# *Datenqualität*

**Felix Naumann**

**Daten von niedriger Qualität sind in kommerziellen und wissenschaftlichen Datenbanken allgegenwärtig.**

Produktcodes werden falsch verwendet, Messreihen werden in unterschiedlichen Einheiten erfasst, Kundendaten werden in Call-Centers fehlerhaft eingetippt etc. Solche Datenfehler und Probleme mindern die Datenqualität und haben wirtschaftliche Konsequenzen: Es gilt das sogenannte *garbage-in-garbage-out* Prinzip. Fehler in den Daten verursachen Fehler in daraus generierten Berichten; mangelndes Vertrauen in Daten führt zu Fehlentscheidungen; Chancen werden verpasst wenn Daten verspätet oder unverständlich sind.

## Motivation

Das Customer-Relationship Management (CRM) ist eines der primären Betätigungsfelder für die Datenreinigung, denn elektronisch erfasste Kundendaten haben viele Fehlerquellen und Fehler in Kundendaten können vielfältige negative Wirkungen haben. Bei der Erfassung von Kundendaten via Telefon entstehen u.a. Tippfehler, Verständnisfehler und fehlende Werte, da Kunden oft nicht bereit sind, alle Daten anzugeben oder bewusst falsche Daten nennen. Zudem entstehen leicht Dubletten (auch "Duplikate"), z.B. wenn Kundendaten mehrfach oder über verschiedene Kanäle (Telefon, www-Formular, Brief) mit dem Unternehmen in Kontakt treten. Auf der anderen Seite verursachen fehlerhafte Daten Kosten. Als einfaches Beispiel dient das Nichterreichen eines Kunden oder das doppelte Versenden eines aufwändigen Katalogs. Schwerer wiegen nicht erkannte Gefahren, etwa wenn ein Kunde unter verschiedenen Identitäten übermäßig hohe Warenkredite erzielt. Und nicht zuletzt schmälern falsche Kontaktangaben das Image des Unternehmens in den Augen des Kunden. Aber auch in anderen Anwendungsgebieten ist eine hohe Datenqualität wichtig oder sogar unabdingbar.

In den folgenden Abschnitten versuchen wir zunächst eine *Definition* der Datenqualität, anschließend betrachten wir Methoden zu *Messung* verschiedener Merkmale der Datenqualität und zuletzt besprechen wir Methoden zur deren *Verbesserung*.

## Datenqualität definieren

Die Qualität von Daten, auch "Informationsqualität", wird oft als die Eignung der Daten für die jeweilige datenverarbeitende Anwendung definiert. Daten von schlechter Qualität enthalten Datenfehler, Dubletten, fehlende Werte, falsche Formatierungen, Widersprüche usw. Rahm und Do geben eine Klassifikation von Datenfehlern, in der sie unterscheiden ob der Fehler auf Schemaebene oder auf Datenebene angesiedelt ist, und ob der Fehler bereits in einer einzigen Datensammlung besteht oder erst durch die Integration mehrere Datensammlungen entsteht [7]. Die Klassifikation ist in Abb. 1 wiedergegeben.

DOI 10.1007/s00287-006-0125-5


Felix Naumann
Hasso-Plattner-Institut an der Universität Potsdam,
Prof.-Dr.-Helmert-Str. 2–3, 14482 Potsdam, Deutschland
E-Mail: naumann@hpi.uni-potsdam.de

*Vorschläge an Prof. Dr. Frank Puppe
<puppe@informatik.uni-wuerzburg.de> oder
Prof. Dr. Dieter Steinbauer <dieter.steinbauer@schufa.de>

