Technische Universität Dortmund
Fakultät Maschinenbau
Lehrstuhl IT in Produktion und Logistik

# Masterarbeit

## Erstellung eines Datenqualitätskonzeptes im Kontext der Eigenschaften von Big Data

Vorgelegt von: Bent Mengering
Matrikel-Nr.: 179569
Studiengang: Wirtschaftsingenieurwesen M. Sc.

Prüferin: Dr.-Ing. Dipl.-Inform. Anne Antonia Scheidler
Betreuer: Florian Hochkamp
Abgabedatum: 31.12.2021

# Inhaltsverzeichnis

## Abbildungsverzeichnis

## Tabellenverzeichnis

# 1. Einleitung

Die Analyse elektronisch verfügbarer Daten ist sowohl in wissenschaftlichen Anwendungsfelder als auch in der unternehmerischen Praxis von großer Bedeutung. Durch diese besteht die Möglichkeit, zukünftige Ereignisse zu prognostizieren, Muster, Trends und Zusammenhänge in den Daten zu finden (vgl. Hackett, 2016, S. 5). Die Verwaltung, Auswertung und Aggregation von Daten stellen Unternehmen immer häufiger vor Herausforderungen. Diese sind jedoch unabdingbar für die Führungsprozesse im Unternehmen, da Entscheidungen schnell und souverän auf Basis verlässlicher Daten zu treffen sind (vgl. Hinrichs, 2002, S. 3). Die wirtschaftliche Bedeutung von Daten wird teilweise als so groß angesehen, dass diese neben Arbeitskraft, Ressourcen und Kapital als „vierter Produktionsfaktor" bezeichnet wird (vgl. Horvath, 2013, S. 1). Anlässlich der wachsenden und häufig unterschiedlich strukturierten Datenmengen rückt die Datenqualität immer weiter in den Fokus von Unternehmen, weil nur Daten geeigneter Qualität die gewünschten Informationen erzeugen (vgl. Gräfe & Maaß, 2021, S. 172; vgl. Apel et al., 2015, S. vii). Eine wichtige Voraussetzung für die Nutzung von Daten für eine Anwendung ist somit ein Verständnis über die jeweilige Datenmenge. Die Bewertung der Datenqualität anhand konzipierter Metriken ermöglicht genauere Analysen und zuverlässigere Entscheidungsfindungen (vgl. Jain et al., 2020, S. 3561). In nahezu allen Fachbereichen wird der Datenqualität, in der Literatur auch Informationsqualität genannt, großer Wert beigemessen, wobei die Definitionen und Eigenschaften oft unterschiedlich sind (vgl. Piro & Gebauer, 2021, S. 144). Die Auswirkungen von unzureichender Datenqualität sind vielseitig und weitreichend. Sie kann zum Verlust von Kunden führen, verhindert korrekte Berichterstattungen, erhöht damit das Risiko von Fehlentscheidungen und vermindert die Wirksamkeit von Bereichen wie Vertrieb und Marketing (vgl. Moser, 2021, S. 423). Die Kosten für unzureichende Datenqualität belaufen sich bei vielen Unternehmen auf bemerkenswerte 15 % bis 25 % des Umsatzes (vgl. Redman, 2017).

Aufgrund der hohen Relevanz des Problemkomplexes und der bislang vernachlässigten Berücksichtigung der Probleme der Datenqualität im wissenschaftlichen Diskurs und in der Praxis untersucht diese Arbeit Dimensionen und Metriken in Big

Data. Ziel der Arbeit ist die Abgrenzung der Definitionen zu Daten- sowie Informationsqualität in diesem Kontext. Außerdem zeigt die Literatur eine Vielzahl von verschiedenen charakterisierenden Dimensionen auf. Diese werden mit Bezug auf Big Data analysiert und den Begriffen der Daten- und Informationsqualität zugeordnet. Den wichtigsten Dimensionen werden vorhandene Metriken zugewiesen und in einem praxisnahen Beispiel erläutert. Diese Qualitätsprüfung ermöglicht quantitative Kennzahlen, mit denen über eine vorhandene Datenbasis eine Qualitätsaussage getroffen werden kann. Um diese Metriken in der Praxis umzusetzen, müssen sie einer wissenschaftlichen Begründung unterliegen. Welche Anforderungen das sind und ob die Metriken diese erfüllen, ist ein weiterer Bestandteil der Arbeit. Durch dieses Konzept soll bestimmt werden, ob die Daten für eine weitere Analyse geeignet sind.

Zunächst werden in Kapitel 2 die Grundlagen der Datenqualität erläutert. Mithilfe der Wissenspyramide werden die in Beziehung stehenden Begriffe Zeichen, Daten, Information und Wissen abgegrenzt. Darauf aufbauend wird in Kapitel 2 die Datenqualität behandelt. Da in der Literatur die Begriffe Datenqualität und Informationsqualität weitestgehend synonym verwendet werden, wird dies bis zur Differenzierung im Kapitel 4 auch so erfolgen. Neben der Erörterung der Definitionen zu dem Begriff werden Dimensionen und Metriken der Datenqualität vorgestellt. Nach der Behandlung des Datenqualitätsregelkreises und der Datenqualitätsmängel wird in Kapitel 3 das Augenmerk auf Big Data gerichtet. In Kapitel 4 wird die Literatur zusammengeführt. Die sich aus Big Data ergebenden Anforderungen an die Datenqualität werden herausgearbeitet. Die Unterschiede zu Daten- und Informationsqualität werden aufgezeigt, ebenso erfolgen die Auswahl der Dimensionen sowie die Aufstellung neuer Definitionen. Im nächsten Kapitel werden die Metriken angewendet und deren Eignung in einer Diskussion analysiert. Im Anschluss wird mit dem Fazit eine Gesamtdarstellung der Ergebnisse durchgeführt und die Arbeit kritisch reflektiert. Im letzten Kapitel wird die Arbeit zusammengefasst und ein Ausblick auf zukünftige Forschungsfragen gegeben.

# 2. Grundlagen der Datenqualität

Im folgenden Kapitel wird das theoretische Grundverständnis dieser Arbeit vermittelt. Die in Verbindung stehenden und voneinander abhängigen Begriffe Zeichen, Daten, Informationen und Wissen werden abgegrenzt und definiert. Darüber hinaus wird die Bedeutung von Qualität veranschaulicht und der zusammengesetzte Begriff Datenqualität erläutert sowie Definitionen aus der Literatur gegenübergestellt. Oft in der Literatur genannte charakteristische Dimensionen der Datenqualität werden näher betrachtet und die wichtigsten mit Metriken aus der Wissenschaft versehen.

## 2.1. Aufbau der Wissenspyramide

Die Wissenspyramide ist ein deskriptives Modell zur Darstellung der Entstehung von Wissen. Die im Zusammenhang stehenden Begriffe Zeichen, Daten, Information und Wissen werden pyramidenförmig als vier Ebenen dargestellt. Das Zeichen bildet die Basis der Pyramide und wird gefolgt von Daten und Information; an der Spitze steht das Wissen. In dieser Reihenfolge werden die Begriffe im weiteren Verlauf behandelt.

### 2.1.1. Zeichen

Das Zeichen kann als eine komplexe semiotische Einheit gesehen werden, die sich aus den drei Bestandteilen von Zeichenträger, Bedeutung und Bezeichnung definiert (vgl. Nöth, 2000, S. 131). Die Lehre der Zeichen kann unter bestimmten Umständen divergieren, wenn ein Objekt zwischen Zeichen und Gegenstand variiert. Ein Buchstabe ist grundsätzlich ein Zeichen, jedoch gelingt es, Zeichen in einer anderen Anordnung zu einem sinngemäßen anderen Gegenstand darzustellen.

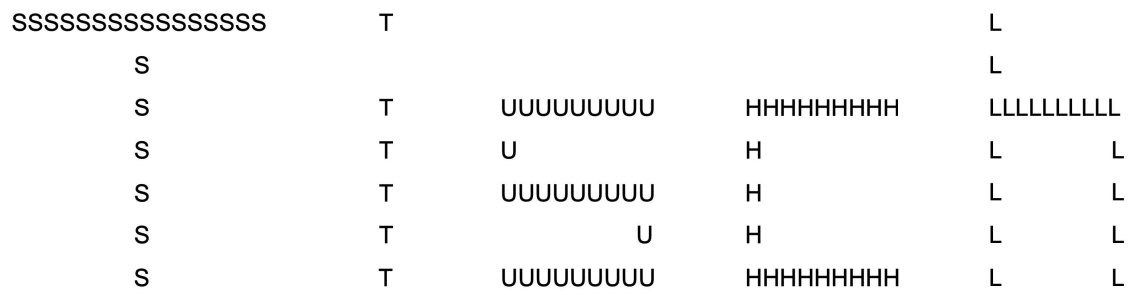


Abbildung 1 Darstellung von Zeichen, in Anlehnung an Borgmann, 1999, S. 18

Die Buchstaben in Abbildung 1, die als kleine Dinge (bloße Tintenflecke) verstanden werden, vermitteln das Wort „Tisch“. Die Zeichen nacheinander gelesen, eines von jeder Gruppe, buchstabieren das Wort „Stuhl“. (vgl. Borgmann, 1999, S. 18,19)

### 2.1.2. Daten

Daten ist der Plural des Wortes Datum, das in der deutschen Sprache auch eine Zeit- bzw. Kalenderangabe ist. Der Duden definiert Daten als „(durch Beobachtungen, Messungen, statistische Erhebungen u. a. gewonnene) [Zahlen]werte, (auf Beobachtungen, Messungen, statistischen Erhebungen u. a. beruhende) Angaben, formulierbare Befunde“ (*Dudenredaktion*, o. J.). Zudem werden Daten im Duden im Kontext der EDV als „elektronisch gespeicherte Zeichen, Angaben, Informationen“ gesehen (*Dudenredaktion*, o. J.). Eine frühe Definition führt Bell 1957 auf; er beschreibt Daten als „jedes numerische oder alphabetische (manchmal auch alphanumerische) Material“ (Bell, 1957, S. 5). Cleve und Lämmel fügen dem die sprachliche Verknüpfung hinzu. Sie sehen Daten als eine Ansammlung von Zeichen mit der dazugehörigen Syntax (vgl. Cleve & Lämmel, 2016, S. 37). Datenarten bilden ein Raster und sind für ein einheitliches Verständnis vonnöten. In diesem Raster werden Informationen niedergelegt, gefunden, erschaffen und gelenkt (vgl. Piro & Gebauer, 2021, S. 144). Im Folgenden werden Beschreibungskriterien erläutert, die als Grundlage für das Verständnis der Datenarten benötigt werden. Die Eigenschaften von Daten werden beschrieben durch Format, Struktur, Stabilität, Inhalt, Businessobjekt und Verarbeitung. Die Kontextinformationen beinhalten die Angaben zu Prozessen und den unterschiedlichen Verwendungszwecken, in denen das Datum genutzt wird. Das *Format* charakterisiert die Daten IT-technisch. In einem Feld kann

das Datum numerisch, integer, alphanumerisch, als Gleitkomma oder ähnlich spezifiziert sein. Auch die Länge ist ein beschreibendes Format (vgl. Piro & Gebauer, 2021, S. 146). Die *Struktur* in den Daten organisiert Informationen so, dass effiziente Algorithmen angewendet werden können. Es gibt drei Kategorien, in denen sich Daten strukturieren; diese sind in Abbildung 2 dargestellt. Bei *strukturierten Daten* sind Informationen zu ihrer Struktur vorhanden, sogenannte Metadaten. Zu den Metadaten können unter anderem Informationen zum Format des Datums, erlaubte Werte für das Datum und die semantische Bedeutung gehören. *Unstrukturierte Daten* sind kurz gefasst Texte, für die ein Textsystem zur Interpretation benötigt wird. Unstrukturierte Daten können jedoch aus strukturierter Information bestehen, wobei diese nicht immer direkt eindeutig ist. Die Informationsgewinnung hängt von der Interpretation ab. *Semistrukturierte Daten* bestehen aus teilweise einzelnen strukturierten Elementen, die im Gesamten jedoch keine eindeutige Struktur aufweisen. Beispiele für semistrukturierte Daten sind Textfelder, in denen nicht angeordnete Daten vorhanden sind (vgl. Piro & Gebauer, 2021, S. 146-147).

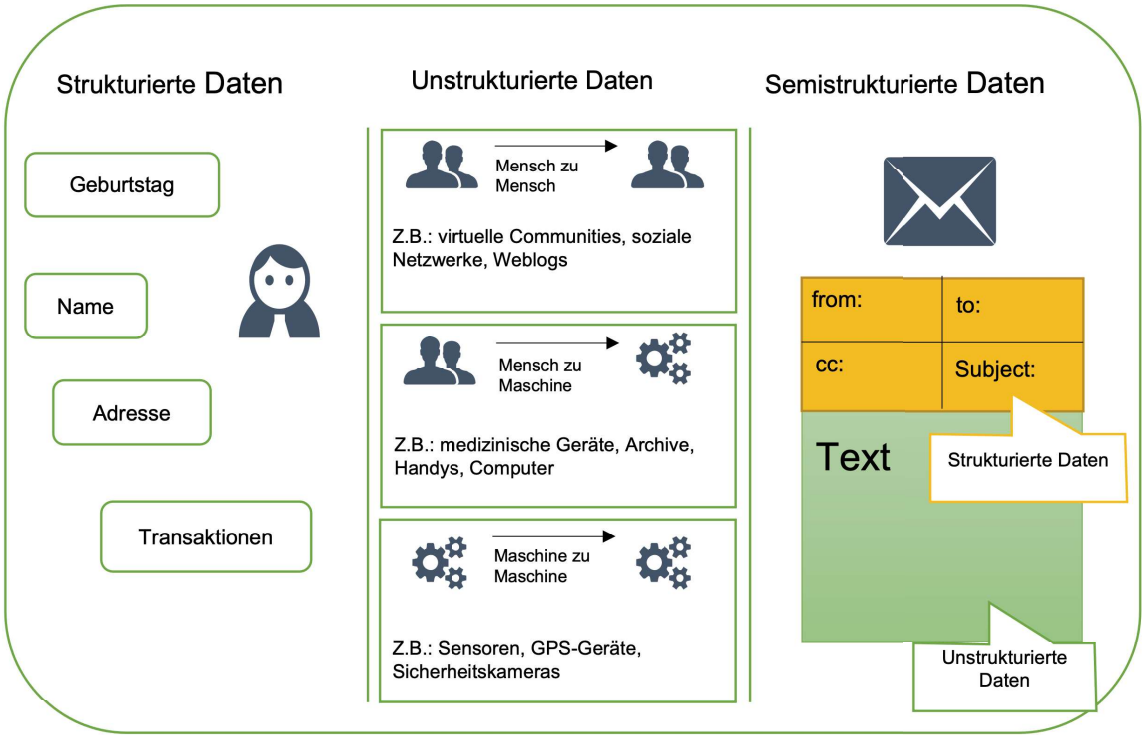


Abbildung 2 Datenvielfalt, in Anlehnung an Klein et al., 2013, S. 321

Der *Inhalt* stellt den Bestandteil der Information in den Daten dar. Er verdeutlicht, ob in den Daten bestimmte Sachverhalte existieren oder ob diese lediglich beschrieben werden. Es gibt eine Abgrenzung zwischen Metadaten und Inhaltsdaten. *Inhaltsdaten* definieren Prio und Gebauer als „Daten, die das Objekt direkt bezeichnen." (Piro & Gebauer, 2021, S. 147). Beispielsweise beschreibt der Name einer Universität inhaltlich direkt die entsprechende Universität. *Metadaten* dokumentieren in strukturierter Form analoge oder digitale Daten. Sie erklären, beschreiben, definieren oder verorten diese (vgl. Jensen et al., 2011, S. 83). Ein weiteres Kriterium, um die Daten zu beschreiben, ist die *Stabilität*. Sie stellt die Zeitdauer, in der die Daten unverändert bleiben, dar. In der Literatur wird zwischen fixen und variablen Daten unterschieden, auch Stamm- und Bewegungsdaten genannt. Stammdaten sind die wesentlichen Grunddaten eines Unternehmens. Sie werden im System selten geändert und normalerweise dauerhaft zentral gespeichert. Dazu gehören zum Beispiel Mitarbeiternamen, Kontonummern oder Artikelbezeichnungen. Bewegungsdaten (variable Daten) ändern sich oftmals in einem Geschäftsprozess. Diese müssen generell eingegeben, aus vorhandenen Daten abgeleitet oder berechnet werden. Dazu gehören zum Beispiel Rechnungsbeträge, Belegdatum oder Bestellmengen (vgl. Piro & Gebauer, 2021, S. 147). Bei der *Verarbeitung* von Daten wird die Position im Datenverarbeitungsprozess auseinandergehalten. Dazu gehören die Eingabedaten, die Speicherdaten und die Ausgabedaten. Die *Eingabedaten* werden in einem System eingepflegt. Beispiele dafür sind Kostenarten, zulässige Belegnummern, Namen oder Bestellmengen. *Speicherdaten* sind bereits im System gespeichert. Es handelt sich zum Beispiel um Daten, die nicht nur für die einmalige Berechnung im System verbleiben, sondern nach der Eingabe im System verbleiben wie Name und Anschrift eines Kunden. Eine weitere Position im Datenverarbeitungsprozess stellen die *Ausgabedaten* dar. Diese sind im System bereits prozessiert, wie zum Beispiel das Ergebnis einer Kostenkalkulation. Im Laufe eines Prozesses kann es zu doppelten Bezeichnungen bei der Verarbeitung der Daten kommen, wenn unterschiedliche Zeitpunkte betrachtet werden (vgl. Piro & Gebauer, 2021, S. 148). Daten, die ein Objekt beschreiben, können dem *Businessobjekt* zugeordnet werden. Das Businessobjekt verbindet Daten mit deren fachlicher Nutzung und ist daher auch für den Inhalt verantwortlich. Ein Geschäftsvorfall benötigt zur Bearbeitung immer Informationen von mehreren Businessobjekten. Datenfelder

sollten immer eindeutig und dauerhaft Businessobjekten zugeordnet sein. Zum Beispiel beschreiben Kundennummer, Kundenname und Adresse das Businessobjekt „Kunde“. Die Zuordnung muss definiert, dokumentiert und in der Regel einmal gültig sein (vgl. Piro & Gebauer, 2021, S. 148).

### 2.1.3. Information

Der Begriff stammt aus dem Lateinischen, wobei dieses vielfach genutzte Wort in der Geschichte kaum verwendet wurde. In einem Häufigkeitswörterbuch aus dem Jahre 1897 wurde das Wort bei 11 Mio. Wörtern lediglich 55-mal verwendet (vgl. Zemanek, 1986, S. 19). Das Wort Information ist jedoch in der heutigen Zeit allgegenwärtig. Es existieren zahlreiche Komposita dieses Begriffes in verschiedenen Kontexten (vgl. Bawden, 2001, S. 96–97). In allen wissenschaftlichen Disziplinen ist es ein fundamentaler Bestandteil und wird unterschiedlich definiert (vgl. Piro & Gebauer, 2021, S. 144). Schon Capurro kam 1978 zu dem Schluss, dass es eine „unüberschaubare Anzahl von einzelwissenschaftlichen Definitionen“ (Capurro, 1978, S. 290) zu dem Begriff Information gebe. Dieses führt unter anderem zu Streitigkeiten, sodass Experten gegenseitig ihre Definitionen als falsch oder ungenügend bezeichnen (vgl. Ingold, 2011, S. 8). Die schiere Anzahl an Definitionen ergibt sich nach Caturro unter anderem aus ständigen Missverständnissen. Dies führt unter anderem zu der Äußerung „Information ist nicht gleich Information und vor allem nicht zu verwechseln mit Information“ (Duxa et al., 2005, S. 119).
Das Begriffsverständnis von Information hat seit der zweiten Hälfte des 20. Jahrhunderts eine große semantische Verschiebung vollzogen. So wird neben der traditionellen Bedeutung Information vielmals als etwas „Quantifizierbares und in Dokumenten, Computern oder auf anderen elektronischen Trägern Gespeichertes“ verstanden (Ingold, 2011, S. 26). Allzu oft werden die nahestehenden Begriffe Daten und Informationen bedeutungsgleich verwendet, sodass es einer Unterscheidung bedarf. Nach Gray hat es in der Vergangenheit nie eine genaue Unterscheidung der Begrifflichkeiten „Daten“ und „Information“ gegeben. In dem im Jahr 1957 erschienenen Buch „An Introduction to Automatic Computers“ von Ned Chapin, das Gray als „earliest textbook on business computing“ betitelt, werden die Begriffe „Daten“ und „Information“ auch augenscheinlich synonym verwendet (vgl. Gray, 2003,

S. 2844–2845). Chapin beschreibt die Information jedoch als Untermenge von Daten. Er sieht alle Informationen als Daten, jedoch nicht umgekehrt. Wenn der Nachrichtengehalt der Daten gering ist, ist die Verwendung von Informationen falsch, wobei dies kontextabhängig ist. In seinem Beispiel erklärt er, dass Daten über Aktienkurse für einen Hausmeister nichts weiter als Daten seien, für einen Börsenmakler jedoch diese Daten als Information zu betrachten seien (vgl. Chapin, 1957, S. 219). Floridi beschreibt Information schlicht als „Daten mit Bedeutung“ (Floridi, 2005, S. 351). In einer späteren Veröffentlichung bekräftigt er, dass Menschen Daten immer mit irgendeiner Form eines semantischen Kontextes aufnehmen und die Daten interpretieren (vgl. Floridi, 2014, S. 161). Dem gegenüber steht eine Behauptung von Strong. Er bekräftigt, dass ohne eine Kenntnis des Kontexts keine Information vorhanden sei und Daten somit nicht bewertet werden könne. Daten werden erst zu Information, wenn sie in einem bestimmten Kontext benutzt werden (vgl. Strong et al., 1997, S. 103). Aufgrund der zahlreichen verschiedenen Ansätze aus unterschiedlichen Bereichen existiert eine Vielzahl von Definitionen. Dies veranlasste Bode dazu, eine Typologie zu entwickeln, die vor allem Ansätze aus der Betriebswirtschaftslehre strukturiert (Tabelle 1).

Tabelle 1 Typologie der Informationsbegriffe, in Anlehnung an Bode, 1997, S. 452

| **Dimensionen** | **Ausprägungen** | | |
|---|---|---|---|
| Semiotik | Syntaktisch | Semantisch | Pragmatisch |
| Träger | Ungebunden | | Menschengebunden |
| Neuheitsgrad | Subjektiv | | Objektiv |
| Wahrheitsgehalt | Wahrheitsabhängig | | Wahrheitsunabhängig |
| Zeitbezogenheit | Statisch | | Prozessual |

Die zentrale Struktur von Informationen ist die Beziehung zwischen einem Zeichen, einem Gegenstand und einer Person. Eine Person wird von einem Zeichen über einen Gegenstand informiert. Es gibt verschiedene Bezeichnungen für die drei Kategorien in dieser Beziehung. Die Person kann auch als Empfänger der Information, als Zuhörer, als Beobachter oder Ähnliches bezeichnet werden. Das Zeichen kann als Signal, Symbol oder Sender verstanden werden. Der Gegenstand ist die Botschaft, die Bedeutung, der Zusammenhang, die Nachricht oder die Information

(vgl. Borgmann, 1999, S. 18). Prinzipiell lassen sich die Informationen in einem Unternehmen in formell strukturierte und informelle Informationen unterteilen. Wie im vorherigen Kapitel ausführlich erläutert sind Daten ohne Kontextbezug nahezu unbrauchbar, sie folgen immer dem Zweck eines Prozesses. Dieser Zusammenhang ist in **Fehler! Verweisquelle konnte nicht gefunden werden.** zur Konkretisierung dargestellt. Die Informationen stammen aus Daten im Hinblick auf eine zielgerichtete Verwendung (vgl. Gebauer & Windheuser, 2021, S. 88).

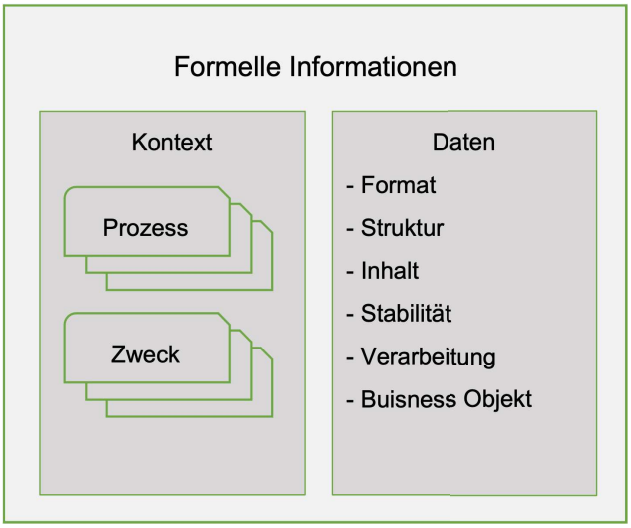


Abbildung 3 Zusammenhang Daten und Information, in Anlehnung an Gebauer & Windheuser, 2021, S. 88

### 2.1.4. Wissen

Nachfolgend wird die Wissenspyramide dargestellt und beispielhaft erklärt; ebenso wird der Begriff Wissen, der die Spitze der Pyramide bildet, erläutert. Dieses Modell besteht aus vier Ebenen und zeigt die Differenzierungen der Begriffe (siehe Abbildung 4).

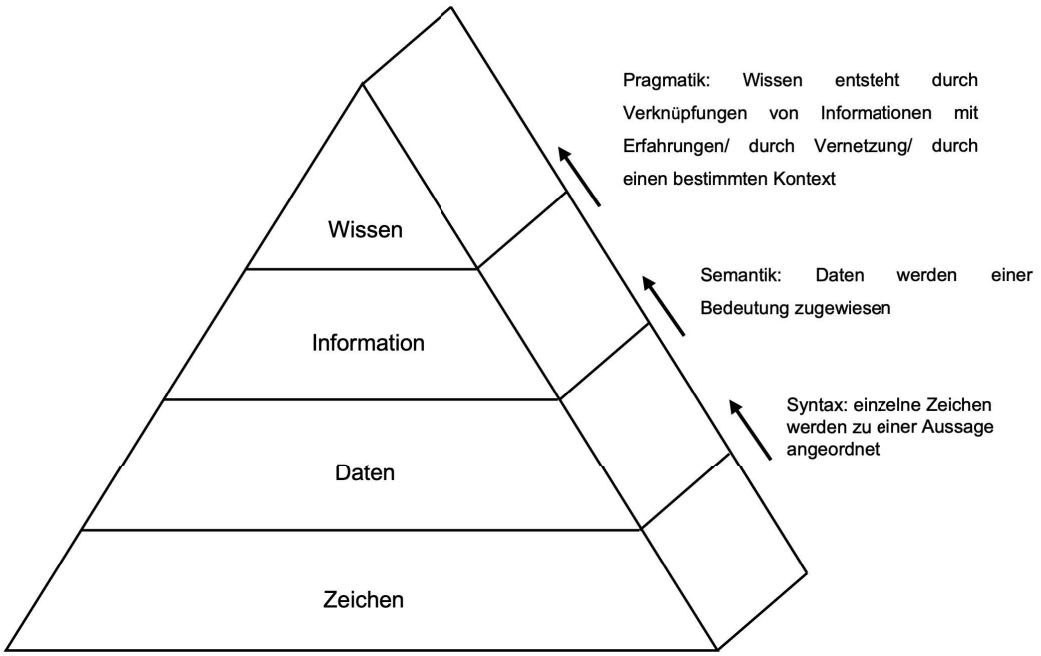


Abbildung 4 Wissenspyramide, in Anlehnung an Bodendorf, 2006b, S. 1

Um diesen Sachverhalt zu verdeutlichen, werden zuerst Beispiele für die Begriffe genannt, deren Bedeutung fortlaufend beschrieben werden.

- Zeichen
    - E, H, A, !, E, R, W, L, H, F, T, E, R, A, F, E
- Daten
    - FEHLERHAFTE WARE!
- Information
    - Die bestellte Ware entspricht nicht den Anforderungen.
- Wissen
    - Es muss eine Retoure erstellt, die Ware neu bestellt werden und nachgelagerte Prozesse verschieben sich.

