1 Datenqualität

Der Begriff Datenqualität ist sehr stark subjektiv geprägt. Sowohl bei der Befragung von Fachleuten als auch in der Literatur erhält man zu diesem Thema sehr unterschiedliche Antworten. Viele Autoren gehen in Ermangelung einer einheitlichen Definition daher auf die beiden Grundbestandteile des Begriffs zurück und definieren sowohl Daten als auch Qualität allgemein und folgen damit Larry English, einem der Pioniere auf dem Gebiet der Datenqualität: »The best way to look at information quality is to look at what quality means in the general marketplace and then translate what quality means for information« (vgl. [English 1999, S. 15ff.]).

In diesem Kapitel werden zunächst die grundlegenden Begriffe Daten und Qualität und daraus abgeleitet der Begriff Datenqualität erläutert. Nach einer ausführlichen Beschreibung der Eigenschaften wird auf unterschiedliche Taxonomien eingegangen. Den Abschluss des Kapitels bildet das Thema Datenqualitätsmanagement.

1.1 Daten

Die aktuelle Situation in den Unternehmen ist durch eine steigende Datenflut gekennzeichnet. Beispielsweise fallen durch die Vernetzung von Scannerkassen in Supermärkten oder die Speicherung von Verbindungsdaten in der Telekommunikationsbranche große Datenmengen an. Dieser Trend wird durch neue Entwicklungen wie Radio Frequency Identification (RFID) noch verstärkt. Nach Schätzungen der Gartner-Gruppe würde die Einzelhandelskette Wal-Mart täglich Daten im Umfang von 7 Terabyte generieren, wenn alle Artikel mit RFID-Marken versehen würden (vgl. [Raskino/Fenn/Linden 2005]). Gemäß einer IDC-Studie (vgl. [IDC 2011]) ist die weltweit produzierte Datenmenge im Jahr 2011 auf ein Volumen von 1,8 Zettabyte[1] angestiegen. Daten allein haben jedoch nur einen begrenzten Wert, erst in einem sinnvollen Kontext werden daraus unternehmensrelevante Informationen.

1. 1 Zettabyte = 1 Billion Gigabyte

Bisher gibt es keine einheitliche Definition des Begriffs Daten. Den meisten Definitionen ist jedoch gemein, dass sie Daten nicht getrennt, sondern im Zusammenhang mit Information und Wissen betrachten, weil sich die Begriffe jeweils ergänzen (vgl. [English 1999, S. 18; Helfert 2002, S. 13; Müller 2000, S. 5ff. u.a.]). Zumeist findet eine Hierarchisierung statt, deren unterstes Glied die Daten darstellen. Hierbei wird häufig die Semiotik als Strukturierungshilfe (Syntaktik – Semantik – Pragmatik) genutzt, die die allgemeine Lehre von den Zeichen, Zeichensystemen und Zeichenprozessen in das Gebiet der Informatik überträgt.

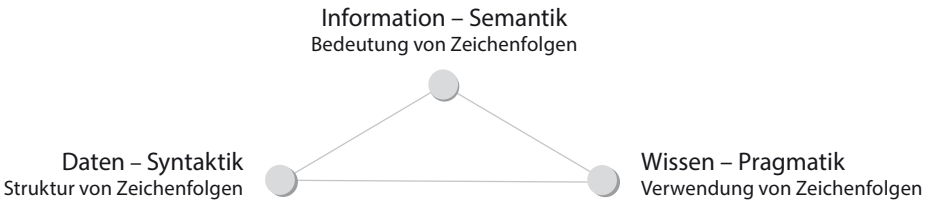

Abb. 1–1 *Semiotisches Dreieck (in Anlehnung an [Hinrichs 2002, S. 27])*

Auf syntaktischer Ebene werden lediglich die **Zeichen** sowie ihre mathematisch-statistischen Beziehungen untereinander (z.B. relative Häufigkeit innerhalb bestimmter Grundstrukturen) untersucht, ohne dabei auf die Bedeutung der Zeichen einzugehen. Diese maschinenlesbaren Zeichenfolgen (Daten) bilden somit die Informationen der realen Welt ab.

