

Edition TDWI

# Datenqualität erfolgreich steuern

Praxislösungen für Business-Intelligence-Projekte

von

Detlef Apel, Wolfgang Behme, Rüdiger Eberlein, Christian Merighi

3., überarbeitete und erweiterte Auflage

dpunkt.verlag GmbH Heidelberg 2015

Verlag C.H. Beck im Internet:

[www.beck.de](http://www.beck.de)

ISBN 978 3 86490 042 6

Zu [Inhaltsverzeichnis](#)

schnell und portofrei erhältlich bei [beck-shop.de](http://beck-shop.de) DIE FACHBUCHHANDLUNG

---

# 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.

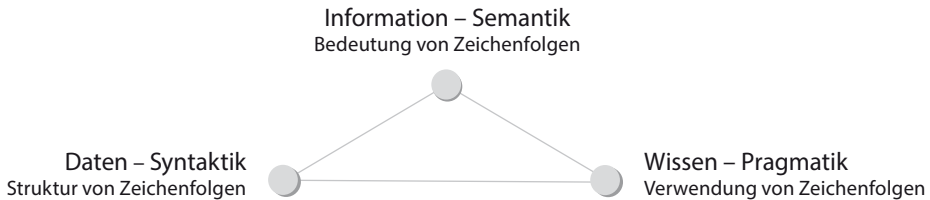
## 1.1 Daten

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<sup>1</sup> 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.

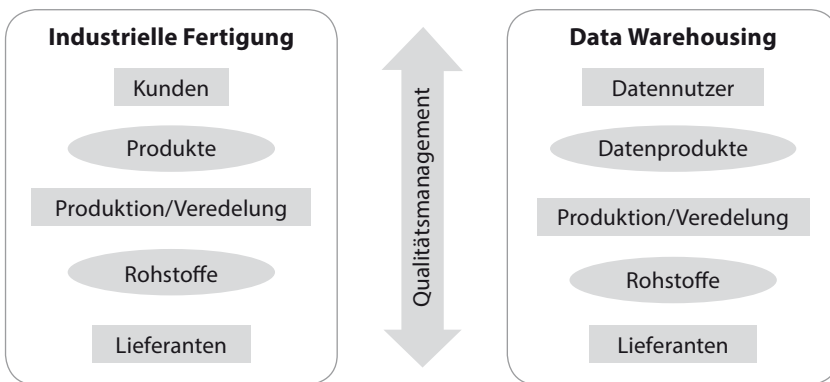
Beim Herstellungsbezug (prozessorientierter Ansatz) wird angenommen, dass Qualität dann entsteht, wenn der Herstellungsprozess optimal und kontrolliert verläuft und alle Vorgaben (Produktspezifikationen) eingehalten werden. Abweichungen von dem definierten Prozess werden als Qualitätsverlust angesehen.

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]).