Eine willkürliche Anordnung von Zeichen aus einem Zeichenvorrat wird im ersten Schritt nach definierten Syntaxregeln zu Daten geordnet. Die geordnete Aneinanderreihung von Zeichen wird zu einer eindeutigen Information. Dies erfolgt mithilfe des Verbindens eines Begriffes mit einem Verständnis aus der realen Welt. Ein Mensch, der die deutsche Sprache spricht, versteht den Kontext und wird die Daten „FEHLERHAFTE WARE!“ interpretieren können. Im Anschluss entsteht durch Verknüpfung von Information das Wissen. Dazu muss ein Zusammenhang zwischen

Informationen hergestellt werden, was mit Ursache-Wirkungs-Beziehungen einhergeht, und die Information muss zweckgerichtet sein, was auch als Pragmatik bezeichnet wird. Nun besteht eine Verbindung zwischen der Information, dass die Ware fehlerhaft ist, und einer weiteren Information, zum Beispiel, dass der Kunde die Ware schnellstmöglich benötigt. Dies führt zu dem Wissen, dass die Ware neu geliefert werden muss (vgl. Bodendorf, 2006, S. 1–2).

$$F_C = \frac{q_1 * q_2}{4 * \pi * \varepsilon_0 * r^2} \tag{1}$$

Ein anderes anschauliches Beispiel, um eine Differenzierung zwischen Information und Daten herzustellen, ist die **Fehler! Verweisquelle konnte nicht gefunden werden.**. Die Zeichen wurden mithilfe der Syntax zu Daten angeordnet. Der Text und die Symbole werden erst zu Information, wenn der Lesende einen Kontext ausmacht. Für manche Menschen hat das Geschriebene keine Bedeutung. Die meisten werden erkennen, dass es eine Gleichung für eine Rechnung ist, zudem anhand von $\pi$ und $r^2$, dass ein Kreis eine Rolle spielt. Andere wissen, dass F für Kraft steht und es sich somit um eine Formel handelt, um eine Kraft zu berechnen. Elektroingenieure erkennen das Coulomb'sche Gesetz und wissen, dass $F_C$ die Anziehungs- bzw. Abstoßkraft von zwei Punktladungen beschreibt. Das $\varepsilon_0$ steht für die elektrische Feldkonstante, $q_1$ und $q_2$ für die elektrische Ladungsmenge der jeweiligen Punktladungen. $r$ ist der Abstand zwischen den Mittelpunkten und $\pi$ ist die Kreiszahl, die das Verhältnis des Umfanges eines Kreises zu seinem Durchmesser angibt.

## 2.2. Zusammensetzung der Datenqualität

Datenqualität ist eine Zusammensetzung der Begriffe Daten und Qualität. Daten wurden bereits im Kapitel 2.1.2. beschrieben, sodass nun der Qualitätsbegriff einer näheren Betrachtung bedarf. Das Wort Qualität leitet sich im Wesentlichen vom lateinischen „qualitas" ab, was sich mit dem Wort Beschaffenheit erklären lässt (vgl. Zech, 2015, S. 23). Die Qualitätsaussage ist generell ein wertloser Hinweis auf das

Wesen der Dinge und schließt positive oder negative Schwankungen nicht ein. Darüber hinaus stellt sich in der erkenntnistheoretischen Forschung zunehmend die Frage, ob die Qualitätsaussage durch den Aussagegegenstand bestimmt wird oder lediglich durch die Person, die eine Qualitätsaussage trifft. Einerseits führt Aussage die subjektive Wahrnehmung ein, andererseits führt es den Unterschied zwischen Erfahrungsobjekten ein und analysiert die Relativität von Begriffen (vgl. Kauffmann, 1996, S. 429-430). Qualität wird Rahmen des Qualitätsmanagements in der Norm DIN EN ISO 9000:2015-11 als „Grad, in dem ein Satz inhärenter Merkmale eines Objekts Anforderungen erfüllt", definiert (DIN EN ISO 9000, 2015). Die Qualität gibt somit an, in welchem Maße ein Produkt (Ware oder Dienstleistung) den bestehenden Anforderungen entspricht. Demnach ist Qualität immer etwas Erreichtes in Relation zu den Qualitätsanforderungen der betrachteten Einheit. Eine kurze und zugleich klare Definition stammt von Geiger und Kotte: „Qualität = Realisierte Beschaffenheit bezüglich geforderter Beschaffenheit" (Geiger & Kotte, 2008, S. 68). Eine in der Literatur immer wieder verwendete Systematisierung der Qualität stammt von Garvin und differenziert fünf verschiedene Vorstellungen des Begriffes im Kontext der Fertigungsindustrie. Der *produktionsorientierte Ansatz* kommt einem objektiven Qualitätsbegriff nahe, denn Qualität wird als messbarer und genau zuordenbarer Parameter verstanden, der ein Produkt beschreibt. Diese Methode bezieht sich nur auf das Endprodukt und hat nichts mit dem Kunden (Benutzer) zu tun. Qualität stellt somit eine objektive Größe dar, die nicht von der subjektiven Wahrnehmung bestimmt wird. Daher können Qualitätsunterschiede auf Unterschiede in den Produkteigenschaften zurückgeführt werden. Des Weiteren gibt es den *anwenderorientierten Ansatz*, der die Qualität des Produkts durch den Produktnutzer definiert. In diesem Fall entscheidet der Kunde (subjektiv), inwieweit das Produkt der geforderten Qualität entspricht (auch „fit for purpose" oder „fit for use" genannt). Endverbraucher können unterschiedliche Bedürfnisse haben, sodass die Qualität desselben Produkts unterschiedlich bewertet werden kann. Der *prozessorientierte Ansatz* folgt der Prämisse, dass ein optimaler und kontrollierter Fertigungsprozess, der alle Anforderungen erfüllt, Qualität hervorbringt. Jegliche Abweichung von dem im Vorhinein definierten Prozess gilt als Qualitätseinbuße. Der vierte Ausgangspunkt ist der *wertorientierte Ansatz.* Dieser betrachtet die Qualität unter Kostengesichtspunkten. Besteht ein akzeptables Verhältnis zwischen Kosten und erhaltener Leistung,

ist das Produkt von hoher Qualität. Der letzte Ansatz ist der *transzendente* und charakterisiert Qualität als festgelegte Einzigartigkeit oder Superlative. Qualität gilt als Synonym für hohe Ansprüche. Dieser Grundgedanke geht von einem philosophischen Verständnis aus, dass die Qualität nicht messbar, sondern nur erfahrbar sei (vgl. Garvin, 1984, S.25-28).

In der Literatur werden die Begriffe Datenqualität und Informationsqualität häufig synonym verwendet (vgl. Eppler, 2006, S. 349; vgl. Gebauer & Windheuser, 2021, S. 87; vgl. Lee et al., 2006, S. 9). Laut Müller lässt sich durch die Semiotik als Strukturhilfe die Datenqualität in eine syntaktische, semantische und pragmatische differenzieren. Die technische Verfügbarkeit und Nutzbarkeit der Daten sind auf *syntaktischer Ebene* einsehbar. Dies gilt neben datenschutzrechtlichen Aspekten auch für Fragen der sachgerechten und einheitlichen Repräsentation des dargestellten Sachverhalts, wie etwa der Einheitlichkeit von Format und Darstellung. Die *semantische Ebene* berücksichtigt die Merkmale, die sich auf den Informationsgehalt der Daten beziehen, wie Genauigkeit, Detaillierung, Validität und Quantifizierbarkeit. Auch verschiedene Aspekte empirischer und logischer Wahrheitsgehalte wie Glaubwürdigkeit, Fehlerfreiheit, Konsistenz und Überprüfbarkeit gehören zu dieser Ebene. Zeitliche und sachliche Eignung zählen zu der *pragmatischen Ebene,* ebenso die Vollständigkeit der Daten für die jeweilige Absicht (vgl. Müller, 2000, S. 15; vgl. Wang et al., 1995, S. 629-632).

Miller stellt bei seiner Erläuterung den Anwender in den Vordergrund. Für ihn liegt die Bedeutung der Informationsqualität darin, wie die Information vom Konsumenten wahrgenommen und genutzt wird. Die Wahrnehmung der Merkmale definiert die Informationsqualität. Die Ermittlung der Informationsqualität erfolgt in zwei Stufen. Zuerst müssen Merkmale für den Konsumenten benannt werden. Daraufhin stellt sich die Frage, wie sich die Merkmale auf den betreffenden Konsumenten auswirken. Folgende spielen für ihn eine Rolle: Korrektheit, Aktualität, Vollständigkeit, Widerspruchsfreiheit, Format, Zugriffsfähigkeit, Kompatibilität, Sicherheit und Validität (vgl. Miller, 1996, S. 79). Auch Holthuis macht den Nutzen von Daten anhand einer Reihe von Merkmalen fest. Für ihn gehören zu den bedeutsamsten Relevanz, Genauigkeit, Vollständigkeit, Zusammenhang, Zugriffsmöglichkeit, Flexibilität, Zeit- und Zeitraumbezug, Transportierbarkeit und Sicherheit (vgl. Holthuis, 1999, S. 33–35). Für Olson hängt die Qualität der Daten sowohl von dem angeforderten Verwendungszweck als auch von den Daten selbst ab. Um den Verwendungszweck zu

erfüllen, müssen die Daten die Attribute Korrektheit, Aktualität, Relevanz, Vollständigkeit, Verständlichkeit und Vertrauenswürdigkeit besitzen (vgl. Olson, 2003, S. 24). Als praktische Definition von Datenqualität kann folgende Ausführung von Gebauer und Windheuser verwendet werden: „Datenqualität ist die Gesamtheit der Ausprägungen von Qualitätsmerkmalen eines Datenbestandes bezüglich dessen Eignung, festgelegte und vorausgesetzte Erfordernisse zu erfüllen“ (Gebauer & Windheuser, 2021, S. 88). Eine weitere Auffassung von Datenqualität im Kontext der Informationssysteme stellt eine Abbildungsgüte zwischen realer Welt und der Repräsentation im Anwendungssystem dar. Eine qualitativ hochwertige Systemabbildung ist somit vollständig, eindeutig, bedeutungsvoll und korrekt (vgl. Wand & Wang, 1996, S. 92–93). Wangs und Strongs Definitionsansatz folgt der „fitness for use by information consumers“ und stellt somit auch den Konsumenten der Information in den Vordergrund. Ihre empirische Erfassung genereller Merkmale strukturiert sich in vier Kategorien: innere Datenqualität, kontextabhängige Datenqualität, Darstellung und Zugriff (vgl. Wang & Strong, 1996, S. 18–21). Eine Definition, die verschiedene Aspekte miteinbezieht, ist die von Würthele. Er bezeichnet die Datenqualität als „mehrdimensionales Maß für die Eignung von Daten, den an ihre Erfassung/Generierung gebundenen Zweck zu erfüllen. Diese Eignung kann sich über die Zeit ändern, wenn sich die Bedürfnisse ändern“ (Würthele, 2003, S. 21). Viele Gemeinsamkeiten bei den Ausführungen zur Daten- bzw. Informationsqualität lassen sich im Hinblick auf den Datenempfänger finden. Diese anwenderbezogene Sicht misst den Wert, den die Daten für den Konsumenten haben. In der Literatur werden zahlreiche Qualitätsmerkmale genannt und an die Definitionen gebunden. Mit Blick auf die im Vorfeld konkretisierten Definitionen lässt sich die undurchsichtige Anzahl an verschiedenen Merkmalen erahnen. Diese Merkmale wurden teils aus Erfahrungen der Autoren, teils aus Expertenwissen oder empirischen Studien zusammengetragen. Die Beschreibungen dieser Merkmale werden im folgenden Kapitel 2.2.1 ausführlicher betrachtet. Die in **Fehler! Verweisquelle konnte nicht gefunden werden.** dargestellte Datenqualitätspyramide stellt die drei Ebenen erfolgreicher Operationen dar. Datenqualität kann als Obermenge aller Datenqualitätsmerkmale verstanden werden, die die zweite Ebene der Pyramide darstellt. Um Datenqualitätsmerkmale bewerten zu können, bedarf es Datenqualitätsmetriken (vgl. Gebauer & Windheuser, 2021, S. 88-89). Die Elemente der Datenqualitätspyramide werden in den folgenden zwei Kapiteln umfassender erläutert.

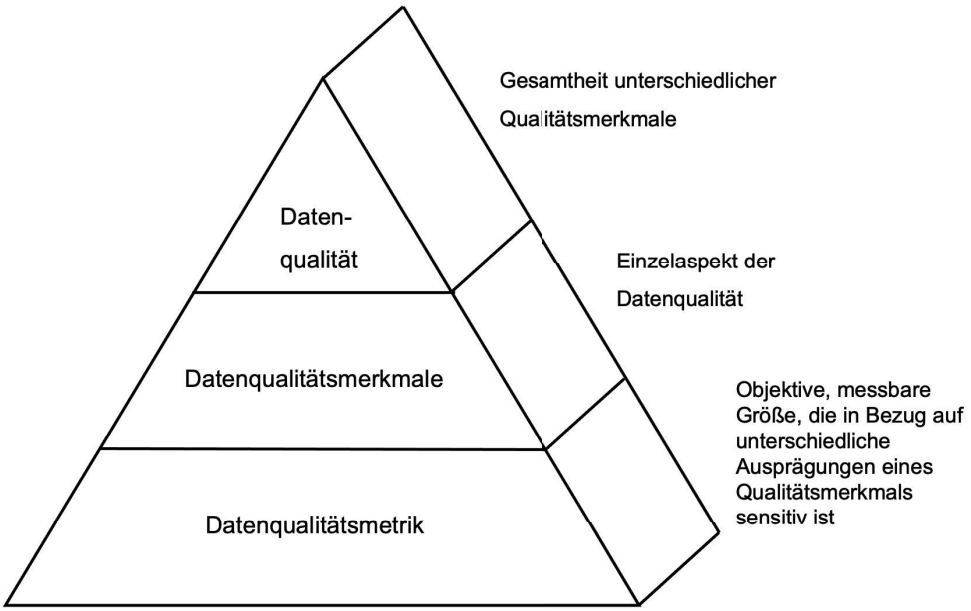


Abbildung 5 Datenqualität, -qualitätsmerkmale und -qualitätsmetrik, in Anlehnung an Gebauer & Windheuser, 2021, S. 88

### 2.2.1. Datenqualitätsdimensionen

Um Datenqualität besser beschreiben und bewerten zu können, werden Dimensionen herangezogen. Dimensionen, auch Merkmale, Attribute oder Kriterien genannt, werden schon lange zur Charakterisierung genutzt. Einige Dimensionen wurden bereits im vorherigen Kapitel genannt, da diese oftmals ein Teil der Definition der Daten- oder Informationsqualität sind. Die Anfänge der Publikationen in diesem Zusammenhang zielten jedoch nicht auf den Begriff der Daten, sondern auf den Informationsbegriff ab. Schon 1983 schlugen King und Epstein mehrere Informationsattribute vor, um ein zusammengesetztes Maß für den Informationswert zu erhalten. Zu den vorgeschlagenen Informationsattributen gehörten Hinlänglichkeit, Verständlichkeit, Unvoreingenommenheit, Zuverlässigkeit, Entscheidungsrelevanz, Vergleichbarkeit und Quantität (vgl. W. R. King & Epstein, 1983, S. 34–35). Auch 1987 wurden zahlreiche Kriterien für die Informationsqualität in das weite Feld der „User Information Satisfaction“ aufgenommen. Das Zufriedenheitsmaß von Iivari und Koskela umfasste drei Konstrukte der Informationsqualität: „Informativität“, bestehend

aus Relevanz, Ausführlichkeit, Aktualität, Genauigkeit und Glaubwürdigkeit, „Zugänglichkeit“, bestehend aus Bequemlichkeit, Aktualität und Interpretierbarkeit, sowie „Anpassungsfähigkeit“ (vgl. Iivari & Koskela, 1987, S. 415).

Einer der bedeutendsten Autoren ist Richard Y. Wang. Viele Werke stützen sich auf eine Studie von ihm, sodass auch in dieser Arbeit seiner Veröffentlichung große Aufmerksamkeit gewidmet wird und diese als Grundlage für Weiterführendes dient (vgl. Müller, 2000, S. 14–16; vgl. Rohweder et al., 2021, S. 24; vgl. Gebauer & Windheuser, 2021, S. 90; vgl. Heinrich & Klier, 2015, S. 52; vgl. Apel et al., 2015, S. 10).

Wang und Strong haben 1996 eine zweistufige Erhebung vorgenommen, um wichtige Datenqualitätsattribute herauszuarbeiten. In der ersten Stufe wurde eine Umfrage bei Datenkonsumenten durchgeführt, um potenzielle Attribute zu identifizieren. Sowohl Personen aus der Industrie, die in verschiedenen Kontexten mithilfe von Daten Entscheidungen treffen müssen, als auch Studenten, die berufliche Erfahrungen mit Daten haben, wurden befragt. In dieser Umfrage wurden 179 Attribute ermittelt (vgl. Wang & Strong, 1996, S. 10).

In der zweiten Stufe wurde eine Umfrage durchgeführt, um Ergebnisse über die Bedeutung jedes dieser Attribute für die Datenkonsumenten zu sammeln. Die Bewertungsskala reichte von 1 für sehr wichtig bis 9 für unwichtig. Bei der ersten Befragung wurden teils ähnliche Attribute angegeben, sodass bei der Bewertung ein Mittelwert gebildet wurde, um entsprechende Ergebnisse zu erhalten. Zum Beispiel bestand die Dimension Leichtigkeit des Verständnisses aus den drei Kategorien: leicht verständlich, lesbar und klar. Der Mittelwert für die Wichtigkeit der Dimension „Verständlichkeit“ war der Durchschnitt der Wichtigkeitsbewertungen für „leicht verständlich“, „lesbar“ und „klar“ (vgl. Wang & Strong, 1996, S. 12).

Im Anschluss wurde zusätzlich eine Zwei-Phasen-Sortierstudie durchgeführt. Wang und Strong empfanden 20 Dimensionen für eine praktische Evaluation als zu viel. Außerdem wurden diese Dimensionen zwar nach ihrer Wichtigkeit geordnet, aber die Dimensionen mit dem höchsten Rang erfassen möglicherweise nicht die wesentlichen Aspekte der Datenqualität. Die erste Phase der Studie bestand darin, diese Zwischendimensionen in eine kleine Gruppe von Kategorien zu sortieren. In der zweiten Phase sollte bestätigt werden, dass diese Dimensionen tatsächlich zu den Kategorien des vorläufigen konzeptionellen Rahmens gehörten. Es wurden zunächst vier Kategorien gebildet, in die die 20 Dimensionen eingeteilt wurden. Im

Anschluss gab es verschiedene Vorgehen von Neusortierungen und Umbenennungen (vgl. Wang & Strong, 1996, S. 16–17).

Die daraus resultierenden Kategorien sind intrinsische, kontextabhängige, begriffliche und Zugangsdatenqualität. Werden Kriterien in Kategorien zusammengefasst, wird dies als Qualitätsmodell bezeichnet. Die Zerlegungssystematik ist ein wesentliches Merkmal dieses Modells (vgl. Apel et al., 2015, S. 10). Intrinsische DQ bedeutet, dass die Daten eine eigenständige Qualität haben. Die Genauigkeit ist nur eine der vier Dimensionen, die dieser Kategorie zugrunde liegen. Die kontextabhängige DQ unterstreicht die Anforderung, dass die Datenqualität im Kontext der jeweiligen Aufgabe betrachtet werden muss, d. h., die Daten müssen relevant, zeitnah, vollständig und von der Menge her angemessen sein, um einen Mehrwert zu schaffen. Begriffliche DQ und Zugangs-DQ betonen die Bedeutung der Rolle von Systemen, d. h., das System muss zugänglich, aber sicher sein, und es muss Daten so darstellen, dass sie interpretierbar, leicht verständlich, prägnant und konsistent dargestellt werden (vgl. Wang & Strong, 1996, S. 20–21). Die Kategorien mit dazugehörigen Dimensionen sind mit den Definitionen sind in Tabelle 2 dargestellt.

Tabelle 2 Kategorien der Datenqualität, in Anlehnung an Treiblmaier, 2011, S. 6; Wang & Strong, 1996, S. 20

| **Kategorie** | **Dimension** | **Definition** |
|---|---|---|
| Intrinsische Datenqualität (Intrinsic Data Quality) | Glaubwürdigkeit (Believability) | Das Ausmaß, in dem Daten als wahr, real und glaubhaft angenommen werden |
| | Genauigkeit (Accuracy) | Das Ausmaß, in dem Daten als korrekt und zuverlässig für konkrete Aufgaben angesehen werden |
| | Objektivität (Objectivity) | Das Ausmaß, in dem Daten als unvoreingenommen und unparteiisch angesehen werden |
| | Reputation (Reputation) | Das Ausmaß, in dem Daten vertraut wird bzw. deren Quelle oder Inhalt als vertrauenswürdig gelten |
| Kontextabhängige Datenqualität (Contextual Data Quality) | Mehrwert (Value-added) | Das Ausmaß, in dem Daten Nutzen stiften und durch ihren Gebrauch Vorteile schaffen |
| | Relevanz (Relevancy) | Das Ausmaß, in dem Daten für konkrete Aufgaben anwendbar und hilfreich sind |

| | | |
|---|---|---|
| | Aktualität (Timeliness) | Das Ausmaß, in dem das Alter der Daten für bestimmte Aufgaben angemessen ist |
| | Vollständigkeit (Completeness) | Das Ausmaß, in dem Daten eine entsprechende Breite, Tiefe und einen bestimmten Umfang für spezielle Aufgaben aufweisen |
| | Angemessene Menge (Appropriate amount of data) | Das Ausmaß, in dem die Quantität der Daten den Aufgaben entspricht |
| Begriffliche Datenqualität (Representational Data Quality) | Interpretierbarkeit (Interpretability) | Das Ausmaß, in dem Daten in einer verständlichen Sprache vorliegen und verwendete Maßeinheiten bzw. Definitionen verständlich sind |
| | Verständlichkeit (Ease of understanding) | Das Ausmaß, in dem Daten ohne Doppeldeutigkeiten vorliegen und einfach begriffen werden können |
| | Konstanz in der Repräsentation (Representational consistency) | Das Ausmaß, in dem Daten ständig in demselben Format repräsentiert werden und mit früheren Daten kompatibel sind |
| | Übersichtlichkeit (Concise representation) | Das Ausmaß, in dem Daten kompakt repräsentiert werden (z. B. kurze und prägnante, jedoch komplette Darstellungsform) |
| Zugangs Datenqualität (Accessibility Data Quality) | Erreichbarkeit (Accessibility) | Das Ausmaß, in dem Daten zur Verfügung stehen bzw. schnell und einfach erhalten werden können |
| | Zugangssicherheit (Access security) | Das Ausmaß, in dem der Zugang zu Daten eingeschränkt und kontrolliert werden kann |

Die von Wang und Strong erarbeiteten Merkmale wurden um zahlreiche Dimensionen ergänzt. Dies resultiert aus der steigenden Relevanz der Thematik mit wachsenden Datenmengen und dem Bedürfnis nach Kontrolle bzw. Messbarkeit dieser vorliegenden Daten. Zum anderen hat sich das Bewusstsein über Wettbewerbsvorteile bei konkurrierenden Unternehmen durch die Nutzung von Daten verändert. Im Anhang befinden sich weitere Dimensionen aus der Literatur, die grundsätzlich dazu beitragen, die Qualität von Daten oder Informationen zu charakterisieren.

### 2.2.2. Datenqualitätsmetriken

Um nachzuvollziehen, ob gewünschte Zustände erreicht werden, bedarf es Kennzahlen. Diese Zahlen fassen messbare relevante Daten zusammen und stellen diese in einen größeren Zusammenhang. Sie bündeln schwer überschaubare Daten zu einer aussagekräftigen Größe und stellen damit komplexe Sachverhalte kurz und prägnant dar (vgl. Vollmuth & Zwettler, 2016, S. 8). „Measurements are key. If you cannot measure it, you cannot control it. If you cannot control it, you cannot manage it. If you cannot manage it, you cannot improve it. It is as simple as that“ (Harrington, 1991, S. 82). Eine Metrik bezeichnet grundsätzlich eine Vorgehensweise zur Messung einer quantifizierbaren Größe und ermöglicht Objektivität. Während für unterschiedliche Qualitätsmerkmale (Länge, Gewicht, Leistung etc.) von (materiellen) Produkten und betriebliche Stärken und Schwächen in der Regel etablierte Kennzahlen, Metriken und Messverfahren existieren, ist dies in der Regel bei (nicht materiellen) Datenprodukten nicht der Fall (vgl. Witte, 2018, S. 1–2). Wie in Kapitel 2.2.1 erwähnt, besteht noch nicht einmal Einigkeit über die zu berücksichtigenden Qualitätsmerkmale. In der Literatur finden sich wenige Empfehlungen zu Datenqualitätsmetriken (vgl. Hinrichs, 2002, S. 44). Das Fehlen geeigneter Metriken führt dazu, dass die in der Literatur beschriebenen Methoden zur Messung der Datenqualität unumgänglich eher informelle Merkmale aufweisen (vgl. Hinrichs, 2002, S. 44). Bevor nachfolgend das Augenmerk auf die Ausarbeitung von Datenqualitätsmetriken gerichtet wird, sollten zuerst die Anforderungen an die Metriken definiert werden, die einer wissenschaftlichen Begründung folgen und eine praktische Anwendung garantieren sollen. Zum einen ist die *Normierung* der Metrikergebnisse vonnöten, um eine Interpretierbarkeit und Vergleichbarkeit der Ergebnisse sicherzustellen. Die *Kardinalität* von Metriken hilft bei der ökonomischen Bewertung von Maßnahmen und bei der Kontrolle der zeitlichen Entwicklung der Ergebnisse. Somit können Ausprägungen in eine Rangordnung gebracht werden und es kann bestimmt werden, inwieweit sich zwei unterschiedliche Merkmalsausprägungen unterscheiden. Die *Sensibilisierbarkeit* gewährleistet, dass das Ergebnis zielgerichtet gemessen werden kann. Metriken müssen dafür für eine bestimmte Anwendung sensibilisiert und für das vorhandene Ziel konfiguriert werden. Verschiedene miteinan-

der in Beziehung stehende Objekttypen werden zu einem höheren Objekttyp zusammengefasst. Dieses nennt sich *Aggregierbarkeit*. Sind die Ergebnisse in Attributwert-, Tupel, Relationen- sowie Datenbankebene gegliedert, wird ein flexibler Einsatz realisiert. Die Metrikergebnisse müssen *fachlich interpretierbar* sein, da für die praktische Anwendung eine Normierung und Kardinalität nicht genügen (vgl. Bamberg et al., 2017, S. 7; vgl. Heinrich & Klier, 2015, S. 49). Die Verwendung von Metriken setzt voraus, dass alle Metadaten zur Qualitätsbewertung (wie Informationen zum Datum) verfügbar und zugänglich sind. In der aktuellen Praxis ist dies nicht immer sichergestellt, da es kostspielig sein kann, alle Metadaten zu beschaffen. Die Bereitstellung hochwertiger Metadaten ist jedoch für eine effektive Bewertung der Datenqualität unerlässlich (vgl. Hinrichs, 2002, S. 69). Ein Verfahren zur Beurteilung der Qualität eines Datenbestandes ist das Hinrichs-Verfahren. Dieses entwickelt Metriken für einige ausgewählte Dimensionen.

In der Literatur werden von vielen Autoren die Dimensionen *Korrektheit, Vollständigkeit, Genauigkeit* und *Aktualität* genannt (vgl. Fox et al., 1994, S. 14–17; vgl. Hinrichs, 2002, S. 30–31; vgl. Pipino et al., 2002, S. 212; vgl. Rohweder et al., 2021, S. 26–27; vgl. Sidi et al., 2013, S. 302–303; vgl. Wand & Wang, 1996, S. 92; vgl. Wang & Strong, 1996, S. 20).