Wird den Daten Bedeutung hinzugefügt, gelangt man auf die semantische Ebene, d.h., die Daten werden in einem bestimmten Kontext gesehen, und man spricht von Information.

Auf der pragmatischen Ebene steht der direkte Benutzer (Interpreter) im Mittelpunkt der Untersuchungen, d.h., hier spielt die **Wirkung** von Information auf die sie verarbeitenden Verwender (Menschen, Maschinen) eine wichtige Rolle. Somit kommt die pragmatische Ebene der Wirklichkeit am nächsten, indem sie sich über die ersten zwei Ebenen hinausgehend noch mit Fragen der jeweiligen Absicht und des Werts für den einzelnen Benutzer befasst. Erst dann wird aus der Information Wissen.

Aus Gründen der besseren Lesbarkeit bezieht sich in den nachfolgenden Kapiteln dieses Buches der Begriff Datenqualität sowohl auf die Qualität der Daten als auch auf die Qualität der Informationen.

1.2 Qualität

Der Begriff Qualität stammt ab vom lateinischen »qualitas« und bedeutet Eigenschaft oder Beschaffenheit. Ursprünglich weder positiv noch negativ belegt, wird der Begriff in der Umgangssprache automatisch als positiv angesehen. Die Suche nach einer einheitlichen Definition führt zu einer Vielzahl von Definitions- und Interpretationsversuchen. Eine allgemein akzeptierte Begriffsbeschreibung ist die DIN-Norm 55 350. Danach ist die »Qualität die Gesamtheit von Eigenschaften und Merkmalen eines Produktes oder einer Tätigkeit, die sich auf deren Eignung zur Erfüllung festgelegter oder vorausgesetzter Erfordernisse beziehen« (vgl. [DIN 55350]).

Einer der ersten Systematisierungsansätze geht auf Garvin (vgl. [Garvin 1984, S. 40ff.]) zurück, der fünf generelle Qualitätsvorstellungen unterscheidet:

- Produktorientierter Ansatz
- Anwenderorientierter Ansatz
- Prozessorientierter Ansatz
- Wertbezogener Ansatz
- Transzendenter Ansatz

Die produktbezogene Sicht entspricht einem objektiven Qualitätsbegriff, weil Qualität als eine messbare, genau spezifizierbare Größe, die das Produkt beschreibt, gesehen wird. Qualität stellt dabei eine objektive Größe dar, die unabhängig von subjektiven Wahrnehmungen bestimmt werden kann, d.h., dieser Ansatz bezieht sich nur auf das Endprodukt, unabhängig von den Kunden (Benutzern). Qualitätsdifferenzen lassen sich damit auf die Unterschiede in den Produkteigenschaften zurückführen.

Der kunden- oder anwenderbezogene Ansatz hingegen definiert die Qualität eines Produkts über den Produktnutzer, und somit entscheidet ausschließlich der Kunde, inwieweit das Produkt der geforderten Qualität entspricht (subjektive Beurteilung des Kunden). In die amerikanische Literatur hat dieser Ansatz Eingang über die Definition »fitness for purpose« oder »fit for use« gefunden. Dabei können verschiedene Endbenutzer unterschiedliche Bedürfnisse haben, sodass die Qualität des gleichen Produkts unterschiedlich bewertet werden kann.

Beim Herstellungsbezug (prozessorientierter Ansatz) wird angenommen, dass Qualität dann entsteht, wenn der Herstellungsprozess optimal und kontrolliert verläuft und alle Vorgaben (Produktspezifikationen) eingehalten werden. Abweichungen von dem definierten Prozess werden als Qualitätsverlust angesehen.

Der wertbezogene Ansatz betrachtet Qualität unter Kostengesichtspunkten. Ein Produkt ist dann von hoher Qualität, wenn die Kosten und die empfangene Leistung in einem akzeptablen Verhältnis stehen.