<ul style="list-style-type: none"> <li>■ Aktualität</li> <li>■ Allgemeingültigkeit</li> <li>■ Alter</li> <li>■ Änderungshäufigkeit</li> <li>■ Aufbereitungsgrad</li> <li>■ Bedeutung</li> <li>■ Benutzbarkeit</li> <li>■ Bestätigungsgrad</li> <li>■ Bestimmtheit</li> <li>■ Detailliertheit</li> <li>■ Effizienz</li> <li>■ Eindeutigkeit</li> <li>■ Fehlerfreiheit</li> <li>■ Flexibilität</li> <li>■ Ganzheit</li> <li>■ Geltungsdauer</li> <li>■ Genauigkeit</li> <li>■ Glaubwürdigkeit</li> <li>■ Gültigkeit</li> <li>■ Handhabbarkeit</li> </ul>	<ul style="list-style-type: none"> <li>■ Integrität</li> <li>■ Informationsgrad</li> <li>■ Klarheit</li> <li>■ Kompaktheit</li> <li>■ Konsistenz</li> <li>■ Konstanz</li> <li>■ Korrektheit</li> <li>■ Neutralität</li> <li>■ Objektivität</li> <li>■ Operationalität</li> <li>■ Performanz</li> <li>■ Portabilität</li> <li>■ Präzision</li> <li>■ Problemadäquatheit</li> <li>■ Prognosegehalt</li> <li>■ Quantifizierbarkeit</li> <li>■ Rechtzeitigkeit</li> <li>■ Redundanzfreiheit</li> <li>■ Referenzielle Integrität</li> <li>■ Relevanz</li> </ul>	<ul style="list-style-type: none"> <li>■ Robustheit</li> <li>■ Seltenheit</li> <li>■ Sicherheit</li> <li>■ Signifikanz</li> <li>■ Testbarkeit</li> <li>■ Unabhängigkeit</li> <li>■ Überprüfbarkeit</li> <li>■ Verdichtungsgrad</li> <li>■ Verfügbarkeit</li> <li>■ Verlässlichkeit</li> <li>■ Verschlüsselungsgrad</li> <li>■ Verständlichkeit</li> <li>■ Vollständigkeit</li> <li>■ Wahrheitsgehalt</li> <li>■ Wiederverwendbarkeit</li> <li>■ Wirkungsdauer</li> <li>■ Zeitbezug</li> <li>■ Zeitnähe</li> <li>■ Zugänglichkeit</li> <li>■ Zuverlässigkeit</li> </ul>
--	--	--

**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
<b>Korrektheit Fehlerfreiheit</b>	Die Attributwerte eines Datensatzes (im Data Warehouse) entsprechen denen der modellierten Entitäten der realen Welt, d.h., die Daten stimmen mit der Realität überein.
<b>Konsistenz</b>	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.
<b>Zuverlässigkeit Nachvollziehbarkeit</b>	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
<b>Vollständigkeit</b>	<p>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.</p> <p>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.</p> <p>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.</p>
<b>Genauigkeit</b>	Abhängig vom jeweiligen Kontext liegen die Daten in der geforderten Genauigkeit (z. B. Anzahl Nachkommastellen) vor.
<b>Aktualität Zeitnähe Zeitbezug</b>	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).
<b>Redundanzfreiheit</b>	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.
<b>Relevanz</b>	Der Informationsgehalt einer Datensatzmenge bezüglich eines definierten Anwendungskontextes deckt sich mit dem Informationsbedarf einer Anfrage.
<b>Einheitlichkeit</b>	Die Repräsentationsstruktur einer Menge von Datensätzen ist einheitlich, d. h., sie werden fortlaufend gleich abgebildet.
<b>Eindeutigkeit</b>	Ein Datensatz muss eindeutig interpretierbar sein, d. h., die vorhandenen Metadaten müssen die Semantik des Datensatzes festschreiben.
<b>Verständlichkeit</b>	Die Datensätze stimmen in ihrer Begrifflichkeit und Struktur mit den Vorstellungen des Fachbereichs überein.
<b>Schlüsseleindeutigkeit</b>	Die Primärschlüssel der Datensätze sind eindeutig.
<b>Referenzielle Integrität</b>	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.

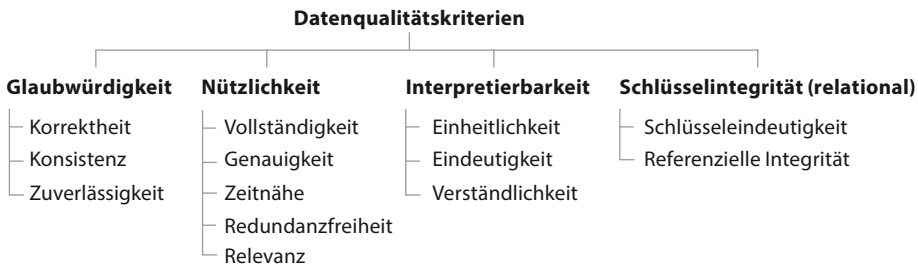


Das folgende Beispiel (in Anlehnung an [Leser/Naumann 2007, S. 354f.]) aus dem BI-Umfeld verdeutlicht die Relevanz der DQ-Kriterien **Vollständigkeit**, **Zeitnähe** und **Glaubwürdigkeit**. Als Entscheidungsgrundlage für das Management eines Industrieunternehmens werden regelmäßig aus einem Data Warehouse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