Alle „Aktuellen Schlagwörter" seit 1988 finden Sie unter:
www.ai-wuerzburg.de/as

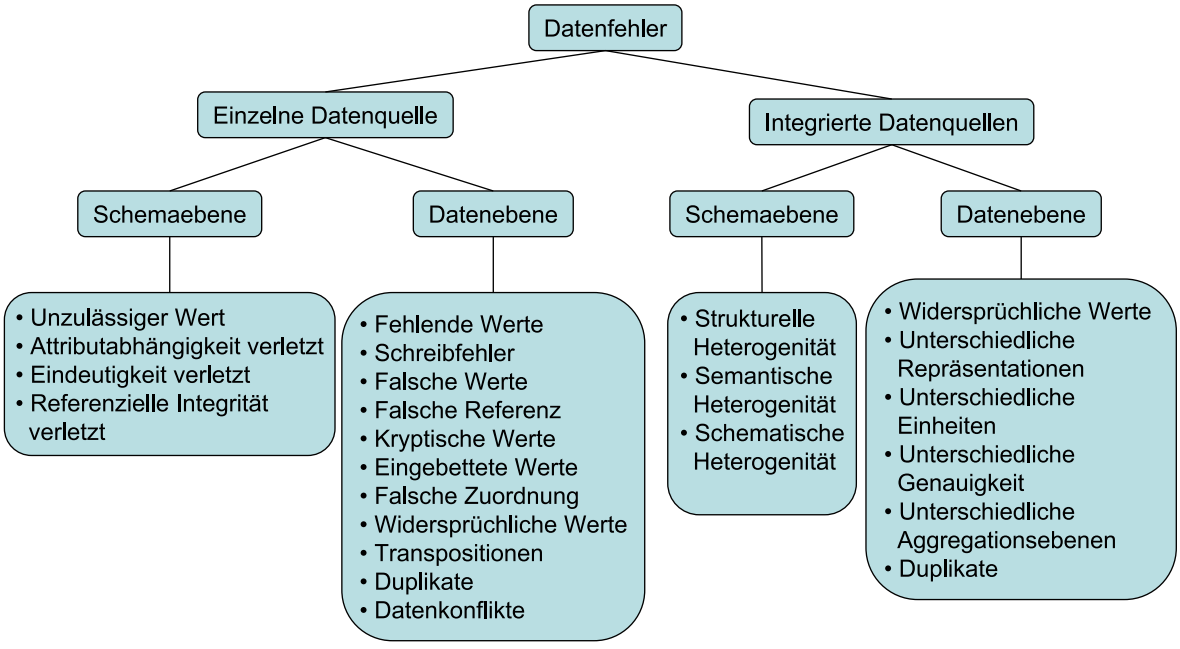


***Abb. 1 Klassifikation von Datenfehlern nach [7]***

Der Begriff der Datenqualität umfasst aber auch abstraktere Merkmale, die sich nicht nur auf einen einzelnen Datenwert oder einen Datensatz beziehen, sondern auf ganze Datenmengen. Beispiele sind die Verständlichkeit einer Datenmenge, deren Vollständigkeit oder auch die Reputation der Datenquelle. Die meistzitierte Aufstellung solcher Informationsqualitätsmerkmale stammt von Wang und Strong [11]. Die Autoren befragten Datenkonsumenten in größeren Unternehmen und destillierten aus einer initialen Menge von 179 Merkmalen die in Tabelle 1 genannten 15 Qualitätsmerkmale. In anderen Arbeiten wurden weitere wichtige Merkmale identifiziert, die, motiviert durch das Gebiet der Informationsintegration, hinzugenommen werden. Dazu gehören u.a. Verifizierbarkeit, Antwortzeit, Latenz und der Preis der Daten.

Zusammenfassend kann Datenqualität also als eine Menge von Qualitätsmerkmalen definiert werden. Die Auswahl der relevanten Merkmale und die genaue Definition der Merkmale bleiben den Experten der jeweiligen Anwendungsdomäne vorbehalten.

## Datenqualität messen

Die Schwierigkeit der konzisen Definition von Qualitätsmerkmalen spiegelt sich auch in den wenig greifbaren Messmethoden wieder. Dennoch sind konkrete, numerische Metriken wichtig – nur so kann man einschätzen, inwieweit die Daten von möglicherweise minderer Qualität Entscheidungen beeinflussen könnten. Und nur so kann man prüfen, ob Datenreinigungs- und Verbesserungsmaßnahmen wirkungsvoll sind und in einem guten Kosten-Nutzen-Verhältnis stehen [6].

Wegen der Subjektivität vieler Qualitätsmerkmale ist der Fragebogen ein wichtiges Instrument zur Qualitätsbestimmung. Beispielsweise schlägt Pierce den Einsatz von Kontrollmatrizen vor, deren Eintragungen von Experten vorgenommen werden und deren Aggregation Auskunft über die Gesamtqualität des "Informationsprodukts" geben [5].