Der transzendente Ansatz kennzeichnet Qualität als vorgegebene Vortrefflichkeit, Einzigartigkeit oder Superlativ. Qualität wird als Synonym für hohe Standards und Ansprüche angesehen. Dieser Grundgedanke setzt ein philosophi-

sches Verständnis voraus, das davon ausgeht, dass Qualität nicht messbar, sondern nur erfahrbar ist. Dieser Ansatz ist für den hier zu betrachtenden Kontext von Business Intelligence nicht geeignet.

Auch wenn die hier beschriebenen Ansätze für die Fertigungsindustrie entwickelt wurden, lassen sie sich ohne Weiteres auf den Bereich der Datenqualität übertragen, wie die folgenden Analogien zeigen (vgl. [Wang/Ziad/Lee 2001, S. 3f.]. Ein Datenverarbeitungsprozess kann auch als Herstellungsprozess im Sinne der Fertigungsindustrie gesehen werden. Die Datenquellen (Lieferanten), die die Rohdaten (Rohmaterialien) bereitstellen, bilden den Ausgangspunkt der Wertschöpfungskette. Sie werden im Zuge der Integration/Transformation (Produktionsprozess) bearbeitet. Das Ergebnis des Prozesses sind die Datenprodukte, die den Datenbeziehern (Kunden) zu Auswertungszwecken zur Verfügung gestellt werden.

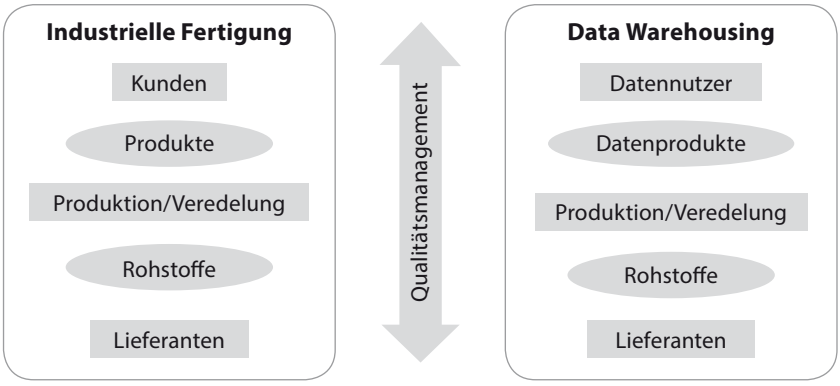

Abb. 1–2 *Analogie zwischen industrieller Fertigung und Datenverarbeitung (Data Warehousing) (in Anlehnung an [Grimmer/Hinrichs 2001, S. 72])*

Der wesentliche Unterschied liegt im Betrachtungsgegenstand sowie dessen Qualitätsmerkmalen. Im industriellen Fertigungsprozess werden physische Produkte erstellt, die Merkmale wie Haltbarkeit, Länge und Gewicht aufweisen. Im dargestellten Kontext der Datenverarbeitung entspricht das Produkt einem bestimmten Ausschnitt des Datenbestands, auch als Datenprodukt (gleichbedeutend mit einem Datensatz) bezeichnet. Zur Bestimmung der Qualität wird einem Produkt eine Menge von Merkmalen zugeordnet. Ein Merkmal ist dabei eine Eigenschaft, die zur Unterscheidung von Produkten in qualitativer oder quantitativer Hinsicht herangezogen werden kann (vgl. [Behme 2002, S. 52]).

Während in der Industrie der Qualitätsbegriff seit Jahrzehnten einen wichtigen Platz einnimmt, taucht der Begriff Datenqualität erst Mitte der 1990er-Jahre vermehrt auf. Die Vorgaben zu Datenqualität liegen damit in ihrer Entwicklung hinter den im Kontext der industriellen Fertigung entwickelten Standards hinsichtlich Qualität deutlich zurück.

1.3 Datenqualität

Es gilt nun, aus den obigen allgemeinen Daten- und Qualitätsdefinitionen den Begriff der Datenqualität abzuleiten. Helfert hat die in der Literatur vorhandenen Ansätze zur Definition von Datenqualität untersucht und einander gegenübergestellt (vgl. [Helfert 2002, S. 69ff.] und [Helfert 2000, S. 62ff.]). Das Ergebnis dieser Untersuchung zeigt, dass der Anwender das Qualitätsniveau festlegt und damit im Kontext der Datenverarbeitung ausschließlich der anwenderorientierte Ansatz (vgl. [Müller 2000, S. 15; English 1999, S. 52ff.]) sinnvoll ist. Datenqualität wird daher nach Würthele definiert 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« (vgl. [Würthele 2003, S. 21]).