Aus diesem Grund werden die entwickelten Metriken von Hinrichs zu diesen Dimensionen nun vorgestellt. Die Metriken beziehen sich auf ein relationales Datenbankschema, sind aber laut Hinrichs auch auf objektorientierte oder objektrelationale Datenmodelle anwendbar. Zum Verständnis erfolgt ein Exkurs zu wichtigen Begriffen aus diesem Bereich.

Eine *Entität*, auch Tabellenname genannt, stellt einen Themenbereich dar, in dem Elemente mit gleichen Merkmalen eingeschlossen sind. Eine *Entitätsmenge* (Datensätze) umfasst alle zu den Merkmalen einer Entität gehörenden Werte. Dazu gehören alle gespeicherten Datensätze der Tabelle. Die Tabelle wird auch *Relation* genannt und besteht aus einer Entität mit der dazugehörigen Entitätsmenge. Berücksichtigt wird die gesamte Tabelle mit Entitätsbezeichnung, Attributen und Tupel. Alle Merkmalswerte eines Elements werden *Tupel* oder auch Datensatz genannt. Alle Tupel einer Entität bilden wiederum die Entitätsmenge. Das *Attribut* (Spaltenname) gleicht einem Merkmal eines Tupels und kennzeichnet eine bestimmte Eigenschaft einer Entitätsmenge. Der *Attributwert*, auch Wert oder Datum genannt,

ist ein Datenwert und beschreibt das zugehörige Attribut eines Tupels. Eine *Datenbasis* beinhaltet alle Tabellen und folglich die auf allen gespeicherten Daten. Die *Datenbank* beinhaltet die Datenbasis sowie das Datenbankverwaltungssystem (vgl. Steiner, 2017, S. 14–16). Eine Diskurswelt ist ein Ausschnitt der Realität; dieser wird benutzt, um nur relevante Aspekte für den Zweck der Anwendung einzubeziehen (vgl. Stock, 2001, S. 5).

Um die Ergebnisse besser zu skalieren, werden Indikatoren für verschiedene Granularitätsebenen (Attributwerte, Tupel, Relationen und Datenbanken) von unten nach oben entwickelt. Eine Metrik auf der n+1-Ebene (z. B. Tupelebene Vollständigkeit) wird basierend auf der zuvor erläuterten n-Ebene (Attributebene Vollständigkeit) definiert. Die Metriken sind auf ein Intervall ]0,1] oder [0,1] normiert. Dies vereinfacht die Aggregation von Metriken auf verschiedenen Ebenen und ermöglicht den Austausch von Formeln innerhalb der Metriken. Bei allen Metriken existiert eine Rangordnung der Werte und bei betrachteten Merkmalen verhalten sie sich monoton steigend bei steigender Qualität, sodass Vergleichsoperatoren wie „größer als“ verwendet werden können. Die Messmethode, die die Operationalisierung misst, wird schließlich der feinkörnigsten Messung (bezogen auf Qualitätsmerkmale) zugeordnet. Die so ermittelten Messwerte werden dann auf einer gröberen Ebene gewichtet und entsprechend der Metrikdefinition gemittelt. Die Richtigkeit der Daten kann nur durch den Vergleich des Zustands der Diskurswelt (Sollzustand) und des Zustands des Informationssystems (Istzustand) bewertet werden. Als Beispiel nennt Hinrich hier die Inventur. Der tatsächliche Artikelbestand wird im Lager ermittelt und mit den Bestandsinformationen im Informationssystem abgeglichen. Die Korrektheit wird anhand der Ähnlichkeit zwischen dem Attributwert des Datenprodukts und dem Attributwert der in der Diskurswelt repräsentierten Entität bewertet (vgl. Hinrichs, 2002, S. 69–71).

Zuerst wird die Dimension *Vollständigkeit* überprüft. Vollständigkeit wird bewertet über die (Nicht-)Verfügbarkeit eines semantisch von „unbekannt“ abweichenden Wertes pro Tupelattribut.

Auf *Attributwertebene* ist $\omega$ ein Attributwert und $NotNull$ eine Funktion mit:

$$NotNull(\omega) := \begin{cases} 1 \; falls \; \omega = NULL \; oder \; \omega \; zu \; NULL \; äquivalent \\ 0 \qquad\qquad sonst. \end{cases} \tag{2}$$

Dann gilt:

$$Q_{Voll}(\omega) := NotNull(\omega)$$

*(3)*

Nicht einbezogen wird, dass ein Wert $\omega$ mit $\omega = NULL$ nicht unbedingt in der Diskurswelt existiert oder es nicht bekannt ist, ob dieser Wert vorhanden sein kann (z. B. Name des Ehepartners bei Ledigen). Es liegt dann kein Qualitätsdefizit vor bzw. es kann keine Qualitätsaussage getroffen werden.
Nun wird die *Tupelebene (einelementiger View)* betrachtet. $t$ ist ein Tupel mit Attributwerten $t.A_1, \ldots, t.A_n$ für Attribute $A_1, \ldots, A_n$. $g_j$ ist die zu bestimmende relative Wichtigkeit $A_j$ im Hinblick auf die Vollständigkeit. Relative Wichtigkeiten müssen für einen konkreten Anwendungsfall charakterisiert werden.

$$Q_{Voll}(t) := \frac{\sum_{j=1}^{n} Q_{Voll}(t.A_j) g_j}{\sum_{j=1}^{n} g_j}$$

*(4)*

Auf *Relationenebene (mehrelementiger View)* ist $T$ ist eine nicht leere Relation (oder ein View über mehrere Relationen). Das arithmetische Mittel der Konsistenzwerte der Tupel $t_i \in T \ \ (i = 1, \ldots, |T|)$ ist der Konsistenzwert von $T$:

$$Q_{Voll}(T) := \frac{\sum_{i=1}^{|T|} Q_{Voll}(t_i)}{|T|}$$

*(5)*

Zuletzt wird die *Datenbankebene* betrachtet. $D$ ist eine Datenbank, die durch eine disjunkte Überdeckung einer Menge von Relationen wie folgt definiert ist: $T_k \in D \ (k = 1, \ldots, p)$. Disjunktheit verhindert, dass Relationeninhalte mehrfach in der Bewertung vorkommen.

$$D := T_1 \cup T_2 \cup \ldots \cup T_P \text{ mit } T_1 \cap T_2 \cap \ldots \cap T_P = \emptyset$$

$R$ ist der durch $D$ modellierte Ausschnitt der Diskurswelt.
Der Vollständigkeitswert von $D$ wird durch das arithmetische Mittel der Vollständigkeitswerte der Relationen definiert:

$$Q_{Voll}(D) := \frac{\sum_{k=1}^{p} Q_{Voll}(T_k)}{p} \tag{6}$$

Im Folgenden wird die Dimension *Korrektheit* mit einer Metrik versehen. Die Betrachtung erfolgt zuerst auf *Attributwertebene.*

$\omega_I$ ist ein Attributwert im Informationssystem, $\omega_R$ ist ein mit $\omega_I$ korrespondierender Attributwert einer modellierten Entität in der Diskurswelt, $d$ ist ein Abstandsmaß und $Q_{Korr}$ die Qualität der Korrektheit

Beispiel 1:

$$d(\omega_1, \omega_2) := \begin{cases} 0 & falls\ \omega_1 = \omega_2 \\ \infty & sonst. \end{cases} \tag{7}$$

Beispiel 2:

$$d(\omega_1, \omega_2) := |\omega_1 - \omega_2| \tag{8}$$

$$Q_{Korr}(\omega_I, \omega_R) := \frac{1}{d(\omega_I, \omega_R) + 1} \tag{9}$$

Auf *Tupelebene (einelementiger View)* ist $Q_{Korr}$ das gewichtete arithmetische Mittel der Qualitätswerte auf Attributebene des Tupels $t$ entsprechend der Entität $e$:

$$Q_{Korr}(t, e) := \frac{\sum_{j=1}^{n} Q_{Korr}(t.A_j, e.A_j)}{\sum_{j=1}^{n} g_j} \tag{10}$$

$t$ ist ein Tupel, das durch eine Selektion einer einelementigen Ergebnismenge sowie eine Projektion auf eine beliebige, nicht leere Teilmenge von Attributen $A_1, \ldots, A_n$ definiert ist. $t.A_j$ ist der Wert von $t$ im Attribut $A_j$ ($j = 1, \ldots, n$). $e$ ist eine Entität der Diskurswelt und $e.A_j$ ist der Wert von $e$ im Attribut $A_j$. $g_j$ ist die relative Wichtigkeit der Korrektheit von $t.A_j$. Mithilfe einer vorab bestimmten Gewichtung lässt sich modellieren, dass Attribute eine größere Bedeutung haben als andere (z. B. Geburtsjahr wichtiger als Körpergröße).

Nun wird die *Relationenebene (mehrelementiger View)* erfasst. Das arithmetische Mittel der Korrektheitswerte der Tupel $t_i \in T$ bzgl. den Entitäten $e_i \in E$ $(i = 1, \ldots, |T|)$ ist der Korrektheitswert von $T$ bzgl. $E$:

$$Q_{Korr}(T, E) := \frac{\sum_{i=1}^{|T|} Q_{Korr}(t_i, e_i)}{|T|} \tag{11}$$

$T$ ist eine nicht leere Relation bzw. ein mehrelemtiger View, der über eine Menge von Relationen definiert ist. $E$ ist eine Menge von Entitäten der Diskurswelt, die von $T$ repräsentiert wird.

Auf der *Datenbankebene* ist $D$ eine Datenbank wie bei der Dimension Korrektheit definiert. Das arithmetische Mittel der Korrektheitswerte der Relationen in $D$ bzgl. ihrer korrespondierenden Entitätsmengen in $R$ ist der Korrektheitswert von $D$ bzgl. $R$:

$$Q_{Korr}(D, R) := \frac{\sum_{k=1}^{p} Q_{Korr}(T_k, E_k)}{p} \tag{12}$$

Die Dimension *Genauigkeit* wird maßgeblich durch die sogenannte Stelligkeit von Attributwerten beeinflusst. Die Stelligkeit meint nicht nur die Anzahl der Nachkommastellen bei numerischen Attributen, sondern auch bei symbolischen Werten die Position in einer Klassifikationshierarchie.

Die Genauigkeit wird auf *Attributwertebene* über das Verhältnis seiner Stelligkeit zur jeweils idealen Stelligkeit bewertet, die vorab nach dem Anwendungskontext definiert wird. Da ideal nicht unbedingt maximal bedeutet, muss das Intervall auf $]0,1]$ normiert werden.

$D$ ist ein numerisches Attribut, $S_{opt}(A)$ die ideale Anzahl von (Nachkomma-)Stellen für $A$ und $\omega$ ein Attribut von $A$. Die Funktion $s : Dom(A) \rightarrow \mathbb{N}$ liefert die Stelligkeit von Werten $\omega \in Dom(A)$ zurück. Dann gilt für die Genauigkeit von $\omega$ im Attribut $A$:

$$Q_{Gen}(\omega, A) := min\left(\frac{s(\omega)}{s_{opt}(A)}, 1\right) \tag{13}$$

Wenn $A$ ein symbolisches Attribut ist und sich $\omega$ in Ebene $i$ einer Klassifikation $K$ (attributbezogenes Metadatum) mit $n$ Ebenen $(K_1, \dots, K_n)$ einordnen lässt (also auf Ebene $K_i$), gelte $s_{opt}(A) \leq n$ und $s(\omega) = i$. Wenn für ein symbolisches Attribut $A$ keine Klassifikationshierarchie gibt, gilt $s_{opt}(A) = 1$ und $s(\omega) = 1$.
Auf *Tupelebene (einelementiger View)* ist $t$ ein Tupel mit Attributwerten $t.A_1, \dots, t.A_n$ für Attribute $A_1, \dots, A_n$. $g_j$ ist eine relative Wichtigkeit von $A_j$ mit Blick auf die Genauigkeit, die vorher zu bestimmen ist. Dann gilt:

$$Q_{Gen}(t) := \frac{\sum_{j=1}^{n} Q_{Gen}(t.A_j, A_j) g_j}{\sum_{j=1}^{n} g_j} \tag{14}$$

$Q_{Gen}(T)$ wird auf *Relationenebene (mehrelementiger View)* zu $Q_{Voll}(T)$ definiert (siehe Kapitel 0).

Auf *Datenbankebene* wird $Q_{Gen}(D)$ analog zu $Q_{Voll}(D)$ definiert (siehe Kapitel 0).

Zur Beurteilung der Dimension *Aktualität (Zeitnähe)* wird das sogenannte Befunddatum von Attributwerten oder Tupel im Verhältnis zur Update-Häufigkeit geprüft. Das Befunddatum ist der Zeitpunkt der Eigenschaft in der Diskurswelt und ist nicht automatisch der Zeitpunkt des Eintrags in die Datenbank. An dieser Stelle wird davon ausgegangen, dass das Diagnosedatum bekannt ist und explizit als Metadatum bereitgestellt wird. Wenn Metadaten nicht vorliegen, kann die Zeitnähe nicht geprüft werden. Durch die Aufnahme der Update-Häufigkeit kann die Dynamik von Wertänderungen einbezogen werden.
Auf Attributwertebene ist $A$ ein Attribut und $\omega$ ein Attributwert in $A$. Das Alter ist $Age(\omega)$ (Differenz zwischen aktuellem Zeitpunkt und Befunddatum) von $\omega$ und $Upd(\omega)$ die Update-Häufigkeit von Werten in $A$. $u$ ist die Zeiteinheit (z. B. min, Tage, Monate), in der die Update-Häufigkeit und das Alter angegeben wurden.
Für die Zeitnähe von $\omega$ im Attribut $A$ gilt dann:

$$Q_{Zeit}(\omega, A) := \frac{1}{\frac{Upd(A)}{u} Age(\omega) u + 1} = \frac{1}{Upd(A) Age(\omega) + 1} \tag{15}$$

$Upd(\omega) = 0$ gilt für Attribute, deren Werte sich nicht ändern (z. B. Geburtsdatum). Für diese gilt stets $Q_{Zeit}(\omega, A) = 1$. Eine hohe Update-Häufigkeit impliziert einen schnell alternden Attributwert (z. B. Börsenkurs).
Eine andere Messung zur Aktualität stammt von Ballou. Anstatt einer Update-Häufigkeit wird mit einer maximalen Gültigkeitsdauer $T_{max}(A)$ gerechnet, die als bekannt angenommen wird (vgl. D. Ballou et al., 1998, S. 468).

$$Q_{Zeit}(\omega, A) \coloneqq \left[max\left\{1 - \frac{Age(\omega)}{T_{max}(A)}; 0\right\}\right] \tag{16}$$

Klier benutzt in seiner Metrik eine Verfallsrate des Attributwerts. $Verfall(A)$ zeigt, wie viele Datenwerte eines Attributs durchschnittlich inaktuell werden (vgl. Klier, 2008, S. 231).

$$Q_{Zeit}(\omega, A) \coloneqq e^{-Verfall(A) * Age(\omega)} \tag{17}$$

Auf *Tupelebene (einelementiger View)* ist $t$ ein Tupel mit Attributwerten $t.A_1, \dots, t.A_n$ für Attribute $A_1, \dots, A_n$. Die relative Wichtigkeit von $A_j$ ist $g_j$. Somit gilt:

$$Q_{Zeit}(t) \coloneqq \frac{\sum_{j=1}^{n} Q_{Zeit}(t.A_j, A_j) g_j}{\sum_{j=1}^{n} g_j} \tag{18}$$

Wenn das Befunddatum nur tupelbezogen als Metadatum vorliegt, kann das Alter dementsprechend nur auf Tupelebene berechnet werden. Die Update-Häufigkeit von $t$ wird dann wie folgt als maximale Update-Häufigkeit der beteiligten Attribute definiert:

$$Upd(A) \coloneqq max_{j=1}^{n}\left(Upd(A_j)\right) \tag{19}$$

Für die Zeitnähe gilt dann:

$$Q_{Zeit}(t) \coloneqq \frac{1}{Upd(t)Age(t) + 1} \tag{20}$$

Auf *Relationenebene (mehrelementiger View)* und auf *Datenbankebene* wird $Q_{Zeit}(T)$ analog zu $Q_{Voll}(T)$ beziehungsweise $Q_{Zeit}(D)$ zu $Q_{Voll}(D)$ definiert.

### 2.2.3. Datenqualitätsregelkreis

Der Regelkreis veranschaulicht den Zusammenhang zwischen Dimensionen, der Datenqualität, den Maßnahmen und den dazugehörigen Kosten sowie dem Nutzen (siehe **Fehler! Verweisquelle konnte nicht gefunden werden.**). Die Maßnahmen s tellen eine Art Regler dar, mit denen der Datenqualitätsregelkreis beeinflusst werden kann. Die Umsetzung von Maßnahmen soll zu Verbesserungen der Datenqualität – gemessen an Metriken – und entsprechendem wirtschaftlichen Nutzen führen. Vielmehr ist es auch möglich, ausgehend von einem bestimmten Niveau den Anstieg der Datenqualität durch geeignete Maßnahmen anhand der Metriken abzuschätzen bzw. zu messen. Bei wirtschaftlichen Kriterien muss die Wahl der Maßnahmen jedoch Kosten-Nutzen-Faktoren berücksichtigen (vgl. Heinrich & Klier, 2021, S. 48).

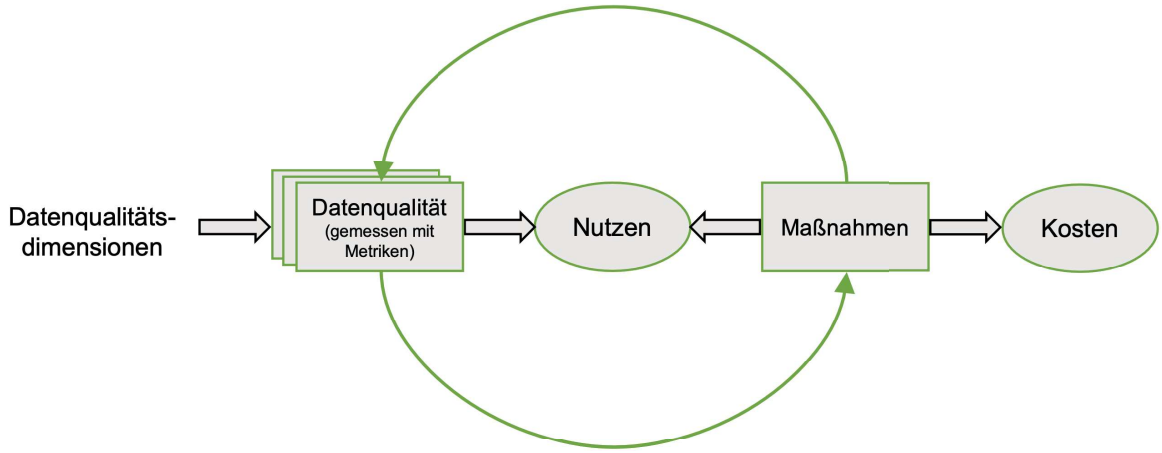


Abbildung 6 Da atenqualitätskreis, in Anlehnung an Heinrich & Klier, 2021, S. 48

### 2.2.4. Datenqualitätsmängel

In diesem Abschnitt werden die bedeutsamsten Datenqualitätsprobleme klassifiziert, die durch Datenbereinigung und Datentransformation zu lösen sind. Datenqualitätsmängel lassen sich in zwei verschiedene Gruppen aufteilen, zum einen Single-Source-Probleme (eine Datenbasis), zum anderen Multi-Source-Probleme (mehrere heterogene, sich ggf. inhaltlich überlappende Datenbasen). In jeder dieser Gruppen kann eine Klassifizierung in schema- und instanzbezogene Probleme vollzogen werden (Tabelle 3) (vgl. Rahm & Do, 2000, S. 2–3).

Tabelle 3 Klassifikation von Datenqualitätsproblemen, in Anlehnung an vgl. Rahm & Do, 2000, S. 3

| | Datenqualitätsprobleme | | | |
|---|---|---|---|---|
| | Single-Source-Probleme | | Multi-Source-Probleme | |
| | Schema- ebene | Instanzebene | Schema-ebene | Instanzebene |
| | Unzureichende Integritäts-bedingungen, mangelhaftes Schemadesign | Dateneingabefehler | Heterogene Datenmodelle und Schemaentwürfe | Inhaltliche Überlappung, unterschiedliche Erfassungsmethoden |
| Beispiele | Nichteindeutigkeit von Schlüsseln<br>Mangelhafte referenzielle Integrität<br>Werte außerhalb des zulässigen Wertebereichs | Duplikate<br>Rechtschreibfehler<br>Falsche Werte<br>Fehlende Werte<br>Widersprüchliche Werte<br>Deplatzierte Werte | Namenskonflikte<br>Strukturelle Konflikte | Redundanz/Duplikate<br>Widersprüchliche Werte |

Die anfängliche Dateneingabe ist eine der wichtigsten Ursachen für eine unzureichende Datenqualität. Es gibt viele Gründe für Fehler bei der Datenerfassung. Es fängt bei einem einfachen Benutzerfehler an und reicht bis hin zum Mangel an Ressourcen, um in unterschiedlichen Anwendungen die korrekte Datenbeschaffung sicherstellen zu können. Benutzerfehler lassen sich beispielsweise in falsche Eingaben, falsche Interpretationen und Bedienung der Anwendung der zu erfassenden Daten unterteilen. Einer der wesentlichen Gründe für eine schlechte Datenqualität

im Datenerhebungsprozess ist das fehlende Bewusstsein über mögliche Folgen von Datenqualitätsmängel. Fehlende oder unvollständig implementierte Prozesse können zu einer fehlerhaften Datenerhebung führen. Der Mangel an Verifikationsprozessen, die zur Vermeidung von Fehleingaben dienen, verstärken Probleme in der Datenerhebung (vgl. Apel et al., 2015, S. 28–30).

# 3. Big Data

Der Begriff Big Data ist bis heute nicht einheitlich definiert und wird verwendet, um eine Reihe von unterschiedlichen Konzepten zu beschreiben. Dies reicht von der Sammlung und Speicherung umfangreicher Datenmengen bis hin zu digitalen Techniken, die Muster im menschlichen Verhalten aufdecken (vgl. Favaretto et al., 2020, S. 1). Viele große Erwartungen richten sich an Big Data, vor allem zielen diese auf neue Möglichkeiten der Speicherung, Verarbeitung und Analyse in datenintensiven Branchen ab (vgl. Buhl et al., 2013, S. 63). Aus technischer Sicht werden unter Big Data große Datenmengen verstanden, die un- oder semistrukturiert sind und in verschiedenen Formaten vorliegen. Diese Datenmenge unterliegt keiner genauen oder einheitlichen Beschreibung sowie keinen Einschränkungen der thematischen Herkunft oder des Datentyps, zum Beispiel Video-, Maschinen- und Audiodaten. (vgl. Dorschel, 2015, S. 307) Die betrachteten Datenmengen sind nicht nur groß und komplex, sondern sie ändern sich auch schnell, sodass bisherige Methoden, diese zu sammeln, zu verarbeiten und auszuwerten, nicht mehr ausreichen (vgl. Jaekel, 2017, S. 92). Das Vorkommen von umfangreichen Datenmengen ist weitaus älter, als es zu sein scheint. Big Data, das in älteren Debatten auch „prozessproduzierte Massendaten“ genannt wurde, gibt es seit über 200 Jahren (vgl. Baur, 2009, S. 10). Nutzer dieser Daten waren damals schon Wissenschaftler, aber auch andere Bereiche verwendeten diese Formen. Beispiele dafür sind Volkszählungen, öffentliche Verwaltungsdaten oder Kirchenbücher (vgl. Baur et al., 2020, S. 210). Einige Definitionen legen einen, wenn auch unklaren, Schwellenwert fest, oberhalb dessen die Datensätze groß werden. Große Datensätze wurden beispielsweise als „zu groß, um auf den Computer eines Analysten zu passen“ (vgl. Vo & Silva, 2016, S. 125), beschrieben. Big Data ist somit ein Sammelbegriff zur Erhebung, Speicherung und Auswertung von Massendaten mittels digitaler Technologien (vgl. Otte et al., 2018,

S. 14). Schätzungen zufolge verdoppeln sich alle Datenmengen weltweit alle zwei Jahre. Anderen Experten zufolge besteht das digitale Universum aus 44 Zettabyte an digitalen Daten im Jahr 2021 und soll bis 2025 auf 163 Zettabyte steigen. Der Begriff Zettabyte ist noch nicht verbreitet, er entspricht 10^21 Byte bzw. einer Milliarde (10^9) Terabyte (10^12) (vgl. Gutierrez, 2021, S. 3; vgl. Otte et al., 2018, S. 14). Kennzeichnend für Big Data sind jedoch die unterschiedlichsten Datenformate, die weltweit auftreten können. Nur ein kleiner Teil aller digitalen Daten ist in einer relationalen Datenbank organisiert. Die meisten Daten sind unstrukturierter Text, Bilder oder Töne. Eine der größten Herausforderungen besteht darin, aus dieser unstrukturierten und sich ständig ändernden Datenmenge möglichst schnell Wissen zu generieren, um eine Entscheidungsgrundlage für Echtzeitaktionen für vernetzte Systeme zu schaffen (vgl. Otte et al., 2018, S. 14–15). Bei komplexen Phänomenen werden kleine Datenmengen keine zuverlässigen Vorhersagen machen können. Zu beachten ist auch, dass nach dieser Definition manche Big-Data-Datensätze nicht als Big Data gelten, das heißt, wenn sie keine verlässlichen Vorhersagen zulassen oder nur einfache Phänomene darstellen. Diese Wahl der Definition konzentriert sich auf erkenntnistheoretische Fragen und stellt klar fest, dass der Begriff „Big Data“ nur auf Situationen anwendbar ist, in denen die Daten wissenschaftlich nützlich sind. Die Beziehung zwischen Small Data und Big Data ist in Abbildung 7 dargestellt. Schließlich betont die Definition, dass die epistemologische Bedeutung von Big Data nur durch spezifische induktive Methoden bestimmt werden könne. Ein Datensatz ist somit „groß“, wenn er groß genug ist, um zuverlässige Vorhersagen auf der Grundlage induktiver Methoden in einem Bereich mit komplexen Phänomenen zu ermöglichen (vgl. Pietsch, 2021, S. 15).

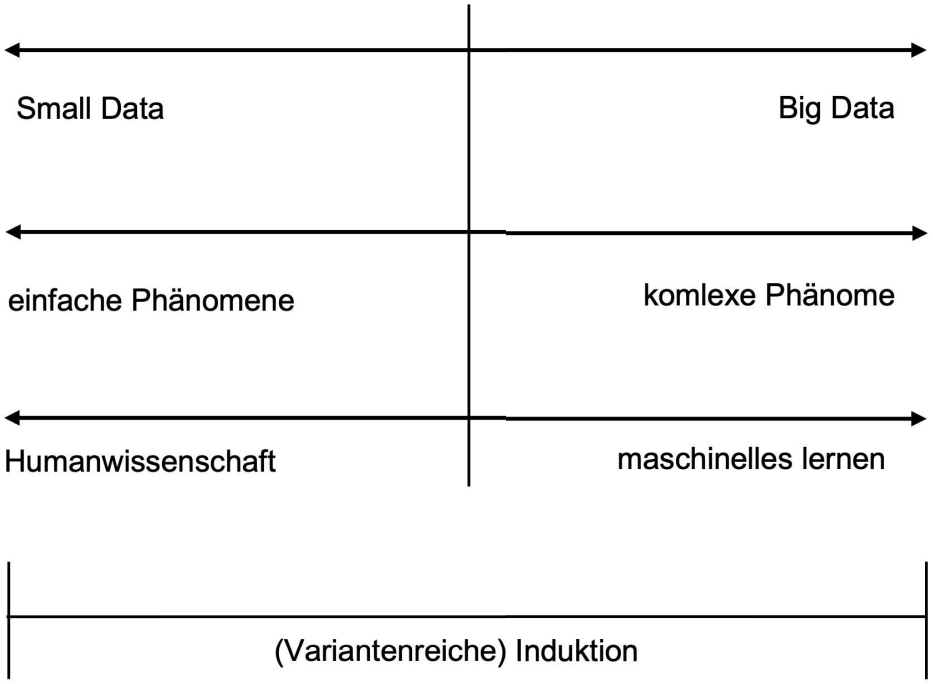


Abbildung 7 Big Data vs. Small Data in der wissenschaftlichen Forschung, in Anlehnung an Pietsch, 2021, S. 15

## 3.1. Merkmale von Big Data

Laney verfasste 2011 mit seiner Ausarbeitung den Ausgangspunkt der beschreibenden Merkmale von Big Data. Die 3 V stehen für die Anfangsbuchstaben der englischen Begriffe Volume (Menge), Velocity (Geschwindigkeit) und Variety (Vielfalt) (vgl. Laney, 2001, S. 949). Big Data wird zwar in der Regel anhand der drei Vs definiert, d. h., dass Daten, die ein großes, temporär anfallendes Volumen haben, in sehr kurzer Zeit verarbeitet werden müssen und in einer Vielzahl unterschiedlicher Formate bereitgestellt werden (vgl. Otte et al., 2018, S. VI). Jedoch zeigt ein Ergebnis verschiedener Interviews, dass die Meinungen über die Anzahl der Vs auch bei Experten auseinandergehen (vgl. Favaretto et al., 2020, S. 1). Gelegentlich werden weitere Merkmale wie Veracity oder Value hinzugefügt (vgl. Vo & Silva, 2016, S. 125).