Andere Merkmale hingegen können leichter mit einer konkreten Metrik versehen werden. Die Vollständigkeit einer Datenmenge beispielsweise kann als die Menge der Datensätze im Verhältnis zur Menge aller möglichen Datensätze definiert werden [3]. So kann die Vollständigkeit der OMIM Datenquelle mit 44% angegeben werden, da sie Informationen über 10.995 der geschätzten 25.000 menschlichen Gene speichert [4]. Auf ähnliche Weise wird die Genauigkeit als Anzahl der fehlerbehafteten Datensätze gegenüber allen Datensätzen berechnet. Da in der Regel nicht alle Fehler bekannt sind, werden Sampling-Methoden verwendet, um die manuelle Fehlersuche auf eine kleine Teilmenge

Tabelle 1

**Qualitätsmerkmale nach [11]**

| Merkmalsklasse | Qualitätsmerkmal |
|---|---|
| Intrinsische Datenqualität | Glaubhaftigkeit (*believability*) |
| | Genauigkeit (*accuracy*) |
| | Objektivität (*objectivity*) |
| | Reputation (*reputation*) |
| Kontextuelle Datenqualität | Mehrwert (*value-added*) |
| | Relevanz (*relevancy*) |
| | Zeitnähe (*timeliness*) |
| | Vollständigkeit (*completeness*) |
| | Datenmenge (*amount of data*) |
| Repräsentationelle Datenqualität | Interpretierbarkeit (*interpretability*) |
| | Verständlichkeit (*understandability*) |
| | Konsistenz der Darstellung (*representational consistency*) |
| | Knappheit der Darstellung (*representational conciseness*) |
| Zugriffsqualität | Verfügbarkeit (*accessibility*) |
| | Zugriffssicherheit (*access security*) |

der Daten zu beschränken. Die Entdeckung und anschließende Elimination von Dubletten, einer der wichtigsten Fehlerarten, besprechen wir im nächsten Abschnitt, da hier in der Regel nicht ein Fehlermaß interessiert, sondern die direkte Behebung des Fehlers angestrebt wird.

Neben den Metriken selbst ist deren Aggregation für die Qualitätsbestimmung interessant. Über Fehler in einzelnen Feldern einer Tabelle kann man die Qualität einer Spalte bestimmen ("Die Vollständigkeit der Spalte FAX beträgt 30%."). Eine andere Aggregationsrichtung aggregiert einzelne Fehler zu der Qualität eines Datensatzes ("Die Genauigkeit der Daten des Kunden BMW beträgt 95%."). Diese können wiederum zu der Qualität einer Menge von Datensätzen aggregiert werde ("Die Genauigkeit aller Kundendaten beträgt 80%.") und schließlich, bei der Integration mehrerer Datenquellen zur Qualität des Gesamtergebnisses ("Durch Zusammenführen unserer drei Kundendatenbestände erzielen wir eine Genauigkeit von 87%.").

Während in den vorigen Beispielen stets nur von einem Qualitätsmerkmal die Rede war, will man meist die Datenqualität gemäß mehrerer Merkmale zugleich bestimmen. Um die Werte mehrerer Merkmale, die verschiedene Einheiten, Skalen und Wertebereiche haben können, zu kombinieren, werden die Qualitätswerte zunächst skaliert und dann z.B. gewichtet zu einem Gesamtqualitätswert summiert. So können verschiedene Datenquellen miteinander verglichen werden.

Software-Werkzeuge zur Unterstützung der Qualitätsbestimmung sind sogenannte Data-Profiling-Tools, die entdeckerisch Datenmengen untersuchen, die Kandidaten für Regeln und Bedingungen an die Daten vorschlagen (z.B. Minimal- und Maximalwerte) und dann die Einhaltung dieser Regeln und Bedingungen prüfen.

## Datenqualität verbessern

Sind qualitative Mängel in den Daten erst festgestellt, bleiben zwei Alternativen. Die erste Alternative, die besonders relevant ist, wenn Daten aus externen Quellen importiert werden, ist es, mit den Daten und ihrer Qualität bewusst umzugehen. Die Kenntnis der minderen Qualität, u.a. durch geeignete Darstellung in Berichten, kann falsche Entscheidungen aufgrund falscher Daten verhindern. Hat man hingegen die Kontrolle über die Daten, ist die zweite Alternative, die aktive Verbesserung der Datenqualität geeignet.

Daten durchlaufen einen komplexen Produktionsprozess von ihrer Entstehung bis hin zu ihrem "Konsum". In [10] wird konsequenterweise vorgeschlagen, Informationen als Produkte anzusehen, deren Produktionsprozesse anhand sogenannter IP-Maps dargestellt werden können. Diese stellen mittels einer graphischen Sprache systematisch

den Werdegang von Daten dar – beginnend bei der Datenproduktion über Veränderungen, Aggregationen und qualitativer Prüfungen bis hin zum Konsum der Daten zur Entscheidungsfindung. Am wirkungsvollsten und nachhaltigsten werden Datenfehler bei ihrer Entstehung bekämpft, also bei der manuellen Dateneingabe bzw. bei der automatischen Datenerhebung. Datenbanksysteme bieten hier insbesondere die Möglichkeit, Integritätsbedingungen zu formulieren, die z.B. die Einhaltung bestimmter Formate erzwingen (z.B. bei Datumsangaben), die Eingabe bestimmter Werte erzwingen (z.B. die E-Mail-Adresse von Kunden) oder die Konsistenz von Datensätzen sicherstellen (z.B. PLZ und Ort). Diese Einschränkungen erschweren jedoch oft die Datenerfassung, z.B. wenn ein Kunde keine E-Mail-Adresse besitzt oder sie nicht preisgeben will.