Diese Definition macht deutlich, dass die Qualität von Daten vom Zeitpunkt der Betrachtung sowie von dem zu diesem Zeitpunkt an die Daten gestellten Anspruchsniveau abhängt.

Um die Datenqualität letztendlich messbar zu machen, bedarf es objektiver Merkmale (auch Qualitätskriterien genannt), die den Daten (Datenprodukten) zugeordnet werden. Diese werden dabei aufgrund der praktischen Erfahrungen intuitiv definiert, auf Basis von Literaturrecherchen erstellt oder anhand von empirischen Untersuchungen zusammengestellt (vgl. [Helfert 2002, S. 69]). Die Qualitätskriterien müssen messbar sein, damit der jeweilige Erfüllungsgrad durch den Datennutzer ermittelt werden kann. In der Praxis wird es einen hundertprozentigen Erfüllungsgrad der Kriterien nicht geben, vielmehr sind jeweils anwendungs- oder kundenbezogene Anspruchsniveaus (Sollwerte) zu definieren, an denen die Datenqualität gemessen wird.

Beispielsweise gelten für Quartals- oder Jahresbilanzen im Bankenbereich, die kurzfristig nach Ablauf des jeweiligen Zeitraums an die Aufsichtsbehörden übermittelt werden, sehr hohe Ansprüche an die Genauigkeit und Aktualität. Dagegen sind bei Auswertungen zum Kundenverhalten geringere Anspruchsniveaus akzeptabel.

Tabelle 1–1 zeigt eine Übersicht über häufig genannte Datenqualitätskriterien (DQ-Kriterien) in alphabetischer Reihenfolge (in Anlehnung an [Helfert/Herrmann/Strauch 2001, S. 7]).

■ Aktualität	■ Integrität	■ Robustheit
■ Allgemeingültigkeit	■ Informationsgrad	■ Seltenheit
■ Alter	■ Klarheit	■ Sicherheit
■ Änderungshäufigkeit	■ Kompaktheit	■ Signifikanz
■ Aufbereitungsgrad	■ Konsistenz	■ Testbarkeit
■ Bedeutung	■ Konstanz	■ Unabhängigkeit
■ Benutzbarkeit	■ Korrektheit	■ Überprüfbarkeit
■ Bestätigungsgrad	■ Neutralität	■ Verdichtungsgrad
■ Bestimmtheit	■ Objektivität	■ Verfügbarkeit
■ Detailliertheit	■ Operationalität	■ Verlässlichkeit
■ Effizienz	■ Performanz	■ Verschlüsselungsgrad
■ Eindeutigkeit	■ Portabilität	■ Verständlichkeit
■ Fehlerfreiheit	■ Präzision	■ Vollständigkeit
■ Flexibilität	■ Problemadäquatheit	■ Wahrheitsgehalt
■ Ganzheit	■ Prognosegehalt	■ Wiederverwendbarkeit
■ Geltungsdauer	■ Quantifizierbarkeit	■ Wirkungsdauer
■ Genauigkeit	■ Rechtzeitigkeit	■ Zeitbezug
■ Glaubwürdigkeit	■ Redundanzfreiheit	■ Zeitnähe
■ Gültigkeit	■ Referenzielle Integrität	■ Zugänglichkeit
■ Handhabbarkeit	■ Relevanz	■ Zuverlässigkeit

Tab. 1–1 *Liste möglicher Datenqualitätskriterien*

Im Folgenden wird lediglich auf eine Auswahl der vorgestellten Qualitätskriterien näher eingegangen, da die Liste zum Teil Doppelungen enthält sowie nicht alle Kriterien als besonders geeignet erscheinen (vgl. [Hinrichs 2002, S. 30f.; Zeh 2009, S. 43f.]):