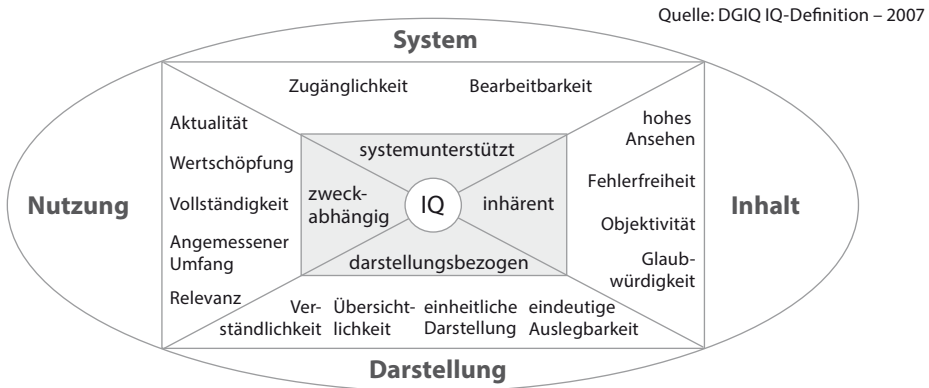
Dieses Beispiel zeigt deutlich, dass Datenqualität stets mehrdimensional zu betrachten ist. Wird die Datenqualität auf ein einzelnes Kriterium (wie beispielsweise Vollständigkeit) reduziert, wird die Datenqualität von den Anwendern dennoch gefühlt als schlecht wahrgenommen, wenn veraltete Daten vorliegen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



**Abb. 1–4** Taxonomie von Datenqualitätskriterien (vgl. [DGIQ 2007])

Ergänzend zu den bereits beschriebenen Kriterien sind vor allem die Zugänglichkeit und die Bearbeitbarkeit hinzugekommen. Unter Zugänglichkeit wird die einfache Abrufbarkeit der Daten für den Anwender verstanden. Inwieweit die Daten leicht für unterschiedliche Zwecke zu bearbeiten sind, wird mit dem Kriterium Bearbeitbarkeit ausgedrückt.<sup>2</sup>

Die Identifikation und Klassifikation von Datenqualitätskriterien allein reicht für die Messung der Datenqualität allerdings nicht aus. Was fehlt, sind konkrete, numerische Metriken. Nur darüber kann später geprüft werden, ob die Verbesserungsmaßnahmen auch wirkungsvoll waren (»You cannot control what you cannot measure« (vgl. [deMarco 1982])).

Die Anwendung geeigneter Metriken ermöglicht eine Quantifizierung von Datenqualitätskriterien und ist somit die Voraussetzung zur Bildung von Qualitätskennzahlen. In Kapitel 7 wird genauer auf die Bildung dieser Kennzahlen auf Basis ausgewählter DQ-Kriterien eingegangen.