*Volume* gilt als erstes der Vs und beschreibt die Menge bzw. den Umfang der Datenmenge, die aufgezeichnet, analysiert und verwaltet werden muss. Die Datenmenge steigt mit der Anzahl der Quellen und der Datentiefe (vgl. King, 2014, S. 35).

*Velocity* ist die Geschwindigkeit, mit der Daten generiert und geändert werden. Dieses erfordert eine Echtzeitanalyse und die Auswertung des Datenflusses. Die schnelllebige Datengenerierung wird durch die Anzahl der Datenquellen und die erhöhte Rechenleistung der datenerzeugenden Geräte beeinflusst (vgl. King, 2014, S. 35). Die Formen der Datenerfassung, die Ladezeiten widerspiegeln und aktualisieren, wirken sich auch auf Änderungszyklen aus. Immer relevanter werden Systeme, in denen Änderungen in Echtzeit erfasst und ggf. sogar ausgewertet werden können (vgl. Deutscher Dialogmarketing Verband e. V, 2016, S. 16). In Big Data bezieht sich die Vielfalt (*variety*) auf die Speicherung von strukturierten, semistrukturierten und unstrukturierten Multimediadaten (Text, Grafiken, Bilder, Audio und Video) (vgl. Meier & Kaufmann, 2016, S. 13). Daten kommen zunehmend aus neuen Quellen innerhalb und außerhalb der Organisation. Sie sind unterschiedlich aufgebaut und können auch bisher unbekannte Strukturformen aufweisen (vgl. King, 2014, S. 35). Zudem müssen unterschiedliche Datenspeicherorte berücksichtigt werden, denn neben der lokalen Datenspeicherung werden immer mehr Daten in der Cloud gespeichert. Darüber hinaus ist die Vielfalt der Datenhaltungssysteme zu berücksichtigen, zum Beispiel hinsichtlich Datenmodellen, Datenbanktypen, Soft- und Hardware. Schließlich haben die Daten je nach Speicherort auch eine unterschiedliche Datenverfügbarkeit (vgl. Deutscher Dialogmarketing Verband e. V, 2016, S. 16). *Veracity* bedeutet in der deutschen Übersetzung Aufrichtigkeit oder Wahrhaftigkeit. Bezüglich Big Data wird davon ausgegangen, dass es Datenbestände unterschiedlicher Datenqualität gibt, die bei der Auswertung berücksichtigt werden müssen. Veracity befasst sich mit der Unsicherheit von Daten aufgrund verschiedener Faktoren wie Dateninkonsistenzen, Unvollständigkeit, Mehrdeutigkeit und absichtlicher Täuschung. Neben statistischen Verfahren gibt es unscharfe Methoden des Soft Computing, die Ergebnisse oder Aussagen Wahrheitswerte zwischen „richtig“ und „falsch“ zuordnen. Da viele Daten missverständlich oder ungenau sind, muss ein bestimmter Algorithmus verwendet werden, um den Wert der Informationen oder die Qualität der Ergebnisse zu bewerten. Eine große Datenbasis allein kann keine bessere Auswertungsqualität garantieren. *Value* bezeichnet den Mehrwert, der durch die Daten geniert werden soll. Es wird oft als fünftes V bezeichnet und ruft die Notwendigkeit hervor, dass Big-Data-Anwendungen aus den rohen und unverarbeiteten Daten einen Nutzen für die jeweilige Verwendung herbeiführen

(vgl. Debattista et al., 2015, S. 92; vgl. S. King, 2014, S. 35; vgl. Meier & Kaufmann, 2016, S. 13).

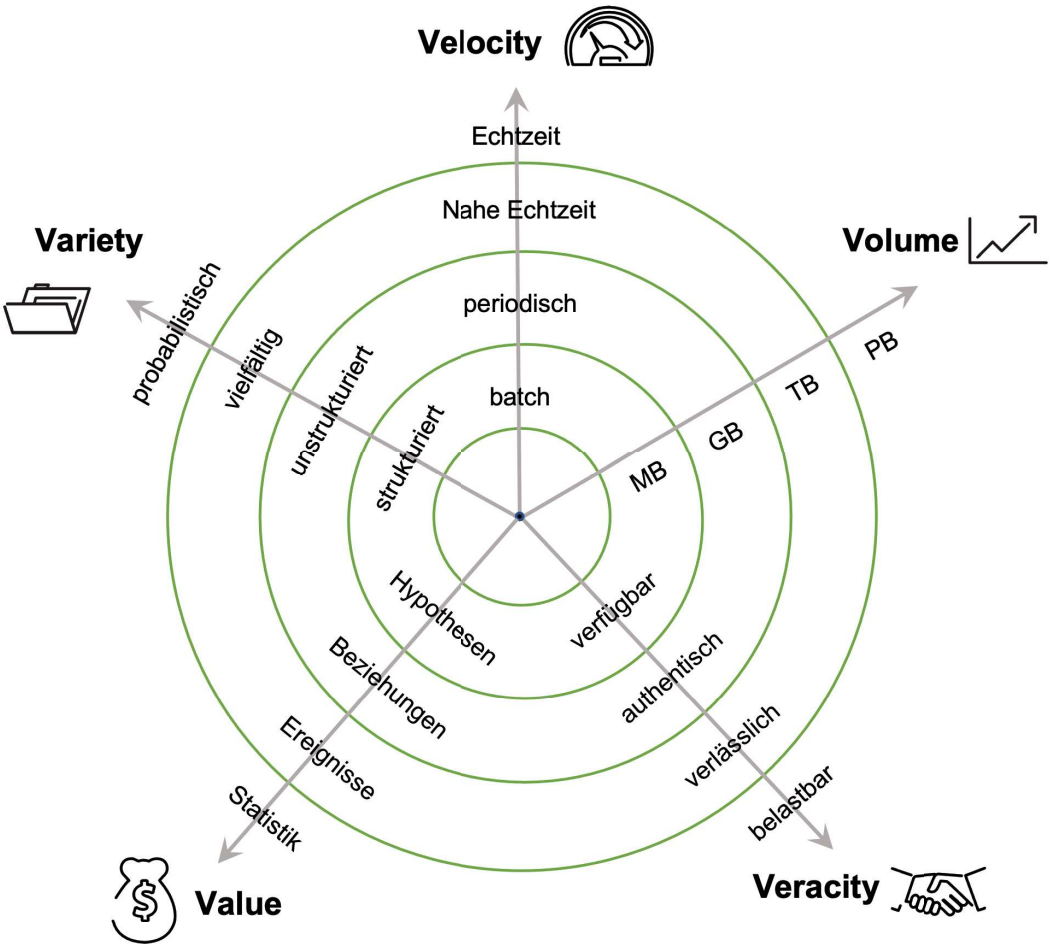


Abbildung 8 Charakteristika von Big Data, in Anlehnung an Gölzer, 2017, S. 49; Klein et al., 2013, S. 320

## 3.2. Datentypen

Neben der ausführlichen Erläuterung des Begriffes „Daten“ und der Kategorisierung innerhalb der Begriffe „Wissen“, „Information“ und „Zeichen“ ist es aufgrund der diffizilen Beschaffenheit notwendig, den Begriff auch im Themenfeld von „Big Data“ aufzugreifen. Computergestützte Definitionen von Daten gehen von der zentralen Rolle aus, dass die Daten für den Aufstieg der Informations- und Kommunikationstechnologie in den vergangenen Jahrzehnten verantwortlich sind. Diese Definitionen versuchen, eine Auffassung von „Daten“ auf der Grundlage der aktuellen Verwendung des Begriffs in vielen empirischen Wissenschaften und in der Informatik

zu erfassen. In der Regel beziehen sich diese Definitionen auf das Medium oder das Format, in dem sie gespeichert sind. Computergestützte Definitionen knüpfen an eine lange Tradition an, in der der Begriff Daten als Markierungen oder Spuren definiert wurde. Damit sind nicht interpretierte Eintragungen gemeint, die zum Beispiel die Ergebnisse eines Experiments oder Manipulationen dieser Ergebnisse sein können (vgl. Hacking, 1992, S. 43-45.; vgl. Leonelli, 2016, S. 75–76) Wie bereits mehrfach betont wurde, sollten grundlegende Begriffe wie Daten nicht unabhängig vom konkreten Kontext, in dem sie verwendet werden sollen, definiert werden. Somit geht es nun um die Beantwortung erkenntnistheoretischer Fragen zu Big Data. Daher muss das richtige Konzept von Daten in das Gesamtbild integriert werden. Nachfolgend werden Merkmale von Daten aufgezählt, die im Kontext von Big Data von Bedeutung sind. Daten sind nicht die Fakten selbst, sondern vielmehr Spuren oder Zeichen dieser Fakten. Spuren oder Zeichen müssen auf ein physisches Medium geschrieben und gespeichert werden. Um Spuren oder Zeichen zu aggregieren und zu analysieren, müssen sie eine gewisse Persistenz auf dem Medium haben. Dies ist bei Fakten nicht der Fall.  Die Fakten beziehen sich auf ein Phänomen, das von seinem Betrachter aus einer Reihe von Möglichkeiten ausgewählt wird. Es handelt sich um singuläre Fakten, die sich auf eine bestimmte Ausprägung des interessierenden Phänomens beziehen, z. B. ein einzelnes Ereignis oder Objekt. Die Spuren oder Zeichen haben eine kausale und definitorische Beziehung zu den Fakten. Aufgrund dieser Beziehungen stellen die Daten die Fakten zumindest teilweise dar. Da sich die Daten von den Fakten unterscheiden, müssen sie interpretiert werden, um aus den Daten etwas über das interessierende Phänomen zu erfahren. Die Daten müssen in einer Form vorliegen, die eine geeignete wissenschaftliche Methode auf die Daten anzuwenden erlaubt. Mithilfe einer wissenschaftlichen Methode können wiederum Daten als Beweis für Phänomene dienen (vgl. Pietsch, 2021, S. 11–12)

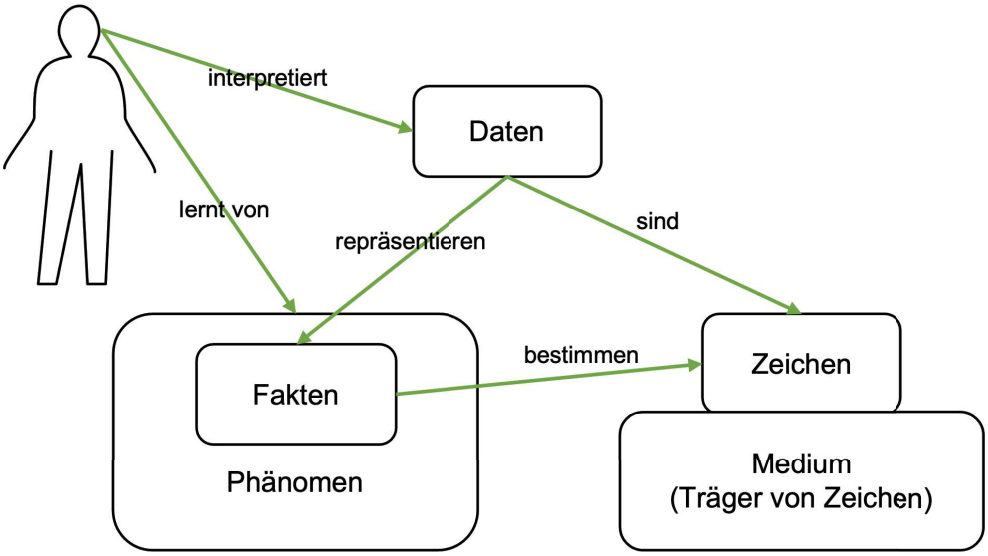


Abbildung 9 Darstellung von Daten, in Anlehnung an Pietsch, 2021, S. 11

Forscher haben unterschiedliche Klassifizierungen für Daten in verschiedenen Bereichen vorgenommen. Im Bereich der Datenqualität unterscheiden die meisten Autoren drei Arten von Daten (vgl. Batini et al., 2009, S. 9). Diese sind strukturierte, unstrukturierte und semistrukturierte Daten und wurden im Kapitel 2.1.2 ausführlicher erläutert. Eine weitere Klassifizierung in der Datenqualitätsliteratur betrachtet Daten als ein Herstellungsprodukt. Zum einen gibt es *Rohdaten*, die seit ihrer Erstellung und Speicherung nicht verarbeitet worden sind. Zum anderen existieren *Informationsprodukte*, die das Ergebnis einer Verarbeitung der Daten sind. Des Weiteren unterteilt Batini Daten in *Komponentendaten*, die jedes Mal erzeugt werden, wenn das entsprechende Informationsprodukt bearbeitet wird, bis das Endprodukt hergestellt ist (vgl. Batini et al., 2009, S. 10). Eine dritte Klassifizierung von Daten basiert auf der Betrachtung von Daten als Produkt. Dieses Modell klassifiziert Daten in drei Typen. Eine weitere Klassifizierung von Daten basiert auf der Strenge, mit der die Datenqualität gemessen und erreicht wird, und umfasst zwei Klassen, nämlich elementare Daten und aggregierte Daten. In einer Organisation werden Daten, die durch operative Prozesse verwaltet werden und atomare Phänomene der realen Welt repräsentieren, als elementare Daten bezeichnet (z. B. Geschlecht, Alter), während Daten, die aus elementaren Daten für die Anwendung der Aggregationsfunktion gesammelt werden, als aggregierte Daten bezeichnet werden (z. B. durchschnittliches Einkommen, das Steuerzahler in einer bestimmten Stadt gezahlt haben). Die dritte Klassifizierung unterscheidet Daten in elementare Daten und ag-

gregierte Daten. *Elementare Daten* werden in Unternehmen durch betriebliche Prozesse verwaltet und stellen Phänomene der realen Welt dar (z. B. Sozialversicherungsnummer, Alter, Geschlecht). *Aggregierte Daten* werden aus einer Sammlung von elementaren Daten gewonnen, indem eine Aggregationsfunktion auf sie angewendet wird (z. B. das Durchschnittseinkommen der Steuerzahler in einer bestimmten Stadt) (vgl. Batini & Scannapieco, 2016, S. 7).

## 3.3. Datenquellen

Klassische Systeme übermitteln strukturierte Datensätze, wie beispielsweise Kundenaufträge und Aufträge aus Enterprise-Resource-Planning-Systemen, Supply-Chain-Management-Systemen und Customer-Relationship-Management-Systemen. Neue Anwendungen sind beispielsweise Sensor- oder Logdaten, um aktuelle Daten aus der Produktion oder aufgezeichnete Wetterdaten zu generieren. Zusätzlich können beispielsweise in mobilen Anwendungen mithilfe der RFID-Technologie mobile Daten aus dem Logistiksystem und Verkehrsdaten generiert werden. Social-Web-Anwendungen wie Facebook und Twitter generieren ebenfalls sehr große Datenmengen, die sich für die Big-Data-Analyse eignen (vgl. Gadatsch, 2017, S. 5). Typische Datenquellen, die in einem Big-Data-Szenario verarbeitet werden müssen, befinden sich in Abbildung 10.

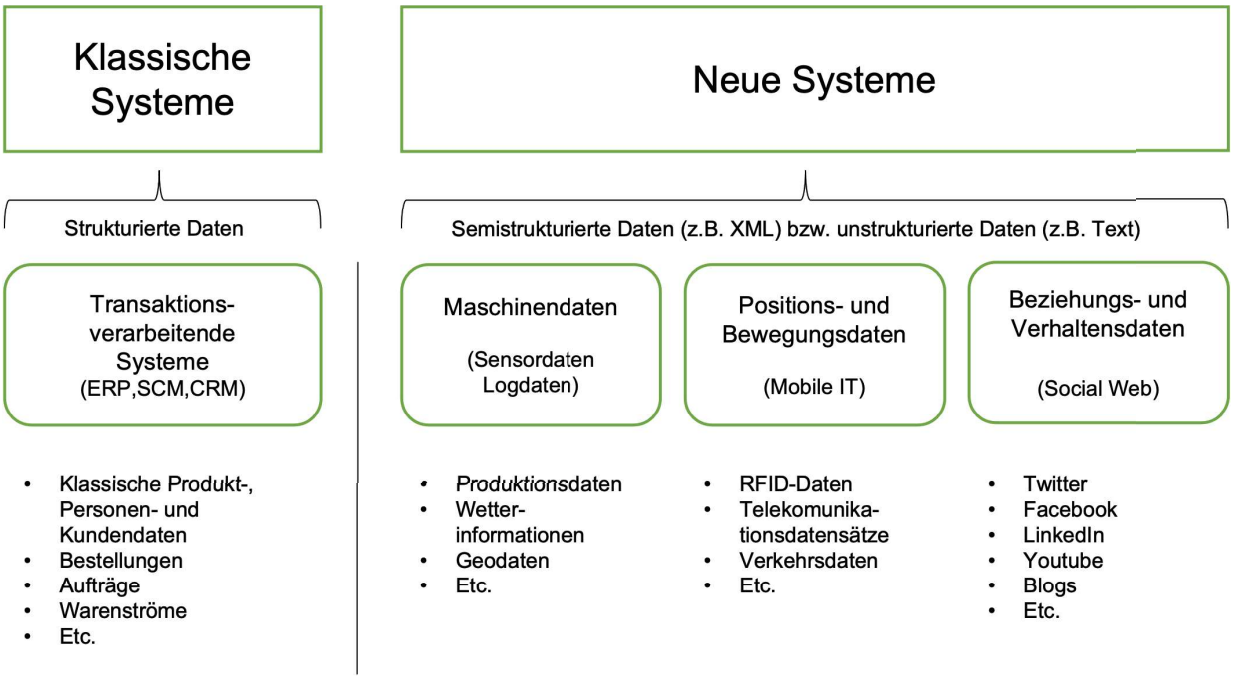


Abbildung 10 Ausgewählte Datenquellen für Big Data, in Anlehnung an Gadatsch, 2017, S. 6

# 4. Datenqualität in Big Data

Im praktischen Teil der Arbeit werden zuerst die Anforderungen von Big Data erarbeitet, da diese die Voraussetzung für die weitere Analyse sind. Nach den Anforderungen wird eine Begriffsabgrenzung von Daten- und Informationsqualität vorgenommen. Im Anschluss werden anhand der Definitionen die Dimensionen aus der Literatur aufgegriffen und zugeordnet. Im letzten Kapitel werden neue Definitionen für die Begriffe aufgestellt. Dieses erfolgt im letzten Schritt, da die Dimensionen die Definitionen charakterisieren.

## 4.1. Anforderungen an die Datenqualität in Big Data

Die Entwicklung eines Datenqualitätskonzeptes für alle Typen von Daten ist nicht zielführend. Aus diesem Grund werden nachfolgend die Anforderungen von Big Data an den Qualitätsbegriff dargestellt. Diese Anforderungen ergeben sich aus den Merkmalen von Big Data aus Kapitel 3.1.. Die Menge an Daten (Volume) übersteigt

die Möglichkeit zur Nutzung von traditionellen relationalen Datenhaltungen. Datenhaltungssysteme wie NoSQL ermöglichen die Analyse dieser Datenmenge mithilfe unterschiedlicher Software- und Hardware-Technologien. Des Weiteren weist die Datenbasis eine Vielfalt (Variety) von verschiedenen Datenquellen auf, die in einer sehr hohen Geschwindigkeit (Velocity) strukturierte, unstrukturierte und semistrukturierte Daten produziert. Vor allem unstrukturierte und semistrukturierte Daten stellen völlig neue Anforderungen an die Erfassung, Speicherung und Analyse. Der Großteil der Daten in Big Data sind unstrukturiert, sodass eine manuelle Überprüfung nicht möglich ist und dieses automatisiert geschehen muss. Diese Daten weisen oftmals einen schnellen zeitlichen Verfall auf. Positions-, Bewegungs-, Beziehungs- und Verhaltensdaten müssen somit in kurzer Zeit analysiert werden, teilweise in Echtzeit. Aufgrund der Notwendigkeit einer schnellen Analyse muss somit die Datenqualität auch vorab in kurzer Zeit bestimmt werden. Die vermehrte Vernetzung in allen Bereichen stellt auch höhere Anforderungen an den schnellen Zugriff auf Daten zur Steuerung von Prozessen. In der Logistik und der Produktion, aber auch in Onlineshops ist dies von großer Relevanz. Gibt zum Beispiel ein Nutzer in den sozialen Medien den möglichen Kauf eines Artikels preis, kann diese Information einem Onlineshop durch gezielte Werbung zu einem schnellen Kaufabschluss verhelfen. Mit dem vierten Merkmal Veracity wird direkt Bezug auf die Datenqualität genommen. Dies beinhaltet die Glaubwürdigkeit der Daten und der Datenquellen. Daten sind nur dann sinnvoll, wenn sie als zuverlässig gelten. Mit Big Data soll ein Mehrwert (Value) geschaffen werden. Somit bezeichnet das Merkmal, ob die Daten einen Wert besitzen, und falls dies der Fall ist, welchen. Weniger wird der monetäre Wert der Daten damit beschrieben, sondern vor allem der inhaltliche Wert. Dieser kann für jeden unterschiedlich sein und zu Wettbewerbsvorteilen oder neuen Geschäftsmodellen für Unternehmen führen. Veracity und Value werden in der Literatur als beschreibende Merkmale von Big Data bezeichnet, jedoch sind diese nicht unbedingt kennzeichnend. Big Data wird analysiert, um einen Mehrwert zu erschaffen; dafür wird die Glaubwürdigkeit benötigt. Sie sind somit eher als resultierende Größen anzusehen.

## 4.2. Beziehung von Datenqualität und Informationsqualität in Big Data

In diesem Kapitel werden die Begriffe Daten und Information aufgegriffen und der Unterschied im Zusammenhang mit Qualität erörtert. Die Begriffe Datenqualität und Informationsqualität werden in der Literatur häufig synonym verwendet. Die Autoren argumentieren dies unter anderem damit, dass Menschen Daten immer mit irgendeiner Form eines semantischen Kontextes aufnehmen. Des Weiteren wird erklärt, dass Daten einen Verwendungszweck besitzen, der vom Zeitpunkt der Betrachtung abhängig ist, und somit Information und Daten in diesem Zusammenhang sinngleich sind. Es stellt sich die Frage, ob Daten nicht auch ein hohes Niveau haben können, wenn sie keine Informationen liefern oder eine Problemstellung nicht beantworten können. Diese Arbeit folgt der Überzeugung, dass Datenqualität und Informationsqualität mit Blick auf Big Data differenziert betrachtet werden müssen. Um die Unterschiede zu verdeutlichen, werden im Folgenden die im Kapitel **Fehler! Verweisquelle konnte nicht gefunden werden.** behandelten Begriffe Daten, Information und Qualität nochmals aufgegriffen und zusammengefasst. Daten sind eine Ansammlung von Zeichen mit der dazugehörigen Syntax. Sie bilden Tatsachen über Dinge der realen Welt (Entitäten) und über Merkmale dieser Dinge (Attribute) ab. Informationen sind Daten und deren Bedeutung in einem konkreten Kontext. Daten werden zu Information, wenn diese in Beziehung zu anderen Inhalten gesetzt werden. Qualität ist die realisierte Beschaffenheit eines Produkts in Relation zu der geforderten Beschaffenheit. Datenqualität kann dem produktionsorientierten Ansatz von Garvin zugeordnet werden. Die Qualität stellt dort eine objektive Größe dar, die nicht von der subjektiven Wahrnehmung bestimmt wird.
Datenqualität kann als Unterpunkt der Informationsqualität angesehen werden. Zuerst muss die generelle Datenqualität gewährleistet sein. Diese ist jedoch keine Gewährleistung für die Informationsqualität. Nach Garvin kann der anwenderorientierte Ansatz der Informationsqualität zugeschrieben werden. Bei dieser Sichtweise wird die Qualität von dem Nutzer definiert, der subjektiv entscheidet, inwieweit das Ergebnis der geforderten Qualität entspricht („fit for use“). So können Nutzer unterschiedliche Bedürfnisse haben, sodass die Informationsqualität für die gleiche Da-

tenmenge unterschiedlich bewertet wird. Informationsqualität setzt eine Datenqualität voraus und ist ein Maß für die Eignung von Daten, einen angeforderten Verwendungszweck zu erfüllen. Der Nutzen ist für Konsumenten verschieden und kann sich über die Zeit ändern, wenn sich die Bedürfnisse ändern.

