Q_{Kons.}(\omega, \mathbb{R}) = \prod_{j=1}^{\mathbb{R}} (1 - r_j(\omega))$$
Consistency	$$Consistency\ (t) = \sum_{r \in \mathbb{R}} \begin{cases} w^+(r), & falls\ t\ r\ erf\ddot{u}llt \\ w^-, & Wenn\ t\ r\ verletzt \\ w^0, & Wenn\ r\ nicht\ zutrifft, \end{cases}$$
Consistent Representation	$$Consistent\ Representation = \frac{\left(\frac{\sum_{j=1}^{N} Datenfeld_j(\sum_{i=1}^{M} brokeRule(i)}{M}\right)}{N}$$
Currency	$$Anteil\ der\ neuen\ Datens\ddot{a}tze = \frac{Anzahl\ der\ neusten\ Datens\ddot{a}tze}{Gesamtanzahl\ der\ Datens\ddot{a}tze}$$
Currency	$$Q_{Akt.}(\omega, A) = \mathrm{e}^{(-Verfall(A)\ x\ Alter\ (\omega, A)}$$
Currency	$$Currency = Zeit, in\ der\ Daten\ im\ System\ gespeichert\ werden - Zeit, in\ der\ Daten\ in\ der\ realen\ Welt\ aktualisiert\ werden$$
Currency	$$Currency = Alter + (Lieferzeit - Eingabezeit)$$
Data Coverage	$$Data\ Coverage = \frac{Anzahl\ der\ g\ddot{u}ltigen\ Messungen\ im\ relevanten\ Bezugszeitraum}{Gesamtanzahl\ der\ potenziellen\ Messungen\ im\ relevanten\ Bezugszeitraum}$$
Data Coverage	$$Data\ coverage = t_v\ x\ d_c$$
Data Decay	$$Data\ Decay = \frac{Anzahl\ der\ Datens\ddot{a}tze\ mit\ negativer\ Ver\ddot{a}nderung}{Gesamtanzahl\ der\ Datens\ddot{a}tze}$$

Data specification	$Anteil\ der\ Datensätze, die\ der\ Spezifikation\ entsprechen = \frac{Anzahl\ der\ Datensätze, die\ der\ Spezifikation\ entsprechen}{Gesamtanzahl\ der\ Datensätze}$
Duplication	$\text{Anteil der duplizierten Werte} = \frac{\text{Anzahl der duplizierten Datensätze}}{Gesamtanzahl\ der\ Datensätze}$
Ease of Manipulation	$Ease\ of\ Manipulation = \frac{Anzahl\ der\ Unterschiede\ zwischen\ der\ Original - und\ der\ bereinigten\ Tab}{Gesamtdaten}$
Free of error	$Free\ of\ error = 1 - \left(\frac{Anzahl\ fehlerhafter\ Daten}{Gesamtanzahl\ an\ Daten}\right)$
Freshness	$Freshness = ts_{cur} - ts_t$
Relevancy	$Relevantestes\ Datenfeld\ f = \frac{Anzahl\ der\ Zugriffe\ auf\ f}{Gesamtzugriffe\ auf\ die\ Tabelle, die\ F\ enthält}$
Relevancy	$Relevancy = \frac{Anzahl\ der\ relevanten\ Daten}{Gesamtanzahl\ der\ Daten}$
Reliability	$Reliability = \sum_{i=1}^{n} Q_i$
Reliability	$Defuzzifzierte\ Reliability = \sum_{i=1}^{n} \frac{a_{1i} + a_{2i} + a_{3i}}{3}$
Security	$Security = \sum_{i=1}^{5} 0{,}2\ x\ Frage\ i$
Timeliness	$\text{Timeliness} = max\left[\left(1 - \frac{\text{Alterswert einer Datenkomponente}}{\text{Haltbarkeit}}\right), 0\right]^s$