## 1.4 Datenqualitätsmanagement

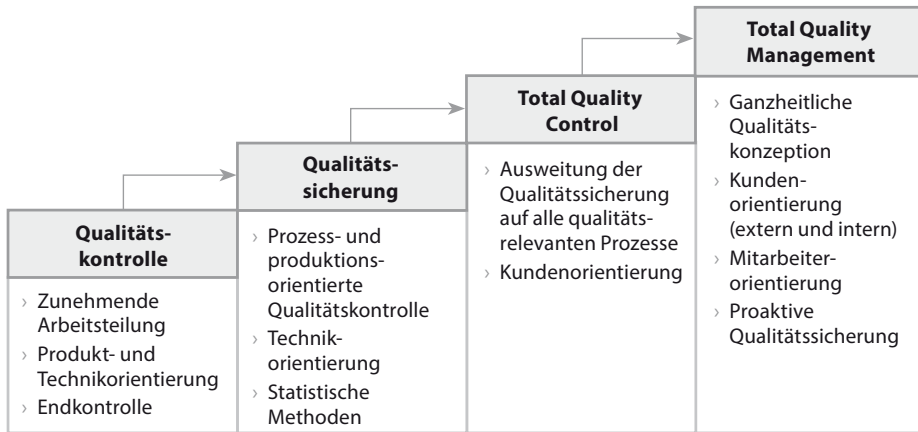
Das nachträgliche Bereinigen von Daten, das durch eine Vielzahl an existierenden Werkzeugen zur Fehlererkennung und -korrektur erleichtert wird, ist im Vergleich zu qualitätssichernden Maßnahmen um den Faktor 5–10 teurer (vgl. [Hankins 1999]). Trotzdem finden in den Unternehmen kaum präventive Maßnahmen statt, sondern es wird erst beim Auftreten von Problemen reagiert (vgl. [Otto et al. 2008, S. 215f.]).

Dieses reaktive Vorgehen führt u.a. dazu, dass Risiken nicht rechtzeitig erkannt werden oder gesetzliche Auflagen nicht zu erfüllen sind (siehe Abschnitt 3.2). Erst langsam kommt es in den Unternehmen zu einem Sinneswandel und

2. Eine ausführliche Beschreibung der einzelnen Kriterien findet sich in [Rohwedder et al. 2011, S. 25ff.].

somit zu einem proaktiven Ansatz mit einem Datenqualitätsmanagement, das von vornherein auf qualitativ hochwertige Daten setzt und kostenintensive nachträgliche Bereinigungen minimiert.

Das dazu erforderliche Qualitätsmanagement umfasst nach DIN ISO 8402 »alle Tätigkeiten der Gesamtführungsaufgabe, die die Qualitätspolitik, -ziele und -verantwortung festlegen sowie durch Mittel wie Qualitätsplanung, -lenkung, -sicherung und -verbesserung im Rahmen des Qualitätsmanagementsystems verwirklichen« (vgl. [DIN ISO 8402]). Hieraus wird deutlich, dass das Qualitätsmanagement in der Gesamtstrategie des Unternehmens verankert sein muss. Diese aus heutiger Sicht sinnvolle Definition entwickelte sich in der Historie seit Beginn des 20. Jahrhunderts gemäß Abbildung 1–5 in vier Stufen:

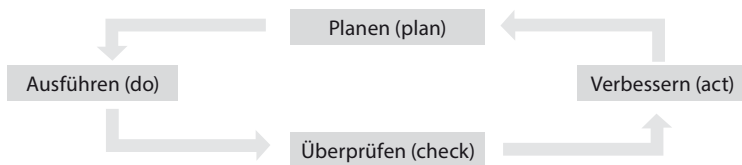


**Abb. 1–5**      *Entwicklungsstufen des Qualitätswesens (in Anlehnung an [Wolf 1999, S. 63])*

In der ersten Stufe (Qualitätskontrolle) wurde eine klare Trennung zwischen der Produktion und der Qualitätskontrolle vorgenommen, d.h., festgestellte Mängel konnten erst nachträglich am bereits fertigen Produkt erkannt und bereinigt werden. Erst in den 1930er-Jahren wurde die Qualitätskontrolle in den Produktionsprozess integriert. Damit war es möglich, die Fehler während des Prozesses am Entstehungsort zu beheben (Qualitätssicherung). In den 1960er-Jahren setzte sich zunehmend die Erkenntnis durch, dass sich durch die prozessbegleitenden Maßnahmen nicht alle Fehlerquellen abstellen lassen. Daher wurde das Qualitätswesen sowohl auf die vorgelagerten Bereiche wie Forschung & Entwicklung oder Konstruktion als auch auf die nachgelagerten Bereiche wie den Vertrieb ausgedehnt. Geprägt wurde diese Stufe durch Feigenbaum (vgl. [Feigenbaum 1961]), der den Begriff »Total Quality Control« einführte. Die ständige Weiterentwicklung der Konzepte führte zu dem heute bekannten, ganzheitlichen Qualitätsmanagement (Total Quality Management), das in der Gesamtstrategie des Unternehmens integriert sein muss.