## 4.3. Auswahl der Dimensionen

In diesem Kapitel werden die in der Literatur genannten relevanten Dimensionen erläutert. Zudem wird der Bezug zur Daten- und Informationsqualität hergestellt und die Relevanz sowie die Zugehörigkeit diskutiert. Die komplette Liste der Dimensionen verschiedener Autoren mit Definitionen befindet sich im Anhang. Die Vorgehensweise zur Untersuchung der Dimensionen erfolgt nach der Reihenfolge der Relevanz. Die Dimension Vollständigkeit ist von großer Bedeutung bei der Datenqualität. Sie wird auch von vielen Autoren in diesem Zusammenhang genannt und ist in Kapitel 2.2.2 mit einer Metrik versehen. In der Literatur sind verschiedene Auslegungen zu finden. Für Ballou ist ein Datensatz vollständig, wenn alle Werte für eine Variable erfasst wurden (vgl. D. P. Ballou & Pazer, 1985, S. 153). Für Wang und Strong ist wiederum der Kontext zur jeweiligen Aufgabe entscheidend (vgl. Wang & Strong, 1996, S. 32). Fehlen in einem Datensatz Werte, die nicht benötigt werden, kann der Datensatz dennoch vollständig sein. Hier wird der Unterschied von Daten- und Informationsqualität wahrnehmbar, was mit einem Beispiel demonstriert wird. Es sei angenommen, es würde eine Petition für ein generelles Tempolimit auf den Autobahnen vorliegen. Neben den verpflichtenden Angaben zu Namen, Wohnort mit Postleitzahl und Unterschrift gab es freiwillige Angaben wie E-Mail, Telefonnummer, Geburtstag, Beruf, Familienstand, Größe und Gewicht. Die freiwilligen Angaben wurden von den Teilnehmern größtenteils ausgelassen. Diese Liste weist im Wesentlichen in Bezug auf die Datenqualität keine gute Vollständigkeit auf, da viele Werte nicht erfasst wurden. Soll diese Liste jedoch dazu dienen, Unterschiede zwischen der Bevölkerung auf dem Land und der Stadt bei diesem Thema herauszufinden, ist die Informationsqualität hoch, da die Postleitzahl eine Region relativ gut zugeordnet werden kann. Timeliness, die von verschiedenen Autoren genannt wird, kann als Aktualität, Pünktlichkeit oder Rechtzeitigkeit übersetzt werden. Für Pipino gibt diese Dimension an, wie aktuell die Daten in Bezug auf die Aufgabe

sind, für die sie verwendet werden (vgl. Pipino et al., 2002, S. 214). Für Wand und Wang hingegen ist es die Zeit zwischen der Änderung der realen Welt und der Änderung in den Daten (vgl. Wand & Wang, 1996, S. 93). Bovee führt eine detailliertere Beschreibung an, bei der Timeliness sich aus den zwei Komponenten Alter und Volatilität zusammensetzt. Das Alter misst, wie lange es her ist, dass eine Information aufgezeichnet wurde. Je kürzer die Information zurückliegt, desto höher ist somit die Wahrscheinlichkeit, dass sie wahr und relevant ist. Die Volatilität zeigt die Informationsstabilität und meint die Frequenz, wie oft sich der Wert ändert (vgl. Bovee et al., 2003, S. 57-58). Die Temperatur hat somit eine hohe Volatilität, da ein gemessener Wert schnell nicht mehr zeitgemäß ist. Da es bei den Dimensionen Timeliness und Currency verschiedene Definitionen und Überschneidungen gibt, werden diese im Folgenden festgelegt. Currency wird mit Aktualität übersetzt und beschreibt, ob die Daten noch der Realität entsprechen, unabhängig von ihrer Entstehung. Timeliness wird mit Rechtzeitigkeit übersetzt und bestimmt den Zeitraum von der Änderung der realen Welt bis zur Änderung von diesem Zustand in einem Datenbestand. Diese Dimension ist nicht von Bedeutung. Als entscheidendes Kriterium ist jedoch die *Aktualität* anzusehen. Diese stellt die Gegenwartsbezogenheit der Daten dar, d. h., ob die Werte noch die Zustände der Realität beschreiben oder alte Gegebenheiten abbilden. Bei der Aktualität sind verschiedene Zeitbezüge einzubeziehen, zum einen, wann der Wert eingetragen wurde, und der Zeitbezug, den der Wert abbildet, und zum anderen der Zeitpunkt, an dem die Analyse stattfindet. Zum Beispiel kann heute ein Bericht über eine Aktie mitsamt dem Kurs von gestern verfasst werden, der morgen veröffentlicht wird. In der Zwischenzeit kann jedoch ein besonderes Ereignis den Kurs immens verändert haben. Aktien sind ein anschauliches Beispiel für turbulente Veränderungen, bei dem die Aktualität eine entscheidende Rolle spielt. Im Gegensatz dazu stehen zum Beispiel Seekarten, da Daten über Meerestiefen einen ganz anderen zeitlichen Verfall vorweisen. Die Aktualität ist signifikant bei der Datenqualität und somit auch für die Informationsqualität. Bei großen Datensätzen ist anzunehmen, dass auch ältere Daten vorhanden sind und diese die tatsächliche Eigenschaft des zu beschreibenden Objekts nicht wiedergeben. Je nach zeitlichem Verfall des Attributes sollte definiert werden, welche Daten einbezogen werden. Eine andere Möglichkeit wäre eine relative Gewichtung von älteren Daten, um neueren eine höhere Priorisierung zu geben. Bei den Dimensionen Fehlerfreiheit, Genauigkeit und Korrektheit von Pipino, Ballou und

Pazer, Wang und Strong sowie Hinrichs muss eine Differenzierung vorgenommen werden (vgl. D. P. Ballou & Pazer, 1985, S. 153; vgl. Hinrichs, 2002, S. 30; Pipino et al., 2002, S. 212; vgl. Wang & Strong, 1996, S. 31). Laut Ballou und Pazer müssen die gespeicherten Datenwerte den realen Werten entsprechen, was gleichzusetzen ist mit der Auffassung der anderen Autoren. Jedoch schließen die anderen Autoren zusätzlich Zuverlässigkeit der Daten mit ein. Wang und Strong setzen sogar eine Zertifizierung der Daten voraus. Im Kapitel 2.1.2 wurde die Relevanz von Metadaten erläutert; diese werden in dieser Arbeit vorausgesetzt. Frei von Fehlern kann synonym zu Korrektheit benutzt werden. Dies wird theoretisch mit der Überprüfung des Attributwerts im Informationssystem mit dem der modellierten Entität in der Realwelt vorgenommen. Die Korrektheit ist im Zusammenhang der Qualität der Daten wichtig. Aus diesem Grund wird sie als Dimension der Datenqualität geführt und ist somit auch entscheidend für die Informationsqualität. Genauigkeit sollte von diesen Dimensionen klar abgegrenzt werden. Genauigkeit liegt vor, wenn ein Attributwert in dem optimalen Detaillierungsgrad und dem richtigen Format vorliegt. Wenn die Maße eines Bauteils zum Beispiel sehr relevant sind, können fehlende Nachkommastellen als fehlende Genauigkeit klassifiziert werden. Auch diese Dimension hat Auswirkung auf die Datenqualität.

Das Merkmal *Glaubwürdigkeit* von Pipino ist die fünfte und letzte entscheidende Dimension für die Datenqualität (vgl. Pipino et al., 2002, S. 212). Glaubwürdigkeit ist oft kontextabhängig und beruht auf verschiedenen Faktoren. Oftmals beruht die Glaubwürdigkeit auf subjektiver Wahrnehmung. So könnte argumentiert werden, dass Sensor-, Log- sowie Positions- und Bewegungsdaten prinzipiell glaubwürdiger sind als Daten aus sozialen Netzwerken. Jedoch unterliegen Maschinen auch Fehlern und es besteht die Möglichkeit, dass die Daten aus den sozialen Netzwerken von Experten und Fachleuten stammen. Andererseits kann angenommen werden, dass bei der Analyse des Wetters Daten von Messinstrumenten, die Maschinendaten zuzuordnen sind, glaubwürdiger sind als eine Filterung von Beiträgen über das Wetter auf Twitter. Die Glaubwürdigkeit kann als Dimension der Datenqualität betrachtet werden und somit ist diese auch wichtig für die Informationsqualität.

Folgende Dimensionen sind relevant für die Informationsqualität. In welchem Maß die Daten zu dem gewünschten Ergebnis führen, definiert McGilvray durch die Dimension *Durchführbarkeit* (vgl. McGilvray, 2008, S. 111). Da die Datenqualität in dieser Ausarbeitung unabhängig vom Kontext betrachtet wird, ist das Ergebnis nicht

relevant. Jedoch spielt die Durchführbarkeit eine Rolle bei der Informationsqualität. Passen die vorliegenden Daten in ihrer Thematik nicht zur Fragestellung, sind diese für die Analyse unbrauchbar. Die *Relevanz* wird von Wang und Strong als bedeutend bei der Informationsqualität angesehen. Daten sind ohne eine Kenntnis des Kontexts keine Informationen, aus diesem Grunde ist die Relevanz bedeutsam (vgl. Wang & Strong, 1996, S. 31). Für die Datenqualität hingegen spielt es keine Rolle, ob die Daten für eine explizite Aufgabe anwendbar oder hilfreich sind. Die Objektivität spielt für Wang und Strong eine Rolle bei der Informationsqualität (vgl. Wang & Strong, 1996, S. 32). Auch in dieser Arbeit wird der Objektivität eine große Bedeutung beigemessen, da unvoreingenommene, vorurteilsfreie und unparteiische Daten den tatsächlichen Umständen entsprechen. Die Auswahl der Daten spielt hier eine große Rolle. Dazu muss bei Big Data die Auswahl der Datenquellen betrachtet werden, was anhand des omnipräsenten Themas der Coronapandemie veranschaulicht wird. Würden als Datenquelle weitestgehend Inhalte aus diversen Telegramgruppen analysiert werden, wäre anzunehmen, dass die Maßnahmen der Bundesregierung von Menschen überdurchschnittlich skeptisch angesehen werden. Dem gegenüber stehen Quellen von Personen aus den Bereichen der Pflege, Gesundheit und Medizin, von denen oftmals eine konsequentere Vorgehensweise zur Eindämmung der Pandemie verlangt wurde. Um eine tatsächliche objektive Meinungsforschung zu dieser Thematik erarbeiten, müsste der Durchschnitt der Bevölkerung als Datenquelle herangezogen werden. Eine weitere Dimension von Wang und Strong ist die *Interpretierbarkeit* der Daten. Die Daten müssen in einer angemessenen Sprache und Einheit vorliegen sowie verständliche Datendefinition vorweisen (vgl. Wang & Strong, 1996, S. 31). Werden Daten aus den sozialen Medien sowie verschiedene Sprachen und Einheiten genutzt, kann dies zu Problemen führen. Auch die Gewichtsangabe kann zu Unstimmigkeiten führen, da ein Pfund im deutschen Sprachgebrauch 500 g entspricht, in England aber ca 450 g. Somit ist die Interpretierbarkeit von großer Bedeutung bei Big Data. Nach Wang und Strong müssen Daten eine Wertschöpfung erbringen (vgl. Wang & Strong, 1996, S. 31). Sie müssen somit nützlich sein und Vorteile durch ihre Verwendung bieten. Diese Dimension ist sinngleich mit dem Begriff „nützlich“ von Knight und Burn und ist für die Informationsqualität relevant (vgl. Knight & Burn, 2005, S. 162).

Die Konsistenz misst die Widerspruchsfreiheit von Attributwerten in einem Informationssystem (vgl. Hinrichs, 2002, S. 30). Wenn zum Beispiel in einem Tupel beim

Attribut Ansprechpartner eine männliche Person gelistet ist, würde beim Attribut Geschlecht mit dem Wert weiblich die Konsistenz verletzt sein. Für relationale Datenbanken ist diese Dimension grundsätzlich relevant. Die Korrektheit überprüft diese Art von Fehlern. Sind Daten korrekt, ist die Konsistenz gewährleistet. Die *Darstellungsqualität* stützt sich auf die Vorstellung, dass die Präsentation der Informationen bzw. der Ergebnisse, wie zum Beispiel das Aussehen und Format, eine korrekte Nutzung ermöglicht (vgl. McGilvray, 2008, S. 111). Grundsätzlich spielt die Darstellungsqualität bei der Aufnahme von Informationen für Konsumenten eine Rolle, jedoch ist dieses Merkmal bei der Analyse der Qualität nicht relevant. Wenn gewonnene Erkenntnisse aus einem Datenbestand präsentiert werden, spielt die Darstellungsqualität eine Rolle. Jedoch ist diese keine Dimension, um die Daten- oder Informationsqualität zu messen. *Kompaktheit* von Wang und Strong bezieht sich auf die Darstellungsqualität der Daten (vgl. Wang & Strong, 1996, S. 31). Diese ist vor allem bei der Darstellung der Ergebnisse für den Konsumenten bedeutsam. Oft werden zu viele Informationen zu einer Belastung eines Konsumenten, sodass diese nicht verarbeitet werden können. Diese Dimension ist nicht der Datenqualität zuzuordnen, sondern der Informationsqualität. Gleichwohl ist das keine ausschlaggebende Dimension, die einen maßgeblichen Einfluss auf die Informationsqualität hat. *Konsistente Repräsentation* kann ebenfalls nicht der Datenqualität zugeordnet werden (vgl. Pipino et al., 2002, S. 212). Es kann nicht davon ausgegangen werden, dass Daten im gleichen Format vorliegen. Für die Darstellung der Daten kann eine konsistente Repräsentation vorausgesetzt werden. Dies fällt aber unter die Darstellungsqualität und gehört somit auch nicht zur Informationsqualität. McGilrays Dimension Konsistenz und Synchronisierung entspricht nahezu der konsistenten Repräsentation von Pipino oder Wang und Strong und entfällt somit (vgl. McGilvray, 2008, S. 111). *Benutzer- und Wartungsfreundlichkeit* setzt sich aus mehreren Dimensionen zusammen (vgl. McGilvray, 2008, S. 111). Diese vereinfachen jedoch lediglich die Nutzung und Bearbeitung für den Konsumenten und somit beeinflusst weder die Benutzer- noch die Wartungsfreundlichkeit die Qualität von Daten oder Informationen.

Eine *angemessene Menge an Daten* und *Umfang der Daten* erübrigt sich bei Big-Data-Anwendungen (vgl. Pipino et al., 2002, S. 212; vgl. Wang & Strong, 1996, S. 32). Im Allgemeinen gilt: Je größer die Datenbasis, auf denen eine Analyse be-

ruht, desto genauer sind die resultierenden Ergebnisse. Liegen jedoch die vorhandenen Daten nicht in einer ausreichenden Qualität vor, ist das nicht der Fall. Da Big Data große Datenmengen impliziert, kann diese in der Literatur genannte Dimension als irrelevant betrachtet werden. Eine Differenzierung, unabhängig von Big Data, ist der Bezug zu der erforderlichen Menge. Im Kontext einer konkreten Fragestellung ist die erforderliche Datenmenge subjektiv zu betrachten und für die Bewertung der Informationsqualität entscheidend festzulegen. Ein *Datenverfall* setzt mit der Zeit ein (vgl. McGilvray, 2008, S. 111). In der Regel verlieren digitale Daten mit der Zeit ihre Gültigkeit. Dies kann eine entscheidende Rolle bei der Datenqualität in Big Data spielen, da bei großen Datenmengen oftmals auch Daten verwendet werden, die nicht mehr zeitgemäß sind. Diese Dimension steht im Einklang mit der Aktualität von Daten, da die Aktualität den Datenverfall miteinbezieht. Definieren Geschäftsregeln zeitgerechte Daten, kann der Datenverfall außer Acht gelassen werden. McGilvray nennt *Duplikate* als Messgröße zur Datenqualität. Weisen Entitäten unterschiedliche Darstellungen in oder zwischen Datenbanken auf, werden diese Duplikate genannt (vgl. McGilvray, 2008, S. 111). Im Idealfall sind in einem System für jeden Datensatz einer Datenbank eindeutige Bezeichnungen vorhanden. Jedoch ist dies in den meisten praktischen Fällen, vor allem bei Big Data, nicht der Fall, und den Daten fehlt eine eindeutige Identifikation. Dies kommt zum Beispiel vor, wenn Daten aus mehreren Quellen über verschiedene Abteilungen oder Organisationen hinweg integriert werden. Diesbezüglich können Differenzen in Datensatzformaten, Standardisierungen oder Tippfehler unweigerlich zu Duplikaten führen. Bei der Datenanalyse von Big Data muss vor allem bei semi- und unstrukturierten Daten darauf geachtet werden. Duplikate könnten bei Textdateien zum Beispiel Synonyme sein, die den gleichen Sachverhalt beschreiben. Bei einer Analyse der Daten ist sorgfältig auf diesen Umstand zu achten. Die Identifizierung doppelter Elemente in den Daten ist ein signifikantes Mittel, die Zuverlässigkeit und Glaubwürdigkeit der Daten zu garantieren. Bei Big Data sind Duplikate jedoch unumgänglich und somit kein Faktor für die Datenqualität. Für Pipino ist eine problemlose Manipulation der Daten ein relevantes Kriterium. Daten sollen auf verschiedene Aufgaben übertragbar sein (vgl. Pipino et al., 2002, S. 212). Dies ist weder für die Datenqualität noch für die Informationsqualität entscheidend. Datenqualität bezieht sich auf den Wert der Daten und die Informationsqualität betrachtet, in welchem Maß die

Daten der Aufgabe dienen. Ergebnisse müssen nicht zwangsläufig auf verschiedene Aufgaben übertragbar sein, wenn sie dem Konsumenten in Bezug auf die Problemstellung helfen. Lernvermögen kann ebenfalls nicht als passende Dimension angesehen werden. Bei der Datenqualität spielt es keine Rolle und bei Informationen ist es selbstredend, dass sich Kenntnisse angeeignet werden können. Peralta hat in ihrer Dissertation die Dimension Freshness (Neuheit) behandelt, die für die Qualität eine Rolle spielt. Zum einen besteht die Neuheit aus der Aktualität der Daten, zum anderen aus der Pünktlichkeit (vgl. Peralta, 2006, S. 7-8). Unter Pünktlichkeit wird der Zeitpunkt verstanden, wann der Datensatz zur Verfügung steht. Die Pünktlichkeit kann bei der Big-Data-Analyse vernachlässigt werden, da nur vorhandene Daten betrachtet werden. Der Aspekt der Aktualität wurde im Vorfeld schon als relevant nachgewiesen. Nach Holthuis spielt für Datenqualität die Sicherheit eine Rolle (vgl. Holthuis, 1999, S. 34). Der Schutz vor unautorisierten Zugriffen gewährleistet dem Unternehmen Wettbewerbsvorteile. Grundsätzlich bieten Daten oder Informationen Vorteile für Unternehmen. Bieten diese in bestimmte Bereichen Nutzen gegenüber Konkurrenten, sollten die Daten entsprechend geschützt werden. Die Logik, dass diese wertvoller sind, wenn kein anderer sie benutzen kann, ist ersichtlich. Jedoch geht der Vorteil gegenüber anderen zu weit, um Sicherheit als Dimension von Daten- oder Informationsqualität zu betrachten. Transportierbarkeit bezieht sich auf die Übertragung von Daten auf verschiedene Systeme (vgl. Holthuis, 1999, S. 34). Orts- oder systembezogene Daten sind für Nutzer weniger vorteilhaft. Dies spielt für die Datenqualität und die Informationsqualität bei Big Data keine Rolle, da davon ausgegangen wird, dass die Daten zur Analyse bereitstehen. Zudem müssen die Daten nicht in andere Systeme übertragen werden, um qualitativ zu sein. Die Transportierbarkeit ist bedeutsam für das Datenqualitätsmanagement in Unternehmen, jedoch nicht Gegenstand in dieser Arbeit. Knight und Burn nennen die Verfügbarkeit als Merkmal. Informationen müssen physisch zugänglich sein, um eine Qualität aufzuweisen (vgl. Knight & Burn, 2005, S. 162). Die Verfügbarkeit ist die Grundlage für eine Analyse und kann daher nicht als Dimension betrachtet werden. Die Dimension *Datenspezifikation* von McGilvray beinhaltet verschiedene Aspekte der Datenqualität (vgl. McGilvray, 2008, S. 111). Die Existenz, Vollständigkeit, Qualität und Dokumentation von Datenstandards, Datenmodellen, Geschäftsregeln, Metadaten und Referenzdaten können nicht als Dimen-

sion von Datenqualität bei Big Data betrachtet werden. Geschäftsregeln für die Datenqualität müssen im Vorfeld definiert werden und bestimmen die Anforderungen von Ausprägungen einer Dimension. Zum Beispiel könnte eine Geschäftsregel lauten, dass lediglich Daten verwendet werden, die nicht älter sind als 20 Tage. Metadaten sind bei der Analyse vorausgesetzt, da diese zum Prüfen einiger Dimensionen elementar sind. Die *Gültigkeit* von Daten oder Informationen liegt vor, wenn sie als zweifelsfrei überprüft werden können und angemessene Standards in Bezug auf andere Dimensionen wie Genauigkeit, Aktualität oder Vollständigkeit erfüllen (vgl. Miller, 1996, S. 81). Zum Beispiel besitzt das Datum 30.02.2021 keine Gültigkeit, da dieser Tag nicht existiert. Diese Dimension ist eher eine resultierende als eine kausale Dimension, weil unter anderem die Korrektheit und Genauigkeit die Gültigkeit implizieren. *Benutzer- und Wartungsfreundlichkeit* setzen sich aus mehreren Dimensionen zusammen (vgl. McGilvray, 2008, S. 111). Diese vereinfachen jedoch lediglich die Nutzung und Bearbeitung für den Konsumenten und somit beeinflusst weder die Benutzer- noch die Wartungsfreundlichkeit die Qualität von Daten oder Informationen. Die *Darstellungsqualität* stützt sich auf die Vorstellung, dass die Präsentation der Informationen bzw. der Ergebnisse, wie zum Beispiel das Aussehen und das Format, eine korrekte Nutzung ermöglicht (vgl. McGilvray, 2008, S. 111). Grundsätzlich spielt die Darstellungsqualität bei der Aufnahme von Informationen für Konsumenten eine Rolle, jedoch besitzt dieses Merkmal bei der Analyse der Qualität keine Bedeutung. Werden gewonnene Erkenntnisse aus einem Datenbestand präsentiert, spielt die Darstellungsqualität eine Rolle. Jedoch ist diese keine Dimension, um die Daten- oder Informationsqualität zu messen. Knight und Burn haben zwei Dimensionen genannt, die auf den ersten Blick keine Unterschiede aufzuweisen scheinen. Diese sind Useability (Nutzbarkeit) und Useful (nützlich). Die Nutzbarkeit zielt allgemein auf das Ergebnis ab, das klar und leicht zu nutzen sein sollte. Nützlich sind die Informationen, wenn diese auf die jeweilige Aufgabe umzusetzen sind und somit einen Mehrwert erbringen (vgl. Knight & Burn, 2005, S. 162). Dies ist von großer Bedeutung als Dimension für die Informationsqualität, da ohne eine Anwendung der Erkenntnisse kein Nutzen vorhanden sein kann. Bei der Frage, woher die Daten kommen und ob die Datenquellen hoch angesehen werden, wird die Dimension *Reputation* herangezogen (vgl. Wang & Strong, 1996, S. 32). Diese ist synonym zu betrachten mit der Dimension Glaubwürdigkeit. Verlässlichkeit steht direkt in Verbindung mit Glaubwürdigkeit, Reputation und Genauigkeit. Bei Big Data

sind diese Punkte nicht immer leicht umzusetzen. Nicht immer sind die Informationen gegeben, woher die Daten stammen und ob diese der Wahrheit entsprechen. Wenn Daten klar, eindeutig und leicht verständlich sind, fasst Pipino dieses als Verständlichkeit zusammen. Diese Dimension beinhaltet Aspekte von Genauigkeit, die klare und eindeutige Daten voraussetzt (vgl. Pipino et al., 2002, S. 212). Leichte Verständlichkeit ergibt sich aus klaren und eindeutigen Daten, sodass diese Dimension nicht weiter betrachtet werden muss. Wirksamkeit setzt sich aus mehreren Dimensionen zusammen beziehungsweise ist das Ergebnis. Für Batini ergibt sich die Wirksamkeit, wenn die Genauigkeit und die Vollständigkeit für eine bestimmte Aufgabe erreicht sind (vgl. Batini et al., 2009, S. 32). Der Wirkungsgrad ist in diesem Zusammenhang, inwieweit die Daten den Informationsbedarf decken und wie schnell dies geschieht.

## 4.4. Definitionen

Nach der Trennung von Daten und Information und somit auch der Trennung von Datenqualität von der Semantik und der Bedeutung für den Konsumenten kann diese im Kontext von Big Data wie folgt definiert werden. *Datenqualität* ist ein mehrdimensionales Maß zur Quantifizierung der Abbildungsgüte zwischen der realen Welt und der Repräsentation im Informationssystem und setzt sich aus den Dimensionen Vollständigkeit, Korrektheit, Glaubwürdigkeit, Genauigkeit und Aktualität zusammen. Nachfolgend wird die Informationsqualität neu definiert. *Informationsqualität* setzt die Datenqualität voraus und ist ein Maß für die Eignung der Daten, einen bestimmten Zweck zu erfüllen. Neben den Dimensionen der Datenqualität sind Interpretierbarkeit, Relevanz, Objektivität, Durchführbarkeit und Wertschöpfung von Bedeutung. Die Beziehung der Daten- und der Informationsqualität ist in Abbildung 11 nochmals übersichtlich dargestellt.

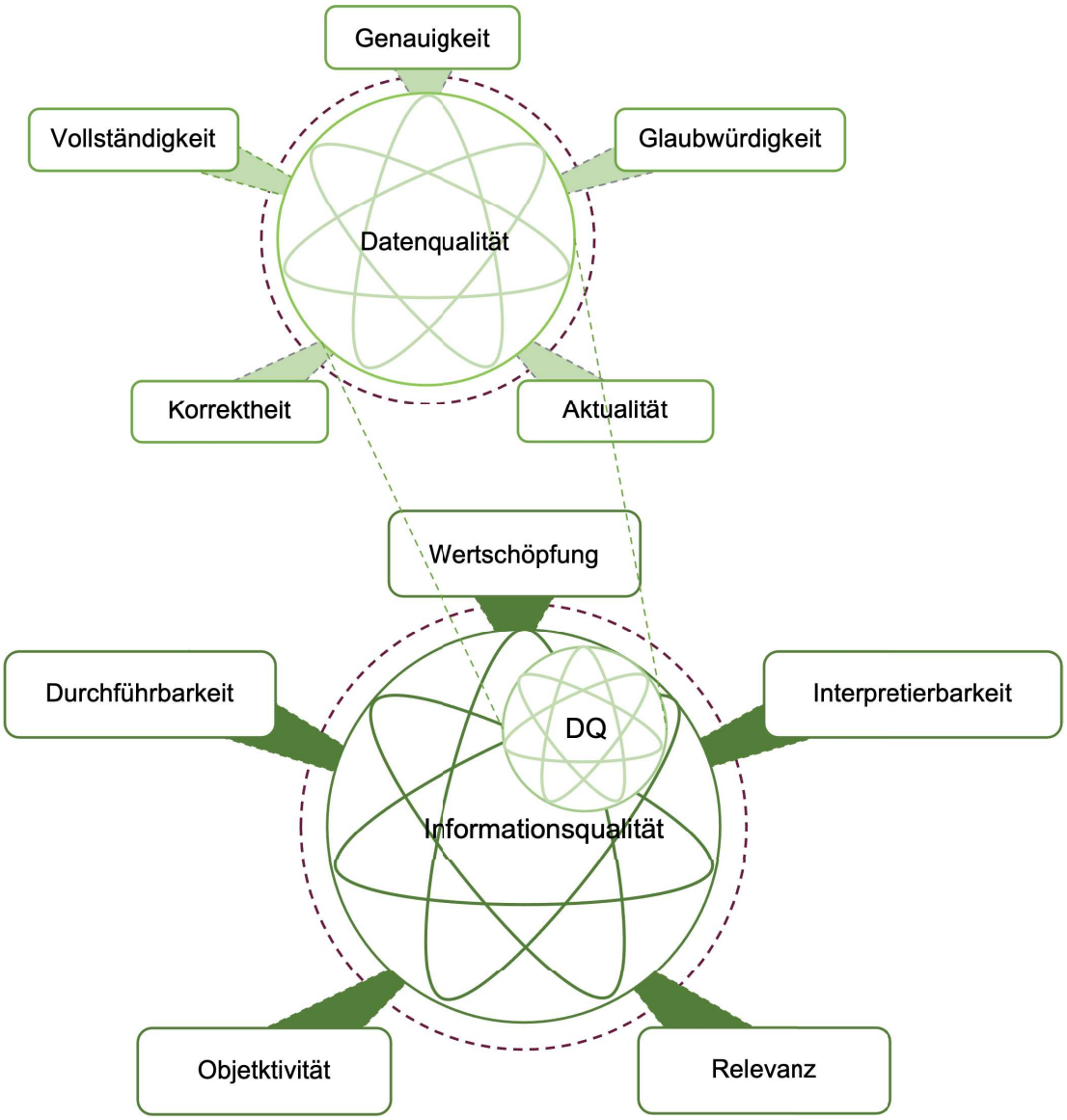


Abbildung 11 Beziehung von Datenqualität und Informationsqualität, eigene Darstellung

# 5. Bewertungsmethode

Im ersten Abschnitt der praktischen Leistung ist die Zusammenführung der Literatur von Datenqualität und Big Data erfolgt. Dimensionen wurden anhand der erarbeiteten Anforderungen von Big Data ausgewählt und irrelevante Dimensionen ausgeschlossen. Nach der Formulierung von neuen Definitionen zu Daten- und Informationsqualität, wird in diesem Kapitel die Anwendung der Metriken behandelt. Dies erfolgt in der gleichen Reihenfolge, wie sie in Kapitel 2.2.2. vorgestellt wurde. Im Anschluss wird im Kapitel 5.2. überprüft, ob die Metriken den geforderten Anforderungen entsprechen. In diesem Kapitel werden zudem die Ergebnisse diskutiert und interpretiert.