Datenqualitätskriterien	Definition
Korrektheit Fehlerfreiheit	Die Attributwerte eines Datensatzes (im Data Warehouse) entsprechen denen der modellierten Entitäten der realen Welt, d.h., die Daten stimmen mit der Realität überein.
Konsistenz	Die Attributwerte eines Datensatzes weisen keine logischen Widersprüche untereinander oder zu anderen Datensätzen auf. Inkonsistente Daten innerhalb der operativen Systeme führen zu massiven Glaubwürdigkeitsproblemen in den analytischen Systemen.
Zuverlässigkeit Nachvollziehbarkeit	Die Attributwerte sind vertrauenswürdig, d.h., die Entstehung der Daten ist nachvollziehbar. Insbesondere bei externen Daten ist auf die Zuverlässigkeit der Quellen zu achten. Aber auch innerhalb des Data Warehouse müssen die verschiedenen Transformationen der Daten nachvollziehbar sein. Dies beginnt bei der Erfassung der Daten und geht bis zur Erstellung der Berichte in den analytischen Systemen.

→

Datenqualitätskriterien	Definition
Vollständigkeit	Die Attributwerte eines Datensatzes sind mit Werten belegt, die semantisch vom Wert NULL (unbekannt) abweichen. Eine andere Definition bezieht sich auf den modellierten Ausschnitt der Welt. Alle wichtigen Entitäten, Beziehungen und Attribute müssen im System repräsentiert sein. Vollständigkeit beschreibt auch die generelle Verfügbarkeit von Inhalten, die der Anwender benötigt, um seine Arbeit überhaupt durchführen zu können. Dies behandelt die Frage, ob beispielsweise alle Datenbereiche in den Business-Intelligence-Systemen integriert sind, um die Anforderungen zu erfüllen. Des Weiteren beschreibt dieses Kriterium auch, ob die Daten komplett im ELT-Prozess oder im Fehlerfall in das Data Warehouse übernommen werden. Besonders schwierig ist dies beispielsweise bei tagesaktuellen Lieferungen aus verschiedenen Zeitzonen.
Genauigkeit	Abhängig vom jeweiligen Kontext liegen die Daten in der geforderten Genauigkeit (z. B. Anzahl Nachkommastellen) vor.
Aktualität **Zeitnähe** **Zeitbezug**	Alle Datensätze entsprechen jeweils dem aktuellen Zustand der modellierten Welt und sind damit nicht veraltet. Die Daten bilden die tatsächlichen Eigenschaften des Objekts zeitnah ab. Mangelnde Aktualität kann einerseits aus der Frequenz der Ladezyklen resultieren (z. B. wöchentlich statt täglich) oder durch die verspätete Pflege der Daten bereits im operativen System (z. B. keine regelmäßige Neubewertung von Sicherheiten).
Redundanzfreiheit	Innerhalb der Datensätze dürfen keine Duplikate vorkommen. Als Duplikate werden hierbei Datensätze verstanden, die dieselbe Entität in der realen Welt beschreiben. Sie müssen aber nicht notwendigerweise in allen Attributwerten übereinstimmen.
Relevanz	Der Informationsgehalt einer Datensatzmenge bezüglich eines definierten Anwendungskontextes deckt sich mit dem Informationsbedarf einer Anfrage.
Einheitlichkeit	Die Repräsentationsstruktur einer Menge von Datensätzen ist einheitlich, d. h., sie werden fortlaufend gleich abgebildet.
Eindeutigkeit	Ein Datensatz muss eindeutig interpretierbar sein, d. h., die vorhandenen Metadaten müssen die Semantik des Datensatzes festschreiben.
Verständlichkeit	Die Datensätze stimmen in ihrer Begrifflichkeit und Struktur mit den Vorstellungen des Fachbereichs überein.
Schlüsseleindeutigkeit	Die Primärschlüssel der Datensätze sind eindeutig.
Referenzielle Integrität	Im relationalen Modell muss jeder Fremdschlüssel eindeutig auf einen existierenden Primärschlüssel referenzieren.