Für den Aufbau eines Qualitätsmanagements (QM) sind vor allem die Bereiche Qualitätsplanung und -lenkung von Interesse (vgl. [English 1999, S. 70ff.]). Aufgabe der Qualitätsplanung ist es, die Qualitätsanforderungen an den Prozess und das Produkt in überprüfbarer Form festzulegen. Dies beinhaltet die Auswahl von Qualitätskriterien sowie die Festlegung von Sollwerten (Anspruchsniveaus) für diese Kriterien. Die Qualitätslenkung, zu der Arbeitstechniken und Tätigkeiten gehören, die zur Erfüllung der Qualitätsanforderungen angewendet werden, setzt die Qualitätsplanung um. Dazu müssen geeignete Prozesse identifiziert und Maßnahmen zum Erreichen einer Prozesskonformität ergriffen werden. Produkt- und Prozessqualität müssen im Rahmen der Qualitätslenkung gemessen werden (vgl. [Helfert 2002, S. 40ff.]). Die Qualitätssicherung umfasst vor allem organisatorische Maßnahmen, die nach außen sicherstellen sollen, dass im Unternehmen ein Qualitätsmanagement existiert.

Die Ausführungen zum Qualitätsmanagement lassen sich auf ein Datenqualitätsmanagement (DQM) für den Datenverarbeitungsprozess übertragen. Besonders erwähnenswert sind in diesem Zusammenhang die Arbeiten von Wang (vgl. [Wang 1998; Wang/Strong 1996]) am Massachusetts Institute of Technology (MIT), der eine Adaption der QM-Konzepte unter der Bezeichnung Total Data Quality Management (TDQM) entwickelte. Der Grundgedanke seiner Methode ist der sogenannte Plan-Do-Check-Act-Zyklus, der die Ideen von Deming als Regelkreis beschreibt (siehe Abb. 1–6).



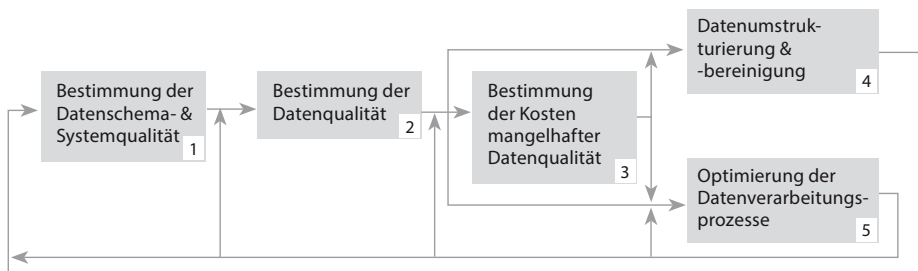
**Abb. 1–6** Plan-Do-Check-Act-Zyklus (vgl. [Redman 1996])

	Redman (1996)	English (1999)
<b>Name der Strategie</b>	Data Quality Policy/Program, Managing Information Chains	Total Quality data Management, Information Quality Environment: the 14 Points of Information Quality
<b>Grundgedanke</b>	Das Managen der Informationskette als Zyklus mit einem Prozessverantwortlichen. Dabei werden die Kundenbedürfnisse durch ein Datenqualitätsprogramm und feststehende Grundsätze für Datenqualität berücksichtigt.	Die Etablierung einer Unternehmenskultur der Informationsqualität als Basis für Datenqualitätsprozessverbesserungen. Die Kultur ist angelehnt an die Theorien des Total Quality data Management und Demings 14-Punkte-Programm für Change Management.
<b>Einzelne Methoden</b>	Error detection and correction, Statistical Quality Control, Measurement-System	Statt einzelner Methoden werden unterschiedliche Prozessabläufe vorgestellt, deren Aktivitäten man schrittweise folgen kann.
<b>Praxisbeispiel</b>	AT & T, Telstra Corporation Limited	Kein Beispiel, an dem die Gesamtstrategie durchgespielt wird, es werden nur einzelne Phasen oder Schritte beispielhaft erläutert

**Tab. 1-3** Vergleich ausgewählter DQM-Strategien

Der Zyklus beginnt mit der Definition der Datenqualitätsziele (Plan). Anschließend wird der Istzustand der Daten aufgenommen und analysiert (Do, Check). Im letzten Schritt muss durch den Einsatz geeigneter Methoden die Datenqualität verbessert werden (Act).