## 5.1. Anwendung der Metriken

Die Metriken der Vollständigkeit, Korrektheit, Genauigkeit und Aktualität von Hinrichs wurden bereits im Kapitel 2.2.2. vorgestellt und werden nun nacheinander an einer Datenbasis getestet. Dies erfolgt mithilfe einer relationalen Datenbank. Um die Metriken der Dimensionen zu testen, wird ein eigener Datensatz geschaffen. Das hat zur Folge, dass die Dimensionen besser auf ihre Praxistauglichkeit überprüft werden können und eventuelle Probleme durch beabsichtigte Fehler hervortreten. Die Tabellen 4 und 5 werden auf ihre Qualität geprüft und gehören zu der Datenbank Lager, die Tabelle 6 dient als Hilfestellung zu Überprüfung der Korrektheit.

Tabelle 4 Produkte

| Art.-Nr. | Hersteller | Bezeichnung | … | Wareneingang | Istmenge | Preis in € |
|---|---|---|---|---|---|---|
| 1 | Continental | Reifen | … | 21.05.2020 | 5 | 58,23 |
| 2 | Michelin | Reifen | … | 21/06/2020 | 4 | 65 |
| 3 | Michelin | | … | 23/7/2020 | 10 | 70.00 |
| 4 | Dunlop | Reifen | … | 23.7.2020 | 8 | 57,50 |
| 5 | Dunlop | Reifen | … | 30.02.2020 | 9 | |
| 6 | Dunlop | | … | 12.13.2020 | 10 | 80 |
| 7 | Continental | Reifen | … | 21/05/2020 | 8 | 85.00 |
| 8 | | Reifen | … | 21.05.2020 | 20 | 90,25 |
| … | | | … | | | |

Tabelle 5 Hersteller

| Hersteller | Kontakt | Telefon | Straße | Ort | Plz |
|---|---|---|---|---|---|
| Continental | Müller | 0151 7261923 | Vahrenwalder St. 9 | Hannover | 30165 |
| Michelin | Meier | 0151 7261923 | | Karlsruhe | 76185 |
| Dunlop | | 06181 6801 | Dunlopstr. | Hanau | |
| … | … | | … | … | … |

Tabelle 6 Inventurliste

| Art.-Nr. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Istmenge | 5 | 4 | 10 | 8 | 9 | 10 | 8 | 20 |
| Sollmenge | 5 | 5 | 10 | 8 | 9 | 11 | 8 | 28 |
| Preis in € | 58,23 | 65,00 | 72,00 | 60,00 | 80,00 | 80,00 | 85,00 | 90,25 |

### 5.1.1. Vollständigkeit

Bei der Vollständigkeit wird lediglich betrachtet, ob ein Attributwert vorhanden ist. Jedoch muss miteinbezogen werden, ob dieser Attributwert tatsächlich vorhanden sein muss. Im vorliegenden Beispiel aus der Entität Hersteller kann es sein, dass es keinen direkten Ansprechpartner gibt. Falls dies der Fall ist, würde kein Qualitätsdefizit vorliegen. Um dieses zu berücksichtigen, kann entweder das Attribut ausgelassen oder die relative Wichtigkeit $g_j$ miteinbezogen werden.

Die Attributwertebene wird mit der Funktion $NotNull(\omega) := \begin{cases} 1 & falls\ \omega = NULL\ oder\ \omega\ zu\ NULL\ äquivalent \\ 0 & sonst. \end{cases}$ beschrieben.

Bei der Entität Produkte sind lediglich vier Attributwerte nicht vollständig.

Auf der *Tupelebene* werden nun die einzelnen Attribute eines Tupels betrachtet. Hier wird auch die relative Wichtigkeit $g_j$ miteinbezogen, die der Tabelle 7 zu entnehmen ist.

Tabelle 7 Tupelebene Vollständigkeit Produkte

| | **Art.-Nr.** | **Her-steller** | **Bezeich-nung** | **…** | **Waren-eingang** | **Ist-menge** | **Preis in €** |
|---|---|---|---|---|---|---|---|
| | 8 | | Reifen | … | 21.05.2020 | 20 | 90,25 |
| $\boldsymbol{Q_{Voll}(t)}$ | 1 | 0 | 1 | … | 1 | 1 | 1 |
| $\boldsymbol{g_j}$ | 1 | 5 | 3 | … | 0,5 | 2 | 1 |

Die relative Wichtigkeit für die Attribute Hersteller, Bezeichnung und Istmenge sind sehr relevant. Weniger wichtig ist der Wareneingang bei diesem Beispiel. Mit der Formel $Q_{Voll}(t) \coloneqq \frac{\sum_{j=1}^{n} Q_{Voll}(t.A_j) g_j}{\sum_{j=1}^{n} g_j}$ ergibt sich $Q_{Voll}(8) \coloneqq \frac{1*1+0*5+1*3+1*0{,}5+1*2+1*1}{12{,}5}$ = 0,6. Die Vollständigkeit für dieses Tupel beträgt somit 60 %.

Auf der *Relationenebene* werden nun alle Tupel der Entität Produkte herangezogen. Die Rechnung erfolgt analog zum Tupel 8 und wird nicht weiter ausgeführt. Die Ergebnisse sind in Tabelle 8 abgebildet.

Tabelle 8 Ergebnis Relationenebene Vollständigkeit Produkte

| **Tupel** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $\boldsymbol{Q_{Voll}(t)}$ | 1 | 1 | 0,76 | 1 | 0,92 | 0,76 | 1 | 0,6 |

Mit $Q_{Voll}(T) \coloneqq \frac{\sum_{i=1}^{|T|} Q_{Voll}(t_i)}{|T|}$ ergibt sich $Q_{Voll}(Produkte) \coloneqq \frac{1+1+0{,}76+1+0{,}92+0{,}76+1+0{,}6}{8} = 0{,}88$. Die Vollständigkeit auf Relationenebene der Entität Produkte beträgt somit 88 %.

Im letzten Schritt wird die Datenbankebene betrachtet. In diesem Beispiel werden nur die Entitäten Produkte und Hersteller betrachtet. Dazu wird zuerst die Vollständigkeit auf Attribut-, Tupel- und Relationenebene der Entität Hersteller vorgenommen. Wie bereits erwähnt gibt es bei der Entität Hersteller und dem Attribut Kontakt Diskussionen über die Machbarkeit und Sinnhaftigkeit einer Qualitätsaussage. Ist

ein Attributwert leer, da das Attribut nicht existiert, bietet sich eine geringe relative Wichtigkeit $g_j$ an. Diese Methode wird in dem Beispiel auch vorgenommen. Die Berechnungen werden nicht weiter ausgeführt; die Ergebnisse sind Tabelle 9 dargestellt.

Tabelle 9 Relationenebene Vollständigkeit Hersteller

| | | $Q_{Voll}(t)$ | | |
|---|---|---|---|---|
| | $g_j$ | **Tupel 1** | **Tupel 2** | **Tupel 3** |
| **Hersteller** | 1 | 1 | 1 | 1 |
| **Kontakt** | 0,5 | 1 | 1 | 0 |
| **Telefon** | 1 | 1 | 1 | 1 |
| **Straße** | 1 | 1 | 0 | 1 |
| **Ort** | 0,5 | 1 | 1 | 1 |
| **PLZ** | 1 | 1 | 1 | 0 |
| $Q_{Voll}(t)$ | | 1 | 0,8 | 0,7 |
| $Q_{Voll}(T)$ | | | 0,83 | |

Die Qualität der Vollständigkeit auf Datenbankebene wird durch die Formel $Q_{Voll}(D) := \frac{\sum_{k=1}^{p} Q_{Voll}(T_k)}{p}$ berechnet. Somit liegt die Qualität $Q_{Voll}(Lager) := \frac{0{,}83+0{,}88}{2} = 0{,}855$ bei 85,5 %.

### 5.1.2. Korrektheit

Die Dimension Korrektheit wird zuerst anhand der Entität Produkte überprüft. Für die Attributwertebene wird die Formel $d(\omega_1, \omega_2) := \begin{cases} 0 & falls\ \omega_1 = \omega_2 \\ \infty & sonst. \end{cases}$ hinzugezogen. Im ersten Tupel ist $\omega_1 = Reifen$ und $\omega_2 = Reifen$ und somit ergibt sich $Q_{Korr}(\omega_I, \omega_R) := \frac{1}{0+1}$ und damit eine Korrektheit von 1 beziehungsweise 100 %.
Im Falle der Überprüfung der Istmenge und der Sollmenge aus Tabelle 6 kann einerseits die Formel 7 verwendet werden. Bei $\omega_1 = 4$ und $\omega_2 = 5$ würde bei der Formel eine Korrektheit von 0 % vorliegen. Zum anderen kann auch die Formel

$d(\omega_1, \omega_2) := |\omega_1 - \omega_2|$ hinzugezogen werden. Nach der Formel 9 liegt die Korrektheit $Q_{Korr}(4{,}5) = \frac{1}{1+1}$ bei 50 %. Bei dieser Berechnung des Abstandsmaßes sollte jedoch eventuell eine andere Formel hinzugezogen werden. Angenommen, die Istmenge liegt bei 2000 und die Sollmenge bei 2001. Nach der Formel 9 liegt die Korrektheit auch bei 50 %, wobei es sich bei einer solchen Anzahl prozentual um eine geringere Abweichung handelt und somit die Korrektheit höher ausfallen sollte. Eine Möglichkeit wäre die folgende Formel:

$$Q_{Korr}(\omega_I, \omega_R) := \frac{\omega_I}{|\omega_I - \omega_R| + \omega_R} \tag{21}$$

Bei $\omega_I = 4$ und $\omega_R = 5$ ergibt sich $Q_{Korr}(4{,}5) = \frac{4}{1+5} = \frac{2}{3} = 0{,}667$ und für $\omega_I = 2000$ und $\omega_R = 2001$ ergibt sich $Q_{Korr}(2000{,}2001) = \frac{2000}{1+2001} = 0{,}999$. Welche Formel tatsächlich im Anwendungsfall eingesetzt werden sollte, obliegt den Analysten, da die Attribute verschiedene Ziele verfolgen und deren Korrektheit somit abhängig sind. Diese Berechnung von Hinrichs kann auch bei Wörtern eingesetzt werden. Mithilfe der Hamming-Distanz können bei der Bewertung die einzelnen Positionen des Wortes verglichen und unterschieden werden. Zur Verdeutlichung wird das Produkt mit der Artikelnummer 5 hinzugezogen.

$Q_{Korr}(Dunlup, Dunlop) := \frac{1}{d(Dunlop,Dunlop)+1} = \frac{1}{|\{5\}|+1} = \frac{1}{1+1}$ = 50 %

Diese Methode ist jedoch unter Umständen unvorteilhaft. Bei den Namen Yanick und Jannic liegt die Korrektheit bei 20 %. Bei den Begriffen Quiz und Zaun liegt die Korrektheit ebenfalls bei 20 %, wobei diese sinngemäß komplett unterschiedlich sind. Im letzten Beispiel ist kein Buchstabe gleich und somit sollte keine Korrektheit vorliegen. Aus diesem Grund muss ein anderes Messverfahren benutzt werden. Vielversprechender wäre die Formel:

$$Q_{Korr}(\omega_I, \omega_R) := 1 - \left(\frac{d(\omega_I, \omega_R)}{\max(\omega_I, \omega_R)}\right) \tag{22}$$

Somit würde eine Korrektheit bei den Namen Yanick und Jannic $Q_{Korr}(Yanick, Jannic) = 1 - \left(\frac{4}{6}\right) = \frac{1}{3}$ von 33,33 % vorliegen und bei Quiz und Zaun

von 0. In der weiteren Ausführung wird die neu entwickelte Formel bei Wörtern verwendet.

Auf *Tupelebene* werden nun die einzelnen Attribute eines Tupels entsprechend einer Entität von der Datenbasis mit der Diskurswelt verglichen. Um dieses zu verdeutlichen, wird die erste Zeile des Beispiels herangezogen. Die Formel $Q_{Korr}(t,e) \coloneqq \frac{\sum_{j=1}^{n} Q_{Korr}(t.A_j,e.A_j)gj}{\sum_{j=1}^{n} g_j}$ zeigt die Korrektheit eines Tupels basierend auf der Korrektheit jedes einzelnen Attributs. Zudem können die Attribute mit einer Gewichtung versehen werden, sodass zum Beispiel die Korrektheit für den Vergleich der Ist- und der Sollmenge höher priorisiert wird als das korrekte Datum des Wareneingangs.

Tabelle 10 Tupelebene Korrektheit Produkte

| | **Art.-Nr.** | **Hersteller** | **Bezeichnung** | **…** | **Wareneingang** | **Istmenge** | **Preis in €** |
|---|---|---|---|---|---|---|---|
| | 5 | Dunlop | Reifen | … | 30.02.2020 | 9 | |
| $\boldsymbol{Q_{Voll}(t)}$ | 1 | 0,833 | 1 | … | 1 | 1 | 0 |
| $\boldsymbol{g_j}$ | 1 | 2 | 2 | … | 1 | 3 | 1 |

Für die Formel $Q_{Korr}(t,e) \coloneqq \frac{\sum_{j=1}^{n} Q_{Korr}(t.A_j,e.A_j)gj}{\sum_{j=1}^{n} g_j}$ liegt dann eine Korrektheit von $Q_{Korr}(5,Produkte) \coloneqq \frac{1*1+0{,}833*2+1*2+1*1+1*3+0*1}{10}$ = 0,866 = 86,66 % vor.

Bei der Relationenebene werden nun alle Tupel einer Entität bewertet. Dazu werden die einzelnen Tupel wie in Tabelle 10 berechnet. In Tabelle 11 sind die Werte für die einzelnen Attribute dargestellt und in Tabelle 12 die Ergebnisse der Relationenebene für die Korrekheit der Produkte. Bei der Berechnung der Ist- und der Sollmenge wird die Formel $Q_{Korr}(\omega_I,\omega_R) \coloneqq \frac{\omega_I}{|\omega_I - \omega_R| + \omega_R}$ hinzugezogen, wobei $\omega_I$ = Sollmenge und $\omega_R$ = Istmenge.

Tabelle 11 Relationenebene Korrektheit Produkte

| Art.-Nr. | Hersteller | Bezeichnung | … | Wareneingang | -Menge | Preis in € |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | … | 1 | 1 | 1 |
| 2 | 1 | 1 | … | 1 | 0,8 | 1 |
| 3 | 1 | 0 | … | 1 | 1 | 0 |
| 4 | 1 | 1 | … | 1 | 1 | 0 |
| 5 | 0,833 | 1 | … | 1 | 1 | 0 |
| 6 | 1 | 0 | … | 1 | 0,917 | 1 |
| 7 | 1 | 1 | … | 1 | 1 | 0 |
| 8 | 0 | 1 | … | 1 | 0,778 | 1 |
| … | | | … | | | |

Tabelle 12 Ergebnis Relationenebene Korrektheit Produkte

| **Tupel** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $\boldsymbol{Q_{Korr}(t,e)}$ | 1 | 0,94 | 0,8 | 1 | 0,866 | 0,975 | 0,9 | 0,733 |

Mit der Formel $Q_{Korr}(T,E) \coloneqq \frac{\sum_{i=1}^{|T|} Q_{Korr}(t_i,e_i)}{|T|}$ ergibt sich dann $Q_{Korr}(T,Produkte) \coloneqq \frac{1+0,94+0,8+1+0,866+0,975+0,9+0,733}{8}$ = 0,902. Auf Relationenebene liegt die Korrektheit dann bei 90,2 %.

Letztlich wird wieder die Datenbankebene betrachtet. In diesem Beispiel werden nur die Entitäten Produkte und Hersteller untersucht. Da die Bestimmung der Korrektheit auf Attribut-, Tupel- und Relationenebene der Entität Produkte bereits vorgenommen wurde, fehlt nur noch die Entität Hersteller. Die Berechnungen werden nicht weiter ausgeführt, es gibt keine unterschiedliche Gewichtung; die Ergebnisse sind in Tabelle 13 dargestellt.

Tabelle 13 Relationenebene Korrektheit Hersteller

| | $Q_{Korr}(\omega_I, \omega_R)$ | | |
|---|---|---|---|
| | **Tupel 1** | **Tupel 2** | **Tupel 3** |
| **Hersteller** | 1 | 1 | 1 |
| **Kontakt** | 1 | 1 | 1 |
| **Telefon** | 1 | 0 | 1 |
| **Straße** | 1 | 0 | 1 |
| **Ort** | 1 | 1 | 1 |
| **PLZ** | 1 | 1 | 0 |
| $Q_{Korr}(t, e)$ | 1 | 0,67 | 0,83 |
| $Q_{Korr}(T, E)$ | | 0,83 | |

Mit der Formel $Q_{Korr}(D, R) := \frac{\sum_{k=1}^{p} Q_{Korr}(T_k, E_k)}{p}$ wird die Datenbankebene überprüft. Mit $Q_{Korr}(T, E) := 0{,}83$ und $Q_{Korr}(T, E) := 0{,}894$ wird dann $Q_{Korr}(Lager, Inventurliste) := \frac{0{,}83+0{,}902}{2}$ = 0,866. Die Korrektheit auf Datenbankebene beträgt dann 86,6 %.

### 5.1.3. Genauigkeit

Die Stelligkeit spielt sowohl bei numerischen Attributen als auch bei symbolischen Werten eine Rolle. Die Genauigkeit wird bewertet über das Verhältnis der Stelligkeit des Attributs zur jeweils idealen Stelligkeit und ist vorher zu definieren. Die Attribute Preis in € und Wareneingang aus der Entität Produkte werden nun auf ihre Genauigkeit auf Attributwerteebene überprüft. Dazu wird zuerst die Funktion $s : Dom(A) \rightarrow \mathbb{N}$ für das Attribut Preis in € betrachtet. Die optimale Stelligkeit ist für das Beispiel ein Wert mit zwei Nachkommastellen. Darüber hinaus sollen mit einem Komma die Dezimalstellen abtrennt werden, was die Funktion ebenso kontrolliert. Für die erste Zeile ergibt sich dann $Q_{Gen}(58{,}23, Preis\ in\ €) := min\left(\frac{s(58{,}23)}{s_{opt}(Preis\ in\ €)}, 1\right) = min\left(\frac{3}{3}, 1\right) = 1.$ Die Genauigkeit beträgt somit 100 %. In der zweiten Zeile wird jedoch schnell ein Problem deutlich. Für $Q_{Gen}(65, Preis\ in\ €)$

$:= min\left(\frac{s(65)}{s_{opt}(Preis\ in\ €)}, 1\right) = min\left(\frac{0}{3}, 1\right) = 0$ ergibt sich eine Genauigkeit von 0 %. Wenn der Preis 65,00 € beträgt, kann der Wert aus diesem Grunde korrekt sein, aber eine Genauigkeit von 0 % vorweisen. Das Fehlen von Nachkommastellen kann bei der Berechnung korrekte Werte als ungenau angeben und irrtümlich ein großes Qualitätsdefizit bescheinigen. Nun erfolgt die Analyse zur Genauigkeit des Attributs Wareneingang; die Werte werden mit Kalendertagen angegeben. Die Funktion $s: Dom(A) \rightarrow \mathbb{N}$ untersucht, ob das Datum in der numerischen Schreibweise TT.MM.JJJJ vorliegt. In der zweiten Zeile beim Wert 21/06/2020 sind zwei Bedingungen verletzt. Somit ergibt sich $Q_{Gen}(21/06/2020, Wareneingang)$ $:= min\left(\frac{s(21/06/2020)}{s_{opt}(Wareneingang)}, 1\right) = min\left(\frac{8}{10}, 1\right) = 0{,}8, \mathrm{d.\,h.}$ eine Genauigkeit von 80 %.

Auf Tupelebene werden nun alle Attribute eines Tupels betrachtet. Bei den Attributen, die nicht überprüft werden müssen, wird eine Genauigkeit mit 1 bei Vollständigkeit und 0 bei Unvollständigkeit angegeben. Darüber hinaus wird wieder mit der relativen Wichtigkeit $g_j$ gerechnet.

Tabelle 14 Tupelebene Genauigkeit Produkte

| | **Art.-Nr.** | **Hersteller** | **Bezeichnung** | **…** | **Wareneingang** | **Istmenge** | **Preis in €** |
|---|---|---|---|---|---|---|---|
| | 2 | Michelin | Reifen | … | 21/06/2020 | 4 | 65 |
| $Q_{Gen}(t)$ | 1 | 1 | 1 | … | 0,8 | 1 | 0 |
| $g_j$ | 1 | 1 | 1 | … | 2 | 1 | 2 |

Mit der Formel $Q_{Gen}(t) := \frac{\sum_{j=1}^{n} Q_{Voll}(t.A_j, A_j) g_j}{\sum_{j=1}^{n} g_j}$ ergibt sich dann $Q_{Voll}(2) := \frac{1*1+1*1+1*1+0{,}8*2*1*1+0*2}{8}$ = 0,7. Die Genauigkeit für dieses Tupel beträgt somit 70 %.

Nun wird die Relationenebene der Genauigkeit betrachtet. In der Tabelle 15 sind alle Genauigkeiten der Attribute aus der Entität Produkte enthalten.

Tabelle 15 Relationenebene Genauigkeit Produkte

| Art.-Nr. | Hersteller | Bezeich-nung | … | Waren-eingang | Ist-menge | Preis in € |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | … | 1 | 1 | 1 |
| 1 | 1 | 1 | … | 0,8 | 1 | 0 |
| 1 | 1 | 0 | … | 0,7 | 1 | 0,666 |
| 1 | 1 | 1 | … | 0,9 | 1 | 1 |
| 1 | 1 | 1 | … | 1 | 1 | 0 |
| 1 | 1 | 0 | … | 1 | 1 | 0 |
| 1 | 1 | 1 | … | 0,8 | 1 | 0,666 |
| 1 | 0 | 1 | … | 0,8 | 1 | 1 |

Um auf die Relationenebene zu gelangen, muss die Genauigkeit aller Tupel berechnet werden. Die Ergebnisse sind in Tabelle 16 enthalten.

Tabelle 16 Ergebnis Relationenebene Genauigkeit Produkte

| **Tupel** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $\boldsymbol{Q_{Korr}(t,e)}$ | 1 | 0,7 | 0,715 | 0,975 | 0,75 | 0,625 | 0,867 | 0,825 |

Mit der Formel $Q_{Gen}(T) \coloneqq \frac{\sum_{i=1}^{|T|} Q_{Gen}(t_i)}{|T|}$ ergibt sich dann $Q_{Genau}(T) \coloneqq \frac{1+0,7+0,715+0,975+0,75+0,625+0,867+0,825}{8}$ = 0,807. Auf Relationenebene liegt die Korrektheit dann bei 80,7 %.

Im letzten Schritt wird die Datenbankebene betrachtet. Dazu wird wieder neben der Entität Produkte auch die Entität Hersteller hinzugezogen. Bei der Entität Hersteller

bedarf es keiner relativen Wichtigkeit und es wieder eine Genauigkeit mit 1 bei Vollständigkeit und 0 bei Unvollständigkeit angegeben.

Tabelle 17 Relationenebene Genauigkeit Hersteller

| | $Q_{Gen}(t)$ | | |
|---|---|---|---|
| | **Tupel 1** | **Tupel 2** | **Tupel 3** |
| **Hersteller** | 1 | 1 | 1 |
| **Kontakt** | 1 | 1 | 0 |
| **Telefon** | 1 | 1 | 1 |
| **Straße** | 1 | 0 | 1 |
| **Ort** | 1 | 1 | 1 |
| **Plz** | 1 | 1 | 0 |
| $Q_{Gen}(t)$ | 1 | 0,833 | 0,667 |
| $Q_{Gen}(T)$ | | 0,833 | |

Mit der Formel auf Datenbankebene ergibt sich mit $Q_{Gen}(D) \coloneqq \frac{\sum_{k=1}^{p} Q_{Gen}(T_k)}{p}$ dann $Q_{Gen}(Lager) \coloneqq \frac{0{,}807+0{,}833}{2}$ = 0,82. Die Genauigkeit beträgt somit für die Datenbank Lager 82 %.

### 5.1.4. Aktualität

Um die Aktualität zu überprüfen, bedarf es eines weiteren Beispiels. Da es drei mögliche Metriken zur Messung der Aktualität gibt, werden diese nun auf ihre Praxistauglichkeit getestet. Dafür werden im ersten Schritt verschiedene Alter für jeden Attributwert angegeben. Es wird davon ausgegangen, dass der Preis sich erfahrungsgemäß alle 300 Tage ändert und dann nicht mehr korrekt ist. Dies bedeutet eine maximale Gültigkeitsdauer $T_{max}(A) = 300$. Die Verfallsrate von Attributwerten kann dann als $Verfall(A) = \frac{1}{300}$ angenommen werden. Zudem ist die Updatehäufigkeit am Betrachtungspunkt somit $Upd(A) = \frac{1}{300}$. Die drei Formeln, die angewendet wurden, sind zum einen $Q_{1Zeit}(\omega, A) \coloneqq \frac{1}{Upd(A)Age(\omega)+1}$, dann $Q_{2Zeit}(\omega, A) \coloneqq$

$\left[max\left\{1-\frac{Age(\omega)}{T_{max}(A)};0\right\}\right]$ und als letzte die Formel $Q_{3Zeit}(\omega,A) \coloneqq e^{-Verfall(A)*Age(\omega)}$. Die Ergebnisse sind in Tabelle 18 dargestellt.

Tabelle 18 Unterscheidung der Formeln auf Attributwertebene

| Art.-Nr. | Alter des Preises in Tagen | $Q_{1Zeit}(\omega,A)$ | $Q_{2Zeit}(\omega,A)$ | $Q_{3Zeit}(\omega,A)$ |
|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 1 |
| 2 | 10 | 0,968 | 0,967 | 0,967 |
| 3 | 50 | 0,857 | 0,833 | 0,847 |
| 4 | 150 | 0,667 | 0,5 | 0,607 |
| 5 | 250 | 0,545 | 0,167 | 0,435 |
| 6 | 300 | 0,5 | 0 | 0,368 |
| 7 | 400 | 0,429 | 0 | 0,264 |

Anhand der Ergebnisse wird deutlich, dass $Q_{1Zeit}(\omega,A)$ und $Q_{3Zeit}(\omega,A)$ die maximale Gültigkeitsdauer nicht wie angenommen berücksichtigen. Für $Q_{1Zeit}(400, Alter\ des\ Preises)$ und für $Q_{2Zeit}(400, Alter\ des\ Preises)$ müsste das Ergebnis $Q_{3Zeit}(400, Alter\ des\ Preises) = 0$ vorliegen. Die maximale Gültigkeitsdauer ist somit nicht auf die Update-Häufigkeit übertragbar. Auch der $Verfall(A)$ ist nicht wie angenommen mathematisch der Kehrwert der maximalen Gültigkeitsdauer. Die beiden ersten Formeln sehen die Gültigkeitsdauer somit nicht als absoluten Punkt an, bei dem eine Qualität von 0 eintreten würde. Die drei vorliegenden Formeln arbeiten demnach mit verschiedenen Annahmen und können nicht direkt verglichen werden. $Q_{1Zeit}(\omega,A)$ kann zum Beispiel bei Sensoren angewendet werden, bei denen ein automatisches Update den Attributwert in regelmäßigen Abständen erneuert. Wenn bekannt ist, wann die Gültigkeit eines Wertes erlischt, kann $Q_{2Zeit}(\omega,A)$ verwendet werden. Dieses erfordert jedoch eine stetige Gleichverteilung. Bei $Q_{3Zeit}(\omega,A)$ wird die Aktualität des Attributwertes exponentiell verteilt berechnet. Der Verfall wird mit der Wahrscheinlichkeit angegeben, inwieweit der Attributwert innerhalb einer Zeiteinheit noch gültig ist. In dieser Arbeit wird die Formel $Q_{3Zeit}(\omega,A)$ von Klier verwendet.

Nach der Untersuchung der verschiedenen Formeln wird mit der Berechnung der *Attributwertebene* fortgefahren. Die Tabelle 19 zeigt das Alter der jeweiligen Attributwerte für die Entität Hersteller. Für die Attribute Hersteller, Straße, Ort und Postleitzahl kann der Verfall 0 angegeben werden, da sich diese in der Regel nicht ändern. Der Ansprechpartner ändert sich nicht so schnell wie die dazugehörige Telefonnummer.