Tab. 1–2 *Definition ausgewählter Datenqualitätskriterien*

Die beiden letzten Kriterien stellen eine spezielle Ausrichtung auf das relationale Datenbankmodell dar. Aufgrund der sehr starken Verbreitung des relationalen Modells ist diese Sichtweise legitim.

Die sechs DQ-Kriterien Korrektheit, Konsistenz, Zuverlässigkeit, Vollständigkeit, Zeitnähe und Relevanz werden in Abschnitt 2.3 nochmals aufgegriffen und im Kontext Business Intelligence näher betrachtet.

Das folgende Beispiel (in Anlehnung an [Leser/Naumann 2007, S. 354f.]) aus dem BI-Umfeld verdeutlicht die Relevanz der DQ-Kriterien **Vollständigkeit**, **Zeitnähe** und **Glaubwürdigkeit**. Als Entscheidungsgrundlage für das Management eines Industrieunternehmens werden regelmäßig aus einem Data Warehouse Berichte erstellt:

- Diese Berichte müssen Daten aus allen Werken vollständig abdecken, sonst sind die Produktionszahlen ungenau.
- Die Berichte müssen zeitnah abrufbar sein, sonst kann nicht schnell genug bei einer veränderten Absatzlage reagiert werden.
- Wenn die Zahlen in den Berichten nicht stimmen, weil in der Vergangenheit nachträglich viele Daten manuell geändert wurden, sind die Kennzahlen unglaubwürdig, und die Akzeptanz der BI-Lösung sinkt.

Dieses Beispiel zeigt deutlich, dass Datenqualität stets mehrdimensional zu betrachten ist. Wird die Datenqualität auf ein einzelnes Kriterium (wie beispielsweise Vollständigkeit) reduziert, wird die Datenqualität von den Anwendern dennoch gefühlt als schlecht wahrgenommen, wenn veraltete Daten vorliegen (DQ-Kriterium Zeitnähe).

Werden die hier vorgestellten DQ-Kriterien strukturiert in Gruppen zusammengefasst, spricht man von einem Qualitätsmodell. Ein wesentliches Charakteristikum eines solchen Modells ist die Zerlegungssystematik. In der Literatur sind diverse Systematiken zu finden (vgl. [Wang/Strong 1996, S. 20; Redman 1996, S. 267]), die bei genauerer Betrachtung gewisse Unstimmigkeiten bezüglich der Zerlegung aufweisen. Ziel dieses Kapitels ist es jedoch nicht, diese Lücke durch ein eigenes Modell zu schließen. Daher sei an dieser Stelle beispielhaft zunächst das Qualitätsmodell von Hinrichs vorgestellt, das sich aus den beschriebenen Qualitätskriterien ableiten lässt:

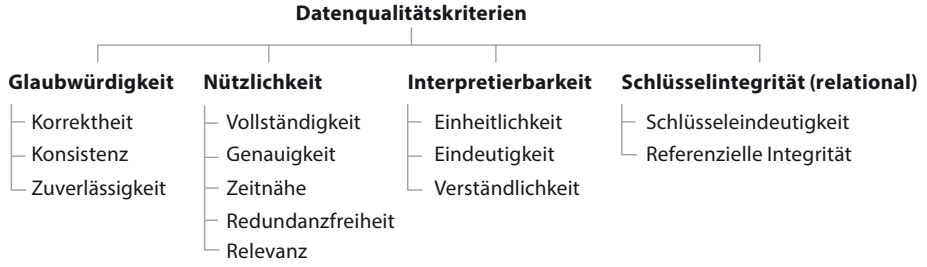

Abb. 1–3 *Taxonomie von Datenqualitätskriterien (vgl. [Hinrichs 2002, S. 30])*

Diesem eher aus theoretischer Sicht entstandenen Qualitätsmodell stellt die Deutsche Gesellschaft für Informations- und Datenqualität (DGIQ) eine Kategorisierung gegenüber, die aus einer Studie (vgl. [Wang/Strong 1996]) durch Befragung von IT-Anwendern hervorgegangen ist (siehe Abb. 1–4).