Das Konzept des TDQM wurde u. a. von English (vgl. [English 1999, S. 69ff.]) weiterentwickelt, der eine Vorgehensweise zur kontinuierlichen Datenqualitätsverbesserung einführte (siehe Abb. 1-7).



**Abb. 1-7** Total-Quality-data-Management-Methodik (TQdM) nach English (vgl. [English 1999, S. 70])

Hinrichs (2002)	Helfert (2002)	Olson (2003)
CLIQ – Data Cleansing mit intelligentem Qualitätsmanagement	Proaktives Datenqualitätsmanagement	Data Quality Assurance Program
Vorgehensmodell für den Datenintegrationsprozess im DWH mit fester Reihenfolge der Integrationsphasen. In Anlehnung an die ISO 9001:2000 werden Datenqualitätsmaßnahmen eingebunden.	Die operative Ebene der Qualitätsplanung und -lenkung wird in die Gesamtführungsaufgabe integriert, die Datenqualitätspolitik, -ziele und Verantwortung für Datenqualität festlegt.	Programm aus drei zu kombinierenden Komponenten: Dimension der Datenqualität, Methodik und Aktivitäten. Fokus sind jedoch nicht alle Dimensionen von Datenqualität, sondern die Genauigkeit der Daten.
Sämtliche Methoden werden innerhalb des Integrationsprozesses vorgestellt, wie z. B. statistische Prozesskontrollen regelbasierte Prüfungen, Duplikatbereinigung.	Forderung der Bereitstellung von Methoden und Maßnahmen für die Ausführung der Datenqualitätsprozesse. Statistische Methoden, Data Mining zur Musterbeschreibung.	Data-Profiling-Methoden zur Generierung korrekter Metadaten, zur Aufdeckung von ungenauen oder falschen Daten: beispielsweise Analyse der Spalteneigenschaften, Strukturanalyse
Data Quality Mining (DaimlerChrysler AG)	Beschreibung des Konzepts eines operativen Datenqualitätsmanagements für Crédit Suisse Financial Services	

Neben der bereits erwähnten Erweiterung gibt es in der wissenschaftlichen Literatur eine Vielzahl von Arbeiten, die sich mit dem Thema Datenqualitätsmanagement, basierend auf den Gedanken von Wang, auseinandersetzen. Die wichtigsten sind in Tabelle 1–3 aufgeführt (vgl. [Behme/Nietzschmann 2005, S. 46]).

Bei der Betrachtung der einzelnen Strategien zeigt sich, dass entweder ein eher technisch orientierter Ansatz gewählt wurde oder der Fokus auf dem Managementkonzept liegt. Einig ist man sich aber, dass die technisch orientierten Ansätze nicht ohne ein begleitendes Managementkonzept möglich sind.

Für die Umsetzung eines Datenqualitätsmanagements für ein Data-Warehouse-System lassen sich drei Bereiche identifizieren (vgl. [Helfert/Herrmann/Strauch 2001, S. 19]):