Tabelle 19 Alter der Herstellerinformation in Tagen

| | **Hersteller** | **Kontakt** | **Telefon** | **Straße** | **Ort** | **PLZ** |
|---|---|---|---|---|---|---|
| | 1500 | 1000 | 1000 | 1500 | 1500 | 1500 |
| | 400 | 100 | 100 | | 400 | 400 |
| | 1000 | | 500 | 1000 | 1000 | |
| | … | … | … | … | … | … |
| $Verfall(A)$ | 0 | 0,0001 | 0,0002 | 0 | 0 | 0 |

In Tabelle 20 wurden mit der Formel $Q_{3Zeit}(\omega, A) \coloneqq e^{-Verfall(A)*Age(\omega)}$ die einzelnen Wahrscheinlichkeiten der Attributwerte der Aktualität berechnet. Zum Beispiel wird die Wahrscheinlichkeit, dass der Kontakt von Continental richtig ist, wie folgt berechnet $Q_{3Zeit}(M\ddot{u}ller, Kontakt) = e^{-0{,}0001*1000} = 0{,}9048$ und liegt bei 90,48 %.

Tabelle 20 Tupelebene Aktualität Hersteller

| | **Hersteller** | **Kontakt** | **Telefon** | **Straße** | **Ort** | **PLZ** |
|---|---|---|---|---|---|---|
| $Q_{3Zeit}(\omega, A)$ | 1 | 0,905 | 0,819 | 1 | 1 | 1 |
| $Q_{3Zeit}(\omega, A)$ | 1 | 0,990 | 0,980 | | 1 | 1 |
| $Q_{3Zeit}(\omega, A)$ | 1 | | 0,905 | 1 | 1 | |
| | … | … | … | … | … | … |
| $g_j$ | 1 | 1 | 2 | 1 | 1 | 1 |

Nun wird mit der Aktualität auf *Tupelebene* fortgefahren. Die Aktualität des Herstellers Continental wird mit $Q_{Zeit}(t) \coloneqq \frac{\sum_{j=1}^{n} Q_{Zeit}(t.A_j, A_j) g_j}{\sum_{j=1}^{n} g_j}$ gemessen. $Q_{Zeit}(1) \coloneqq$

$\frac{1*1+0,905*1+0,819*2+1*1+1*1+1*1+1*1}{7}$ = 0,935 ergibt somit eine Aktualität bzw. Zeitnähe von 93,5 %.

Auf *Releationenebene* werden nun alle Tupel mit der Formel $Q_{Voll}(T) := \frac{\sum_{i=1}^{|T|} Q_{Voll}(t_i)}{|T|}$ betrachtet. Ist der Wert nicht vorhanden, wird dieser auch nicht auf Aktualität überprüft. Für das zweite Tupel ergibt sich mit $Q_{Zeit}(2) := \frac{1*1+0,99*1+0,98*2+1*1+1*1}{6} =$ 0,992 eine Aktualität von 99,2 % und für das dritte Tupel 96,2%. Die Aktualität der gesamten Relation ist dann $Q_{Voll}(Hersteller) := \frac{0,935+0,992+0,962}{3}$ = 0,963 und liegt damit bei 96,3 %.

Zum Schluss wird die *Datenbankebene* betrachtet. In diesem Beispiel wird angenommen, dass die Entität Produkte eine Aktualität von 100 % vorweist. Somit ergibt sich dann mit $Q_{Zeit}(Lager) := \frac{1+0,963}{2} =$ eine Aktualität der Datenbank von 98,2 %.

Im letzten Schritt muss noch die gesamte Datenqualität berechnet werden. Die Dimension Glaubwürdigkeit wurde nicht mit einer Metrik versehen. Aus diesem Grunde wird diese Qualität mit $Q_{Glaub}(D) = 0,95$ angegeben. Die gesamte Datenqualität der Datenbank berechnet sich mit dem Durchschnitt der jeweiligen Qualität der Dimensionen auf Datenbankebene und wird wie folgt definiert:

$$Q_{Gesamt}(D) := \frac{Q_{Voll} + Q_{Korr} + Q_{Gen} + Q_{Zeit} + Q_{Glaub}}{5}$$

Somit ergibt sich mit $Q_{Gesamt}(Lager) = \frac{0,855+0,866+0,82+0,982+0,95}{5} = 0,8946$ eine gesamte Qualität von 89,46 %

## 5.2. Eignung der Metriken

Im Kapitel 2.2.2 wurden Anforderungen an Metriken vorgestellt. Diese haben den Zweck, Metriken einer wissenschaftlichen Begründung zu unterziehen und ihnen die praktische Nutzung zu attestieren. Aus diesem Grund werden die behandelten

Metriken nun dahin gehend überprüft. Darüber hinaus wird diskutiert, ob die Metriken für eine Anwendung auf Big Data tatsächlich nutzbar sind und welche Probleme auftreten können.

Bei der Vollständigkeit auf Attributwertebene ist das Ergebnis auf 0 und 1 normiert. Jedoch können bei der Formel 2 die Ergebnisse nicht kardinal skaliert werden, da keine Werte zwischen 0 und 1 ausgegeben werden können. Die Ergebnisse können nicht in eine Rangordnung gebracht werden. Zusätzlich kann bei zwei Merkmalen nicht bestimmt werden, inwieweit diese sich unterscheiden. Auf Tupel-, Relationen- und Datenbankebene ist dies wiederum möglich. Mit der relativen Wichtigkeit ist auf diesen Ebenen die Sensilisierbarkeit gegeben. Auch die Aggregierbarkeit ist gewährleistet. Die Messung der Datenqualität kann von der Attributwertebene auf Tupel-, Relationen- und anschließend auf Datenbankebene aggregiert werden. Es ist verständlich, wie sich die Vollständigkeit zusammensetzt; somit ist eine fachliche Interpretation gewährleistet.

Bei der Vollständigkeit wird eine formale Spezifikation geprüft und es kann problematisch werden, wenn der Wert nicht existieren muss. So kann der Wert fälschlicherweise als unvollständig angenommen werden und es liegt ein Qualitätsdefizit vor. Ein Attribut dafür ist zum Beispiel der Name des Ehepartners. Dieses muss vor dem Anwenden der Metrik berücksichtigt werden. Es kann gegebenenfalls durch Auslassen bestimmter Attribute oder mit Berücksichtigung der relativen Wichtigkeit realisiert werden. Die Vollständigkeit ist vor allem dann wichtig, wenn relationale Datenbanken betrachtet werden. Diese Art der Strukturierung liegt bei Big Data jedoch oftmals nicht vor. Somit kann diese Dimension schnell in den Hintergrund geraten, da eine Messung der Vollständigkeit bei unstrukturierten Daten schwierig erscheint. Falls unstrukturierte Daten in strukturierte Daten umgewandelt werden können, wäre die Rolle einflussreicher. Datenpunkte müssten fehlerfrei und automatisiert umgewandelt werden, ohne wichtige Gesichtspunkte zu verlieren.

Bei den Messungen der Korrektheit ist die Normierung gegeben. Auf Attributwertebene ist die fachliche Interpretierbarkeit indessen nicht bei Formeln gewährleistet. Falls das Abstandsmaß mit der Formel 8 gewählt wird, kann mit der Formel 9 lediglich eine Korrektheit von 0 oder 1 aufgezeigt werden. Somit fehlen die kardinale Skalierung und die damit einhergehende Vergleichbarkeit. Mit den Formeln 8 und 9 kann zudem keine Korrektheit von 0 ausgegeben werden. Es sei denn, das Ab-

standsmaß beträgt unendlich, was in der Praxis nicht vorkommt. Bei der neu entwickelten Formel 21 für die Überprüfung numerischer Werte liegt das gleiche Problem vor. Somit sind diese nicht zielführend. In einem weiteren Beispiel wurden die Namen Jannic mit Yanick sowie die Begriffe Quiz und Zaun mit der Hamming-Distanz verglichen. Beide wiesen eine Korrektheit von 20 % auf, obwohl die Unterschiede sinngemäß groß sind. Zudem müsste das Abstandsmaß unendlich sein, um eine Korrektheit von 0 anzuzeigen. Dieses ist in der Praxis nicht möglich und somit mit der kardinalen Skalierung und der fachlichen Interpretation nicht zu vereinen. Bei der neu entwickelten Formel 22 für die Überprüfung von Wörtern kann sowohl eine Korrektheit von 0 als auch eine Korrektheit von 1 vorkommen. Aus diesem Grund sind eine Normierung und die kardinale Skalierung gewährleistet. Es gibt jedoch Begriffe, die gleich lauten, jedoch sinngemäß sehr unterschiedlich sind. Diese würden irrtümlich eine gewisse Korrektheit vermitteln. Zudem wird es problematisch, wenn zum Beispiel die Reihenfolge der verglichenen Buchstaben durch das Fehlen des ersten Buchstabens falsch ist. Die fachliche Interpretierbarkeit bei Wörtern scheint nicht möglich zu sein, da die Frage nach dem Sinngehalt der verglichenen Wörter nicht automatisiert gestellt werden kann. Die Metrik der Dimension Korrektheit verlangt den Abgleich der Attribute im Informationssystem mit denen der Realwelt. Dies ist bei Big Data nicht möglich. Big-Data-Analysen werden vorwiegend an Datenmengen vorgenommen, deren Inhalt nicht klar ist. Ein Abgleich und die Kontrolle nach der Korrektheit sind somit nahezu ausgeschlossen. Zudem ist die Messung der Korrektheit schon in einer relationalen Datenbank aufwendig zu realisieren. Der Kostenaufwand scheint immens und die technische Umsetzung sehr schwierig. Die Korrektheit führen viele Autoren als wichtige Dimension auf und auch diese Arbeit ist der Auffassung bis zu dieser Erkenntnis gefolgt. Darüber hinaus stellt sich die Frage, ob bei 100 % Korrektheit die anderen Dimensionen nicht entfallen. Sind Daten korrekt, sind diese auch aktuell und vollständig. Des Weiteren entfällt die Prüfung der Glaubwürdigkeit bei der Messung der Korrektheit. Bei der Genauigkeit kann diskutiert werden, ob ein Wert zwar korrekt ist, aber nicht genau in dem Maß, wie es gefordert ist. Auch die genannten Datenqualitätsmängel aus Kapitel 2.2.4. würden bei einer Korrektheit von 100 % nicht mehr vorliegen. Somit ist die Dimension Korrektheit zur Überprüfung der Datenqualität in Big Data als problematisch zu betrachten. In Kapitel 4.3. wurde die Konsistenz aufgrund der Korrektheit ausgeschlossen. Nach den neuen Erkenntnissen wäre es aussichtsreich,

anstatt der Korrektheit die Konsistenz von Daten zu überprüfen. Die Überprüfung der Widerspruchsfreiheit in Datensätzen kann technisch realisiert werden, dient der Datenqualität und scheint zielführend.
Fortlaufend wird die Genauigkeit überprüft. Eine Normierung der Ergebnisse ist durch die Formel 13 auf Attributwertebene gegeben. Auch die kardinale Skalierung ist gewährleistet und in den vorliegenden Beispielen können die Ergebnisse leicht fachlich interpretiert werden. Wie in den beiden Dimensionen davor führt die relative Wichtigkeit zur Erfüllung der Sensilisierbarkeit. Diese kann auf allen Ebenen angewendet werden und zusätzlich kann die Metrik von der Attributwertebene auf Tupel-, Relationen- und anschließend auf Datenbankebene aggregiert werden. Bei der Analyse in Kapitel 5.1.3. wurde deutlich, dass die Metrik auch Beeinträchtigungen vorweist. Ein korrekter Wert kann bei einer Überprüfung der Nachkommastellen eine Genauigkeit von 0 % ausgeben. Wenn weggelassene Kommastellen zu einem großen Qualitätsverlust führen, ist das Qualitätsmaß mit Vorsicht zu beurteilen. Bei der Betrachtung von Kalenderdaten kann eine Ungenauigkeit zu Problemen bei der Analyse führen. Neben unterschiedlichen Satzzeichen sind unterschiedliche Darstellungsformate zu betrachten. In den USA zum Beispiel wird zuerst der Monat genannt, dann der Tag und das Jahr. Der Entstehungsort der Daten kann somit auch zu Problemen bei der Genauigkeit führen. Die Stelligkeit von Hinrichs, die bei der Genauigkeit überprüft wird, ist auch auf die Klassifikationshierarchie anwendbar. Somit kann überprüft werden, ob Attributwerte den Attributen zuzuordnen sind und inwieweit diese bestimmten Klassen oder Gruppen zuzuordnen sind.
Bei der Aktualität wurden drei verschiedene Formeln auf Attributebene vorgestellt. Die Formel 15 von Hinrichs verletzt die kardinale Skalierung sowie die fachliche Interpretierbarkeit. Werte mit verschiedenen Zeiteinheiten lassen sich nach kardinaler Skalierung nicht vergleichen. Zudem kann keine Aktualität von 0 auftreten, da das Alter dazu unendlich sein müsste. Ballous Ansatz zum Messen der Aktualität (Formel 16) ist auf [0;1] normiert und diese Grenzen können auch in der Praxis erreicht werden. Eine kardinale Skalierung und die fachliche Interpretierbarkeit sind in dem Sinne nur bei einer gleichen maximalen Gültigkeitsdauer gewährleistet. Die Formel 17 zeigt die Gültigkeitsdauer der Datenwerte mit einem exponentiellen verteilten Parameter $Verfall(A)$. Auch diese Formel ist normiert und zeigt Probleme bei der Kardinalität und der fachlichen Interpretierbarkeit. Die Aktualität wird als Wahrscheinlichkeit dargestellt, sodass eine fachliche Interpretierbarkeit im weiteren

Sinne vorgenommen werden kann. Außerdem lassen sich die Wahrscheinlichkeiten bei Betrachtung unterschiedlicher Parameter kardinal skalieren, auch wenn unterschiedliche Parameter auf die Wahrscheinlichkeit zurückzuführen sind. Wie alle vorgestellten Dimensionen erfüllt die Aktualität auch die Sensibilisierbarkeit und die Aggregierbarkeit. Die Aktualität ist sehr wichtig bei Betrachtung der Datenqualität. Um die Aktualität zu messen, muss in den Metadaten das Alter der betrachteten Daten vorhanden sein. Das Alter darf nicht mit dem Zeitpunkt verwechselt werden, wann die Daten in das Informationssystem eingetragen wurden. In Kapitel 4.3. wurde bei der Auswahl der Dimensionen der Unterschied zwischen verschiedenen Daten anhand von Aktienkursen und Seekarten erläutert. Je nach Auswahl der Formel müssen Informationen über die Verfallsrate $Verfall(A)$, die maximale Gültigkeitsdauer $T_{max}(A)$ oder die Update-Häufigkeit $Upd(A)$ vorliegen. Liegen diese nicht exakt vor, muss der genaue Hintergrund der Daten festgestellt werden. Eine Schätzung des jeweiligen Parameters führt dann zur Messung der Aktualität. Vor der Bestimmtheit der Datenqualität können bestimmte Kategorien gebildet werden, damit diese Schätzung automatisch erfolgt. Zum Beispiel kann festgehalten werden, dass Telefonnummern durchschnittlich 5 Jahre gültig sind und Adressen 10 Jahre. Dies bedarf umfangreicher Vorbereitungen, sodass alle Daten einem Wert zugeordnet werden können. Zusammenfassend kann festgehalten werden, dass die Messung der Aktualität im Zusammenhang von Big Data aufwendig zu realisieren ist.

Abschließend wird die Glaubwürdigkeit der Datenqualität betrachtet. In der Studie von Wang und Strong wurde die Glaubwürdigkeit als wichtigste Dimension ausgemacht. Zudem stellt sie eine Anforderung an Big Data dar und wurde bei der Auswahl der Dimensionen als äußerst relevant eingestuft. In Kapitel 4.3. wurden außerdem bereits Beispiele und Probleme dazu behandelt. Eine Klassifikation der möglichen Datenquellen oder auch Datenarten nach Glaubwürdigkeit kann das Ergebnis positiv beeinflussen. Es besteht die Möglichkeit, eine Gewichtung der Glaubwürdigkeit vorzunehmen, falls verschiedene Datenquellen zur Datenmenge beitragen. Daten, die eher als wahr angenommen werden, werden mit einer hohen Glaubwürdigkeit bewertet. Prinzipiell ist die Glaubwürdigkeitsbeurteilung schwierig. Sie ist jedoch wichtig, wenn die Richtigkeit der Daten nicht verifiziert ist. Beurteilungen beruhen teilweise auf Erfahrungen, die nicht unbedingt objektiv vorgenommen wurden. Falls in den Metadaten die Herkunft der Daten nicht aufgeführt ist, kann eine Beurteilung nicht erfolgen und die Daten sollten nicht zur Analyse herangezogen

werden. Die Glaubwürdigkeit bei Big Data zu überprüfen, wird mit einem hohen zeitlichen Aufwand verbunden sein, spielt jedoch eine große Rolle bei der Betrachtung der gesamten Datenqualität.

Zwischen dem Interesse, so viele Daten wie möglich zur Analyse zu verwenden, und dem Anliegen, Daten von höchster Qualität zu nutzen, liegt unweigerlich ein Spannungsverhältnis. Um die gewünschte Datenqualität zu erreichen und diese zu überprüfen, sollten Dimensionen sowie Ist- und Solldatenqualität festgehalten werden. Falls die geforderte Datenqualität nicht erreicht wird, sind Maßnahmen zur Verbesserung der jeweiligen Dimension einzuleiten. Liegt die geforderte Datenqualität vor, kann die Datenanalyse stattfinden. Wenn mögliche Maßnahmen nicht realisiert werden können und die Datenqualität nicht auf ein akzeptables Maß steigt, müssen andere Daten für die Analyse verwendet oder Teile der Daten entfernt werden.

# 6. Fazit

In diesem Abschnitt werden die Ergebnisse der vorliegenden Arbeit bewertet. Es wurde festgestellt, dass sich die Anforderungen an die Datenqualität von Big Data gegenüber anderen Datenbeständen unterscheiden. Die charakterisierenden Eigenschaften von Big Data, die mit den Vs dargestellt werden, erschweren die Bewertung der Datenqualität. Dazu gehören unter anderem der hohe Anteil an unstrukturierten Daten und die große Datenmenge. Darüber hinaus muss aufgrund des zeitlichen Verfalls der Daten die Datenqualitätsmessung sehr schnell erfolgen, damit die Analyse der Daten ebenso zeitnah erfolgen kann. Aus diesem Grund setzt sich die Datenqualität aus anderen Dimensionen zusammen, als sie teilweise in der Literatur vorgegeben wurden. Zudem wurde die Forschungsfrage zur Unterscheidung von Daten- und Informationsqualität beantwortet. Die Datenqualität kann unabhängig betrachtet werden, da der Kontext zur Beurteilung der Qualität nicht zwingend erforderlich ist. In einer Betrachtung der zahlreichen in der Literatur genannten Dimensionen wurde erörtert, welche der Datenqualität und welche der Informationsqualität zugeordnet werden können. Zu den bedeutsamen der Datenqualität wurden vor der Analyse Vollständigkeit, Genauigkeit, Aktualität, Glaubwürdigkeit sowie Korrektheit ermittelt. Außerdem wurde eine Abhängigkeit von Datenqualität und Informationsqualität beschrieben. Ohne die Datenqualität kann keine Informationsqualität vorliegen. Neben den Dimensionen der Datenqualität gehören zur Informationsqualität Interpretierbarkeit, Objektivität, Relevanz, Durchführbarkeit und die Wertschöpfung. Die erfassten Dimensionen zur Datenqualität wurden mit Metriken versehen und mithilfe eins praxisnahen Beispiels erläutert. Die Bewertung zur Quantifizierung der Datenqualität ist vielversprechend, da sie eine größtenteils objektive und automatisierte Messung verspricht. Durch die Durchführung wurden Probleme der Metriken identifiziert und Konflikte zwischen einzelnen Dimensionen deutlich.
Ein Kritikpunkt an dieser Arbeit ist die Durchführung an einer relationalen Datenbank. Es wurde ausgemacht, dass Big Data zu einem großen Teil aus unstrukturierten Daten besteht, die nicht in einer relationalen Datenbank gegliedert sind. Außerdem wurde keine bestehende Datenmenge verwendet, sondern eine eigene modelliert. Die Praxistauglichkeit kann somit nicht vollends sichergestellt werden, je-

doch konnten aufgrund dieser Vorgehensweise alle Bewertungsmethoden angewendet und Probleme identifiziert werden. Die Anforderungen an die Metriken, die zu einer wissenschaftlichen Begründung verlangt werden, konnten größtenteils nachgewiesen wiesen. Zu diesen gehören Normierung, Kardinalität, Sensibilisierbarkeit, Aggregierbarkeit und die fachliche Interpretation. Umstritten bleiben die Normierung, die kardinale Skalierung und die damit verbundene fachliche Interpretation bei den Abstandsmaßen. Die Überprüfung der formalen Spezifikation Vollständigkeit kann vorwiegend bei strukturierten Daten vorgenommen werden. Können unstrukturierte Daten in dieses Format überführt werden, besitzt die Metrik eine größere Relevanz. Die Korrektheit erwies sich als ungeeignet bei der Betrachtung der Datenqualität in Big Data. Es ist unmöglich, die Messung der Dimension zu realisieren. Werte der Realwelt, die in Big Data nicht zu überprüfen sind, müssten mit den Werten des Informationssystems verglichen werden. Außerdem würden andere Dimensionen, deren Messungen umzusetzen sind, entfallen. Vor der Analyse wurde die Konsistenz aufgrund der Korrektheit ausgeschlossen, nach genauer Betrachtung kann der Konsistenz somit eine höhere Bedeutung zugewiesen werden. Die Genauigkeit überprüft die Stelligkeit von Daten. Diese Dimension ist vor allem bei numerischen Werten sinnvoll und leicht zu realisieren. Die Überprüfung der Aktualität ist dagegen anspruchsvoller und kann unter Umständen bei Schätzungen des Verfalls subjektive Einflüsse einbeziehen. Bei der letzten zu untersuchenden Dimension ist eine erforderliche neutrale Bewertung nicht zweifelsfrei gegeben. Bei der Messung der Glaubwürdigkeit muss gegebenenfalls mit Erfahrungseigenschaften der Daten und Datenquellen gearbeitet werden. Dieser Art der Einflüsse sollte bei einer Bewertung im Grunde genommen keine große Bedeutung beigemessen werden, sind in diesem Fall dennoch vonnöten. Eine vollständige Antwort auf die Frage, ob diese Metriken im Anwendungsfall von Big Data mit all den Anforderungen, die gestellt werden, wirklich zur effektiven Beurteilung der Datenqualität herangezogen werden können, kann nicht abschließend gegeben werden.

# 7. Zusammenfassung und Ausblick

Die vorliegende Arbeit mit dem Ziel zur Erstellung eines Datenqualitätskonzeptes im Kontext der Eigenschaften von Big Data zeigt die Komplexität des Wunsches zur Kontrolle der Datenflut, der nahezu alle Bereiche ausgesetzt sind. Zu Beginn in Kapitel 2 wurde mithilfe der Wissenspyramide, die Zeichen, Daten, Information und Wissen beinhaltet, eine Begriffserklärung vorgenommen. Darauf aufbauend wurde die Datenqualität erläutert sowie die Studie von Wang und Strong, die unumstritten der Baustein vieler Ausarbeitungen in diesem Themenfeld ist, behandelt. Anschließend wurden Metriken aus der Dissertation von Hinrich, die oftmals in der Literatur bei Beiträgen zu Messmethoden in Erscheinung treten, detailliert dargestellt. Nach der Betrachtung des Datenqualitätsregelkreises und der möglichen Mängel wurde in Kapitel 3 der theoretische Rahmen zu Big Data geschaffen. Neben den Merkmalen von Big Data, die größtenteils die Anforderungen an die Dimensionen stellen, wurden sowohl Datentypen als auch Datenquellen thematisiert und damit der Abschluss der notwendigen Grundlagen geschaffen. In Kapitel 4 wurden die Grundlagen zusammengetragen, um eine neue Definition für die Datenqualität im Kontext von Big Data zu erarbeiten. Dazu wurden die Anforderungen, die Big Data stellt, zusammengetragen sowie der genaue Unterschied zwischen Daten- und Informationsqualität erarbeitet. Bei der Auswahl der Dimensionen wurden die in der Literatur genannten Dimensionen im Zusammenhang von Big Data auf die Daten- sowie die Informationsqualität analysiert. Im anschließenden Kapitel 4.3. konnten somit neue Definitionen für beide Begriffe erarbeitet und diese mit den dazugehörigen Dimensionen charakterisiert werden. Das Kapitel 5 diente zur Bewertung der Datenqualität mit den Metriken aus der Literatur. Dazu wurde eine Datenbasis geschaffen und die Dimensionen Vollständigkeit, Korrektheit, Genauigkeit und Aktualität geprüft. Es wurden darüber hinaus neue Formeln entwickelt und diverse Schwierigkeiten festgestellt. Dieses wurde auf Attribut-, Tupel-, Relationen- und Datenbankebene durchgeführt und somit konnte abschließend eine gesamte Datenqualität für den vorliegenden Fall erarbeitet werden. Im Kapitel 5.2 wurde neben der Kontrolle der Anforderungen an die Metriken auch die Diskussion über die Praxistauglichkeit der Bewertungsmethode in Big Data durchgeführt. Abschließend wurde eine Handlungsempfehlung aufgestellt, wie mit der Datenqualität umzugehen ist und welche

Maßnahmen zur Erhöhung der Datenqualität beitragen. Mit dem Fazit wurden die Forschungsfragen, die mit der Arbeit beantwortet werden sollten, und die erreichten Ziele der Ausarbeitung zusammengefasst. Insbesondere durch die literarische Arbeit konnten die Begriffe Daten- und Informationsqualität abgegrenzt und neue charakterisierende Dimensionen gefunden werden. Auch eine Bewertungsmethode durch die Metriken konnte veranschaulicht und Schwierigkeiten zur automatisierten Messung aufgezeigt werden. Offene Fragen blieben jedoch bei der Thematik der unstrukturierten Daten, da lediglich strukturierte Daten mithilfe der Bewertungsmethode behandelt wurden. Diese Arbeit konnte grundsätzliche Probleme in diesem Zusammenhang aufdecken, jedoch keine automatisierte Messung der Datenqualität an einer Datenmenge vornehmen, die kennzeichnend für Big Data ist. Darüber hinaus wird Big Data oftmals zu einer Analyse einer bestimmten Fragestellung herangezogen, sodass der Informationsqualität eine wichtigere Rolle als der Datenqualität zugeschrieben werden kann. Aus diesem Grund sollte aufbauend auf dieser Arbeit die Informationsqualität mit den charakterisierenden Dimensionen betrachtet und mögliche Metriken analysiert werden.

# Literaturverzeichnis


Apel, D., Behme, W., Eberlein, R., & Merighi, C. (2015). *Datenqualität erfolgreich steuern: Praxislösungen für Business-Intelligence-Projekte*. dpunkt.verlag.

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

Ballou, D., Wang, R., Pazer, H., & Tayi, G. (1998). Modeling Information Manufacturing Systems to Determine Information Product Quality. *Management Science*, *44*, 462–484. https://doi.org/10.1287/mnsc.44.4.462

Bamberg, G., Baur, F., & Krapp, M. (2017). Statistik: Eine Einführung für Wirtschafts- und Sozialwissenschaftler. In *Statistik*. De Gruyter Oldenbourg. https://doi.org/10.1515/9783110495720

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for Data Quality Assessment and Improvement. *ACM Computing Surveys*, *41*(3), 16-16.52. https://doi.org/10.1145/1541880.1541883

Batini, C., & Scannapieco, M. (2016). *Data and Information Quality: Dimensions, Principles and Techniques*. Springer International Publishing. https://doi.org/10.1007/978-3-319-24106-7

Baur, N. (2009). Measurement and Selection Bias in Longitudinal Data. A Framework for Re-Opening the Discussion on Data Quality and Generalizability of Social Bookkeeping Data. *Historical Social Research / Historische Sozialforschung*, *34*(3 (129)), 9–50.

Baur, N., Graeff, P., Braunisch, L., & Schweia, M. (2020). The Quality of Big Data. Development, Problems, and Possibilities of Use of Process-Generated Data in the Digital Age. *Historical Social Research*, *45*(3), 209–243. https://doi.org/10.12759/hsr.45.2020.3.209-243

Bawden, D. (2001). The shifting terminologies of information. *Aslib Proceedings*, *53*(3), 93–98. https://doi.org/10.1108/EUM0000000007043

Bell, W. D. (1957). *A management guide to electronic computers.* McGraw-Hill. https://catalog.hathitrust.org/Record/001118356

Bodendorf, F. (2006). *Daten- und Wissensmanagement*. Springer-Verlag.

Borgmann, A. (1999). *Holding On to Reality: The Nature of Information at the Turn*

*of the Millennium*. University of Chicago Press. http://ebookcentral.proquest.com/lib/dortmundtech/detail.action?docID=408313

Bovee, M., Srivastava, R. P., & Mak, B. (2003). A conceptual framework and belief-function approach to assessing overall information quality. *International Journal of Intelligent Systems*, *18*(1), 51–74. https://doi.org/10.1002/int.10074

Buhl, H. U., Röglinger, M., Moser, F., & Heidemann, J. (2013). Big Data. *WIRTSCHAFTSINFORMATIK*, *55*(2), 63–68. https://doi.org/10.1007/s11576-013-0350-x

Capurro, R. (1978). *Information: Ein Beitrag zur etymologischen und ideengeschichtlichen Begründung des Informationsbegriffs*. Saur.

Chapin, N. (1957). *An Introduction to Automatic Computers*. Van Nostrand.

Cleve, J., & Lämmel, U. (2016). Data Mining. In *Data Mining*. De Gruyter Oldenbourg. https://www.degruyter.com/document/doi/10.1515/9783110456776/html

Debattista, J., Lange, C., Scerri, S., & Auer, S. (2015). Linked „Big“ Data: Towards a Manifold Increase in Big Data Value and Veracity. *2015 IEEE/ACM 2nd International Symposium on Big Data Computing (BDC)*, 92–98. https://doi.org/10.1109/BDC.2015.34

Deutscher Dialogmarketing Verband e. V (Hrsg.). (2016). *Dialogmarketing Perspektiven 2015/2016*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-12924-8

DIN EN ISO 9000 (Hrsg.). (2015). *Qualitätsmanagementsysteme: Grundlagen und Begriffe (ISO 9000:2015-11) ; deutsche und englische Fassung EN ISO 9000:2015 ; DIN EN ISO 9000*. Beuth Verlag GmbH.

Dorschel, J. (Hrsg.). (2015). *Praxishandbuch Big Data*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-07289-6

*Dudenredaktion*. (o. J.). Abgerufen 3. August 2021, von https://www.duden.de/rechtschreibung/Daten

Duxa, S., Hu, A., & Schmenk, B. (2005). *Grenzen überschreiten: Menschen, Sprachen, Kulturen : Festschrift für Inge Christine Schwerdtfeger zum 60. Geburtstag*. Gunter Narr Verlag.

Eppler, M. J. (2006). *Managing Information Quality: Increasing the Value of Information in Knowledge-intensive Products and Processes*. Springer Science

& Business Media.

Favaretto, M., Clercq, E. D., Schneble, C. O., & Elger, B. S. (2020). What is your definition of Big Data? Researchers' understanding of the phenomenon of the decade. *PLOS ONE*, *15*(2), e0228987. https://doi.org/10.1371/journal.pone.0228987

Floridi, L. (2005, November). *Is Information Meaningful Data?* [Preprint]. http://philsci-archive.pitt.edu/2536/

Floridi, L. (2014). *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. Oxford University Press.

Fox, C., Levitin, A., & Redman, T. (1994). The notion of data and its quality dimensions. *Information Processing & Management*, *30*(1), 9–19. https://doi.org/10.1016/0306-4573(94)90020-5

Gadatsch, A. (2017). Big Data – Datenanalyse als Eintrittskarte in die Zukunft. In A. Gadatsch & H. Landrock (Hrsg.), *Big Data für Entscheider: Entwicklung und Umsetzung datengetriebener Geschäftsmodelle* (S. 1–10). Springer Fachmedien. https://doi.org/10.1007/978-3-658-17340-1_1

Garvin, D. A. (1984). What Does „Product Quality" Really Mean? *Sloan Management Review*, *26*(1), 25–43. Periodicals Archive Online; Periodicals Index Online.

Gebauer, M., & Windheuser, U. (2021). Strukturierte Datenanalyse, Profiling und Geschäftsregeln. In K. Hildebrand, M. Gebauer, & M. Mielke (Hrsg.), *Daten- und Informationsqualität: Die Grundlage der Digitalisierung* (S. 87–100). Springer Fachmedien. https://doi.org/10.1007/978-3-658-30991-6_5

Geiger, W., & Kotte, W. (2008). *Handbuch Qualität: Grundlagen und Elemente des Qualitätsmanagements: Systeme - Perspektiven* (5., vollst. überarb. und erw. Aufl). Vieweg.

Gölzer, P. (2017). *Big Data in Industrie 4.0—Eine strukturierte Aufarbeitung von Anforderungen, Anwendungsfällen und deren Umsetzung* [Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)]. https://opus4.kobv.de/opus4-fau/frontdoor/index/index/docId/8106

Gräfe, G., & Maaß, C. (2021). Bedeutung der Informationsqualität bei Kaufentscheidungen im Internet. In K. Hildebrand, M. Gebauer, & M. Mielke (Hrsg.), *Daten- und Informationsqualität: Die Grundlage der Digitalisierung* (S. 171–193). Springer Fachmedien. https://doi.org/10.1007/978-3-658-30991-6_10

Gray, R. L. (2003). Brief Historical Review of the Development of the Distinction Between Data and Information. *in the Information Systems Literature," in J. Ross and D. Galletta (Eds.), Proceedings of the 9th Americas Conference on Information Systems*, 2843–2849.

Gutierrez, F. (2021). *Spring Cloud Data Flow: Native Cloud Orchestration Services for Microservice Applications on Modern Runtimes*. Apress. https://doi.org/10.1007/978-1-4842-1239-4

Hackett, D. (2016). *Big Data in Life Insurance* (S. 12). MLC Life Insurance. https://www.mlc.com.au/content/dam/mlc/documents/pdf/media-centre/big-data-report.pdf

Hacking, I. (1992). The self-vindication of the laboratory sciences. In A. Pickering, *Science as Practice and Culture* (S. 29–64). University of Chicago Press. http://ebookcentral.proquest.com/lib/dortmundtech/detail.action?docID=625217

Harrington, J. H. (1991). *Business Process Improvement: The Breakthrough Strategy for Total Quality, Productivity, and Competitiveness | H. Harrington | download*. McGraw-Hill. https://b-ok.cc/book/696725/a3eb79

Heinrich, B., & Klier, M. (2015). Datenqualitätsmetriken für ein ökonomisch orientiertes Qualitätsmanagement. In K. Hildebrand, M. Gebauer, H. Hinrichs, & M. Mielke (Hrsg.), *Daten- und Informationsqualität: Auf dem Weg zur Information Excellence* (S. 49–67). Springer Fachmedien. https://doi.org/10.1007/978-3-658-09214-6_3

Heinrich, B., & Klier, M. (2021). Datenqualitätsmetriken für ein ökonomisch orientiertes Qualitätsmanagement. In K. Hildebrand, M. Gebauer, & M. Mielke (Hrsg.), *Daten- und Informationsqualität: Die Grundlage der Digitalisierung* (S. 47–65). Springer Fachmedien. https://doi.org/10.1007/978-3-658-30991-6_3

Heravizadeh, M., Mendling, J., & Rosemann, M. (2008). *Dimensions of Business Processes Quality (QoBP)*. *17*, 80–91. https://doi.org/10.1007/978-3-642-00328-8_8

Hinrichs, H. (2002). *Datenqualitätsmanagement in Data Warehouse-Systemen* [PhD Thesis].

Holthuis, J. (1999). Informationen für das Management. In J. Holthuis (Hrsg.), *Der*

*Aufbau von Data Warehouse-Systemen: Konzeption—Datenmodellierung—Vorgehen* (S. 16–35). Deutscher Universitätsverlag. https://doi.org/10.1007/978-3-322-92336-3_3

Horvath, S. (2013). *Aktueller Begriff: Big Data*. Wissenschaftliche Dienste des Deutschen Bundestages. https://www.bundestag.de/resource/blob/194790/c44371b1c740987a7f6fa74c06f518c8/big_data-data.pdf

Iivari, J., & Koskela, E. (1987). The PIOCO Model for Information Systems Design. *MIS Quarterly*, *11*(3), 401–419. https://doi.org/10.2307/248688

Ingold, M. (2011). Information als Gegenstand von Informationskompetenz. Eine Begriffsanalyse. In *Ingold, Marianne (2011). Information als Gegenstand von Informationskompetenz. Eine Begriffsanalyse. Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft: Vol. 294. Berlin: Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin* (Bd. 294). Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin. https://doi.org/10.7892/boris.84816

Jaekel, M. (2017). *Die Macht der digitalen Plattformen*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-19178-8

Jain, A., Patel, H., Nagalapatti, L., Gupta, N., Mehta, S., Guttula, S., Mujumdar, S., Afzal, S., Sharma Mittal, R., & Munigala, V. (2020). Overview and Importance of Data Quality for Machine Learning Tasks. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3561–3562. https://doi.org/10.1145/3394486.3406477

Jensen, U., Katsanidou, A., & Zenk-Möltgen, W. (2011). *Metadaten und Standards*.

Kauffmann, C. (1996). Qualität. In F.-P. Burkard & P. Prechtl (Hrsg.), *Metzler-Philosophie-Lexikon* (S. 429–430). Metzler.

King, S. (2014). *Big Data*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-06586-7

King, W. R., & Epstein, B. J. (1983). Assessing Information System Value: An Experimental Study. *Decision Sciences*, *14*(1), 34–45. https://doi.org/10.1111/j.1540-5915.1983.tb00167.x

Klein, D., Tran-Gia, P., & Hartmann, M. (2013). Big Data. *Informatik-Spektrum*, *36*(3), 319–323. https://doi.org/10.1007/s00287-013-0702-3

Klier, M. (2008). Metriken zur Bewertung der Datenqualität – Konzeption und praktischer Nutzen. *Informatik-Spektrum*, *31*(3), 223–236. https://doi.org/10.1007/s00287-007-0206-0

Klier, M., & Heinrich, B. (2016). Datenqualität als Erfolgsfaktor im Business Analytics. *Controlling*, *28*(8–9), 488–494. https://doi.org/10.15358/0935-0381-2016-8-9-488

Knight, S., & Burn, J. (2005). Developing a Framework for Assessing Information Quality on the World Wide Web. *Informing Science*, *8*, 159–172. https://doi.org/10.28945/493

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety. In *Application Delivery Strategies*. META Group. https://www.bibsonomy.org/bibtex/742811cb00b303261f79a98e9b80bf49?lang=de

Lee, Y. W., Funk, J. D., Pipino, L. L., & Wang, R. Y. (2006). *Journey to Data Quality*. MIT Press. http://ebookcentral.proquest.com/lib/dortmundtech/detail.action?docID=3338616

Leonelli, S. (2016). Data-Centric Biology: A Philosophical Study. In *Data-Centric Biology*. University of Chicago Press. https://doi.org/10.7208/9780226416502

Loshin, D. (2010). *The Practitioner's Guide to Data Quality Improvement*. Elsevier.

McGilvray, D. (2008). Chapter 3—The Ten Steps Process. In D. McGilvray (Hrsg.), *Executing Data Quality Projects* (S. 62–236). Morgan Kaufmann. https://doi.org/10.1016/B978-012374369-5.50005-2

Meier, A., & Kaufmann, M. (2016). *SQL- & NoSQL-Datenbanken*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-47664-2

Miller, H. (1996). THE MULTIPLE DIMENSIONS OF INFORMATION QUALITY. *Information Systems Management*, *13*(2), 79–82. https://doi.org/10.1080/10580539608906992

Moser, H. (2021). Datenqualitäts-Modell der Volkswagen Financial Services AG. In K. Hildebrand, M. Gebauer, & M. Mielke (Hrsg.), *Daten- und Informationsqualität: Die Grundlage der Digitalisierung* (S. 421–435). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-30991-6

Müller, J. (2000). *Transformation operativer Daten zur Nutzung im Data Warehouse*. Deutscher Universitätsverlag. https://doi.org/10.1007/978-3-663-09052-6

Nöth, W. (2000). Zeichen und System. In W. Nöth (Hrsg.), *Handbuch der Semiotik* (S. 131–226). J.B. Metzler. https://doi.org/10.1007/978-3-476-03213-3_3

Olson, J. (2003). Data Quality: The Accuracy Dimension. In *Data Quality: The Accuracy Dimension*.

Otte, R., Wippermann, B., & Otte, V. (2018). *Von Data Mining bis Big Data: Handbuch für die industrielle Praxis*. Hanser.

Peralta, V. (2006). *Data Quality Evaluation in Data Integration Systems* [Phdthesis, Université de Versailles-Saint Quentin en Yvelines ; Université de la République d'Uruguay]. https://tel.archives-ouvertes.fr/tel-00325139

Pietsch, W. (2021). *Big Data* (1. Aufl.). Cambridge University Press. https://doi.org/10.1017/9781108588676

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data Quality Assessment. *Communications of the ACM*, *45*(4), 211–218. https://doi.org/10.1145/505248.506010

Piro, A., & Gebauer, M. (2021). Definition von Datenarten zur konsistenten Kommunikation im Unternehmen. In K. Hildebrand, M. Gebauer, & M. Mielke (Hrsg.), *Daten- und Informationsqualität: Die Grundlage der Digitalisierung* (S. 143–156). Springer Fachmedien. https://doi.org/10.1007/978-3-658-30991-6_8

Rahm, E., & Do, H. H. (2000). *Data Cleaning: Problems and Current Approaches*. 11.

Redman, T. C. (2017). *Seizing Opportunity in Data Quality*. MIT Sloan Management Review. https://sloanreview.mit.edu/article/seizing-opportunity-in-data-quality/

Rohweder, J. P., Kasten, G., Malzahn, D., Piro, A., & Schmid, J. (2021). Informationsqualität – Definitionen, Dimensionen und Begriffe. In K. Hildebrand, M. Gebauer, H. Hinrichs, & M. Mielke (Hrsg.), *Daten- und Informationsqualität: Auf dem Weg zur Information Excellence* (S. 23–43). Springer Fachmedien. https://doi.org/10.1007/978-3-658-09214-6_2

Sidi, F., Hassany Shariat Panahy, P., Affendey, L., A. Jabar, M., Ibrahim, H., & Mustapha, A. (2013). *Data quality: A survey of data quality dimensions*. 300–304. https://doi.org/10.1109/InfRKM.2012.6204995

Steiner, R. (2017). *Grundkurs Relationale Datenbanken*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-17979-3

Stock, S. (2001). *Modellierung zeitbezogener Daten im Data Warehouse*. Deutscher Universitätsverlag. https://doi.org/10.1007/978-3-322-90963-3

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. *Communications of the ACM*, *40*(5), 103–110. https://doi.org/10.1145/253769.253804

Treiblmaier, H. (2011). Datenqualität und Validität bei Online-Befragungen. *der markt*, *50*(1), 3–18. https://doi.org/10.1007/s12642-010-0030-y

Vo, H., & Silva, C. (2016). Programming with Big Data. In *Big Data and Social Science* (S. 125–144).

Vollmuth, J. H., & Zwettler, R. (2016). *Kennzahlen*. Haufe-Lexware GmbH & Co. KG; (c) Haufe-Lexware GmbH & Co. KG. https://www.wiso-net.de/document/AHAU__9783648081662252

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, *39*(11), 86–95. https://doi.org/10.1145/240455.240479

Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, *12*(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099

Witte, F. (2018). *Metriken für das Testreporting*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-19845-9

Würthele, V. G. (2003). *Datenqualitätsmetrik für Informationsprozesse: Datenqualitätsmanagement mittels ganzheitlicher Messung der Datenqualität* [Doctoral Thesis, ETH Zurich]. https://doi.org/10.3929/ethz-a-004650156

Zech, R. (2015). *Qualitätsmanagement und gute Arbeit*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-07504-0

Zemanek, H. (1986). Information und Ingenieurwissenschaft. In C. Hackl, *Der Informationsbegriff in Technik und Wissenschaft* (Bd. 18).

# Anhang

| Dimension | Definition |
|---|---|
| Aktualität (Currency) | Die Aktualität ist der Grad, in dem ein Datum aktuell ist. Ein Bezugswert ist aktuell, wenn er trotz möglicher Abweichungen, die durch zeitbedingte Änderungen des korrekten Wertes verursacht werden, korrekt ist (vgl. Batini et al., 2009, S. 8)<br>Aktualität beschreibt die Gegenwartsbezogenheit eines Datenbestandes. Bilden die Datenwerte nach den aktuellen Gegebenheiten den Realwert ab oder sind diese durch den zeitlichen Verfall veraltet (vgl. Klier & Heinrich, 2016, S. 490). |
| Angemessene Menge an Daten (Appropriate amount of data) | Zu erweitern, welche Datenmenge für die jeweilige Aufgabe geeignet ist (vgl. Pipino et al., 2002, S. 212). |
| Benutzer- und Wartungsfreundlichkeit (Ease of Use and maintainability) | Ein Maß dafür, inwieweit Daten zugänglich sind, verwendet, aktualisiert, gepflegt und verwaltet werden können (vgl. McGilvray, 2008, S. 111). |
| Darstellungsqualität (Presentation Quality) | Ein Maß dafür, wie die Informationen den Nutzern präsentiert und von ihnen erfasst werden. Format und Aussehen unterstützen die angemessene Nutzung der Informationen (vgl. McGilvray, 2008, S. 111). |
| Daten Spezifikation (Data specification) | Ein Maß für die Existenz, Vollständigkeit, Qualität und Dokumentation von Datenstandards, Datenmodellen, Geschäftsregeln, Metadaten und Referenzdaten (vgl. McGilvray, 2008, S. 111). |

| | |
|---|---|
| Datenabdeckung (Data Coverage) | Ein Maß für die Verfügbarkeit und den Umfang von Daten im Vergleich zum gesamten Datenuniversum oder der betrachteten Menge (vgl. McGilvray, 2008, S. 111). |
| Datenverfall (Data Decay) | Ein Maß für die Rate der negativen Veränderung von Daten (vgl. McGilvray, 2008, S. 111). |
| Duplikate/ Vervielfältigung (Duplication) | Ein Maß für die unerwünschte Duplizierung innerhalb oder zwischen Systemen für ein bestimmtes Feld oder Datensatz (vgl. McGilvray, 2008, S. 111). |
| Durchführbarkeit (Transactability) | Ein Maß dafür, inwieweit die Daten zu dem gewünschten Geschäftsvorgang oder Ergebnis führen (vgl. McGilvray, 2008, S. 111). |
| Frei von Fehlern (Free-of-error) | Welche Daten korrekt und zuverlässig sind (vgl. Pipino et al., 2002, S. 212). |
| Genauigkeit (Accuracy) | Daten sind genau, wenn die in der Datenbank gespeicherten Datenwerte den realen Werten entsprechen (vgl. D. P. Ballou & Pazer, 1985, S. 153).<br>Das Ausmaß, in dem Daten korrekt, zuverlässig und zertifiziert sind (vgl. Wang & Strong, 1996, S. 31).<br>Die Eigenschaft, dass die Attributwerte in dem jeweils „optimalen Detaillierungsgrad vorliegen (vgl. Hinrichs, 2002, S. 30). |
| Glaubwürdigkeit (Believability) | Ausmaß, in dem Informationen als wahr und glaubwürdig angesehen werden (vgl. Pipino et al., 2002, S. 212). |
| Grundlagen der Datenintegrität (Data integrity fundamentals) | Ein Maß für die Existenz, die Gültigkeit, die Struktur, den Inhalt und andere grundlegende Merkmale der Daten (vgl. McGilvray, 2008, S. 111). |
| Gültigkeit (Validity) | Eine Information hat dann Gültigkeit, wenn sie als wahrhaftig überprüft werden kann und angemessene Standards in Bezug auf andere Dimensionen hat (vgl. Miller, 1996, S. 81). |
| Interpretierbarkeit (Interpretability) | Das Ausmaß, in dem die Daten in angemessener Sprache und in angemessenen Einheiten vorliegen und die Datendefinitionen klar sind (vgl. Wang & Strong, 1996, S. 31). |

| | |
|---|---|
| Kompaktheit (Concise) | Ausmaß, in dem Informationen kompakt dargestellt werden, ohne überwältigend zu sein (d. h. kurz, aber vollständig und auf den Punkt gebracht) (vgl. Wang & Strong, 1996, S. 31). |
| Konsistente Repräsentation (Consistent Representation) | Das Ausmaß, in dem die Daten in demselben Format präsentiert werden (vgl. Pipino et al., 2002, S. 212).<br>Das Ausmaß, in dem die Daten im gleichen Format präsentiert werden und mit früheren Daten kompatibel sind (vgl. Wang & Strong, 1996, S. 32). |
| Konsistenz (Consistency) | Ein Maß für die logische Widerspruchsfreiheit von Attributwerten eines Datenprodukts (vgl. Hinrichs, 2002, S. 30). |
| Konsistenz und Synchronisierung (Consistency and Synchronization) | Ein Maß für die Äquivalenz von Informationen, die in verschiedenen Datenspeichern, Anwendungen und Systemen verwendet werden, sowie für die Prozesse zur Herstellung der Datenäquivalenz (vgl. McGilvray, 2008, S. 111). |
| Korrektheit | Attributwerte eines Datenprodukts im Informationssystem stimmen mit denen der modellierten Entität (in der Diskurswelt) überein (vgl. Hinrichs, 2002, S. 30). |
| Leichtigkeit der Manipulation (Ease of manipulation) | Das Ausmaß, in dem Daten leicht zu verarbeiten und auf verschiedene Aufgaben anzuwenden sind (vgl. Pipino et al., 2002, S. 212). |
| Lernvermögen (Learnability) | Es bedeutet die Fähigkeit, dem Benutzer zu ermöglichen, sie zu erlernen (vgl. Heravizadeh et al., 2008, S. 84). |
| Navigation (Navigation) | Ausmaß, in dem Daten leicht auffindbar und mit ihnen verknüpft sind (vgl. Knight & Burn, 2005, S. 162). |
| Neuheit (Freshness) | Neuheit stellt eine Familie von Qualitätsfaktoren dar, von denen jeder einen Neuheitsaspekt repräsentiert und seine eigene Metrik hat (vgl. Peralta, 2006, S. 7–8). |
| Nutzbarkeit (Useability) | Das Ausmaß, in dem Informationen klar und einfach zu nutzen sind (vgl. Knight & Burn, 2005, S. 162). |

| | |
|---|---|
| Nützlich (Useful) | Ausmaß, in dem die Informationen für die jeweilige Aufgabe anwendbar und hilfreich sind (vgl. Knight & Burn, 2005, S. 162). |
| Objektivität (Objectivity) | Ausmaß, in dem Informationen unvoreingenommen, vorurteilsfrei und unparteiisch sind (vgl. Wang & Strong, 1996, S. 32). |
| Pünktlichkeit und Verfügbarkeit (Timeliness and Availability) | Ein Maß dafür, inwieweit die Daten aktuell sind und für die Nutzung wie angegeben und in dem Zeitrahmen, in dem sie erwartet werden, zur Verfügung stehen (vgl. McGilvray, 2008, S. 111). |
| Rechtzeitigkeit oder Aktualität (Timeliness) | Die Aktualität gibt an, wie aktuell die Daten in Bezug auf die Aufgabe sind, für die sie verwendet werden (vgl. Pipino et al., 2002, S. 214).<br>Die Aktualität bezieht sich nur auf die Verzögerung zwischen der Änderung eines Zustands der realen Welt und der daraus resultierenden Änderung des Zustands des Informationssystems (vgl. Wand & Wang, 1996, S. 93).<br>Die Aktualität hat zwei Komponenten: Alter und Volatilität. Alter oder Aktualität ist ein Maß dafür, wie alt die Information ist, basierend darauf, wie lange sie aufgezeichnet wurde. Die Volatilität ist ein Maß für die Informationsinstabilität, d. h. die Häufigkeit der Änderung des Werts eines Entitätsattributs (vgl. Bovee et al., 2003, S. 57-58).<br>Die Aktualität bezieht sich auf die Zeit, die für die Zugänglichkeit von Informationen erwartet wird. Die Aktualität kann gemessen werden als die Zeitspanne zwischen dem Zeitpunkt, an dem die Information erwartet wird, und dem Zeitpunkt, an dem sie zur Nutzung bereitsteht (vgl. Loshin, 2010, S. 141). |
| Relevanz (Relevancy) | Ausmaß, in dem die Informationen für die jeweilige Aufgabe anwendbar und hilfreich sind (vgl. Wang & Strong, 1996, S. 31). |

| | |
|---|---|
| Reputation (Reputation) | Ausmaß, in dem Informationen in Bezug auf ihre Quelle oder ihren Inhalt hoch angesehen sind (vgl. Wang & Strong, 1996, S. 32). |
| Sicherheit (Safety) | Daten sind vor einem unautorisiertem Zugriff zu schützen, da sie für Unternehmen Wettbewerbsvorteile gegenüber der Konkurrenz darstellen können (vgl. Holthuis, 1999, S. 34). |
| Transportierbarkeit | Im Vergleich zu Daten, die übertragen werden können und damit an jedem Ort verfügbar sind, sind orts- oder system-bezogene Daten für Entscheider deutlich weniger nutzbringend (vgl. Holthuis, 1999, S. 34). |
| Umfang der Daten (Amount of data) | Inwieweit die Menge oder der Umfang der verfügbaren Daten angemessen ist (vgl. Wang & Strong, 1996, S. 32). |
| Verfügbarkeit (Availability) | Ausmaß, in dem Informationen physisch zugänglich sind (vgl. Knight & Burn, 2005, S. 162). |
| Verlässlichkeit (Reliability) | Ausmaß, in dem Informationen korrekt und zuverlässig sind (vgl. Wand & Wang, 1996, S. 93). |
| Verständlichkeit (Understandability) | Ausmaß, in dem die Daten klar, eindeutig und leicht verständlich sind (vgl. Pipino et al., 2002, S. 212). |
| Vollständigkeit (Completeness) | Alle Werte für eine bestimmte Variable wurden erfasst (vgl. D. P. Ballou & Pazer, 1985, S. 153).<br>Das Ausmaß, in dem Daten von ausreichender Breite, Tiefe und Umfang für die jeweilige Aufgabe sind (vgl. Wang & Strong, 1996, S. 32).<br>Die Fähigkeit eines Informationssystems, jeden sinnvollen Zustand des dargestellten Systems der realen Welt zu repräsentieren (vgl. Wand & Wang, 1996, S. 93). |
| Wertschöpfung (Value added) | Das Ausmaß, in dem Informationen nützlich sind, bietet Vorteile durch ihren Gebrauch (vgl. Wang & Strong, 1996, S. 31). |
| Wirksamkeit (Effectiveness) | Die Fähigkeit einer Funktion, den Benutzern zu ermöglichen, bestimmte Ziele mit Genauigkeit und Vollständigkeit |

| | |
|---|---|
| | in einem bestimmten Nutzungskontext zu erreichen (vgl. Batini et al., 2009, S. 32). |
| Wirkungsgrad (Efficiency) | Ausmaß, in dem die Daten in der Lage sind, den Informationsbedarf für die jeweilige Aufgabe schnell zu decken [15]. |
| Zugänglichkeit (accessibility) | Ausmaß, in dem Informationen verfügbar oder leicht und schnell abrufbar sind (vgl. Wang & Strong, 1996, S. 32). |
| Zugangssicherheit (access security) | Ausmaß, in dem der Zugang zu Informationen in angemessener Weise eingeschränkt wird, um ihre Sicherheit zu gewährleisten (vgl. Wang & Strong, 1996, S. 32). |