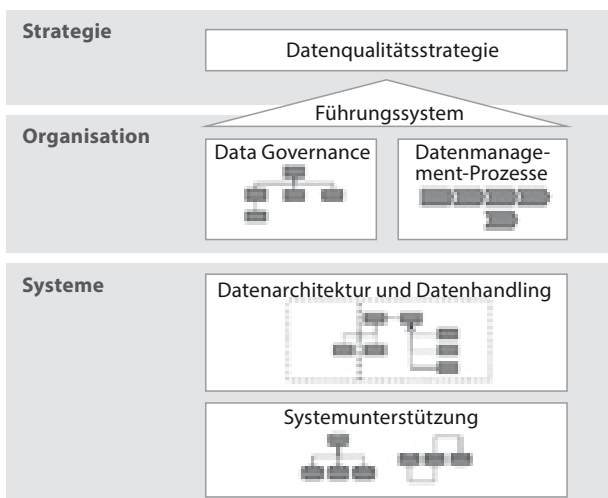
- Datenqualität als Unternehmenskultur und -philosophie ist als Verpflichtung des Managements anzusehen.
- Ein Qualitätsmanagementsystem (nach ISO 8402) ist zu etablieren, das in allen Bereichen geeignete Prozesse, Richtlinien, Pläne sowie Test- und Prüfverfahren aufsetzt, um die geforderte Datenqualität dauerhaft zu erreichen.

- Zur Ausübung der Qualitätsprozesse sind geeignete Methoden, Verfahren und Werkzeuge zur Verfügung zu stellen.

Auf Basis der bestehenden Datenqualitätsmanagement-Ansätze haben Otto u. a. einen Ordnungsrahmen entwickelt, der dem Ansatz des Business Engineering folgt (vgl. [Otto et al. 2008, S. 215ff.]). Dieser Ordnungsrahmen setzt sich aus den folgenden sechs Gestaltungselementen zusammen:

- Datenqualitätsstrategie
- Führungssystem
- Data Governance
- Datenmanagement-Prozesse
- Datenarchitektur und Datenhaltung
- Systemunterstützung

Den Aufbau des Ordnungsrahmens zeigt Abbildung 1–8. Die Datenqualitätsstrategie befindet sich auf der Ebene »Strategie«, wodurch die enge Verzahnung mit der Unternehmensstrategie deutlich wird. Zwischen den Ebenen »Strategie« und »Organisation« liegt das Führungssystem, das die Umsetzung der Strategie steuert. Die Ebene »Organisation« enthält zum einen die Data Governance (Zuordnung der Aufgaben und Verantwortlichkeiten im Rahmen des DQM) sowie die Datenmanagement-Prozesse (umfasst die Prozesse Anlage, Pflege und Ausphasen der Daten). Auf der Ebene »Systeme« wird im Rahmen der Datenarchitektur der Geltungsbereich für einzelne Datenobjekte definiert (z.B. globale versus lokale Lieferantenummer). Außerdem legt die Datenarchitektur fest, welche Rolle die Systeme bei der Anlage, Änderung und Verteilung der Stammdaten spielen.



**Abb. 1–8** Ordnungsrahmen für das Datenqualitätsmanagement (vgl. [Otto et al. 2008, S. 218])

## 1.5 Zusammenfassung

Die Ausführungen in diesem Kapitel machen deutlich, dass die Anforderungen an die Datenqualität jeweils von dem Unternehmen und insbesondere von den Anwendern der Daten abhängen. Eine allgemein gültige Festlegung, was eine gute Datenqualität im Rahmen von Business-Intelligence-Lösungen ausmacht, gibt es nicht.

Um Datenqualität aber nicht nur als abstraktes Gebilde stehen zu lassen, wurde mithilfe der Datenqualitätskriterien eine Operationalisierung vorgenommen. Die Aufgabe im konkreten Projekt besteht darin, eine Auswahl dieser Kriterien zu treffen und jeweils auf das Unternehmen und den Projekthintergrund passende Anspruchsniveaus festzulegen und diese zu realisieren.

Insgesamt wird auch deutlich, dass Datenqualität keine Einmal-Aktion im Rahmen eines Projekts ist, sondern vielmehr als eine permanente Aufgabe angesehen werden muss. Dazu ist bei den Mitarbeitern ein Bewusstsein für Datenqualität zu schaffen sowie die Motivation, diese auch herzustellen. Nur so kann langfristig und nachhaltig eine bessere Datenqualität sichergestellt werden.