Viele Datenfehler lassen sich auch durch domänenspezifische Normalisierung/Transformation (Nachname, Vorname → Vorname Nachname) und durch Standardisierung (MM/TT/YY → TT/MM/YYYY) der Daten beheben. Fehlende Werte können so ergänzt werden, dass statistische Größen wie Durchschnitt oder Standardabweichung unverändert bleiben. Diverse Methoden zur Ausreißererkennung finden und eliminieren wahrscheinliche Datenfehler.

Zu den kostenträchtigsten Datenfehlern gehören die schon erwähnten Dubletten. Um sie aufzuspüren muss zunächst ein Maß entwickelt werden, das die Ähnlichkeit zweier Datensätze bestimmt. Ausgehend von Einzelähnlichkeiten, etwa der Namensähnlichkeit und der Adressähnlichkeit, wird eine Gesamtähnlichkeit ermittelt und gegen einen Schwellwert verglichen. Sind die Datensätze hinreichend ähnlich, werden sie als Dublette gekennzeichnet und z.B. zur Vorlage an einen Experten gereicht. Neben der Schwierigkeit, ein der Anwendungsdomäne angemessenes Ähnlichkeitsmaß zu definieren, bedarf es auch Algorithmen, die den paarweisen Vergleich *aller* Datensätze vermeiden. Nur so kann man den quadratischen Aufwand umgehen, der schon für 10.000 Datensätze inakzeptabel hoch wäre. Zu diesem Zweck partitioniert man üblicherweise die Daten und vergleicht z.B. nur noch Personen, die in der gleichen Postleitzahl wohnen. Ein anderer Ansatz ist es, Daten geeignet zu sortieren, ein Fenster einer bestimmten Größe über die Daten zu schieben und nur Datensätze, die innerhalb eines Fensters auftauchen, zu vergleichen [2].

Nachdem Dubletten erkannt sind, kann eine Verbesserung der Datenqualität durch die Fusion der Dubletten erzielt werden. Im einfachsten Fall werden bis auf einen alle Datensätze einer Dublettengruppe gelöscht. Bessere Verfahren kombinieren Datenwerte der einzelnen Datensätze und gelangen so zu einem vollständigeren und verlässlicheren Datensatz.

## Community und Ausblick

Es hat sich international und auch in Deutschland eine Community rund um die Datenqualität gebildet. Das MIT veranstaltet in diesem Jahr zum mittlerweile elften Mal die International Conference on Information Quality (ICIQ), die zwar einen Schwerpunkt in der Wirtschaftsinformatik hat, aber auch "reinen" Informatikthemen einen Platz bietet. In 2004 wurde die Deutsche Gesellschaft für Informationsqualität (DGIQ) gegründet, die ebenfalls regelmäßig eine Konferenz veranstaltet. Hinzu kommen zahlreiche internationale Workshops, die vornehmlich im Rahmen von Datenbankkonferenzen wie der VLDB oder SIGMOD abgehalten werden. Zuletzt sei auf eine jüngere Ausgabe des Datenbankspektrums hingewiesen, die sich ganz der Datenqualität widmet [9].

Neben dem in diesem Überblick vorgestellten Informatikaspekt der Datenqualität, über den z.B. Batini und Scannapieco einen tieferen Überblick geben [1], ist die Managementsicht auf mangelnde Datenqualität und deren Verbesserung ebenso wichtig. Redman gibt in seinem Buch einen guten Einstieg in diese andere Facette des Themas [8].

Datenqualität bzw. deren Mangel ist ein Problemfeld, das die Informatik auch noch in ferner Zukunft beschäftigen wird. Wo immer Daten entstehen oder verarbeitet werden, entstehen auch Datenfehler. Methoden der Informatik können viele dieser Fehler entdecken und korrigieren, aber nur gepaart mit einer Bekämpfung der Fehlerursache ist der Kampf um für hohe Datenqualität zu gewinnen.

## Literatur

1. Batini, C., Scannapieco, M.: Data Quality: Concepts, Methods and Techniques. Heidelberg: Springer Verlag (2006)
2. Hernández, M.A., Stolfo, S.J.: Real-world data is dirty: Data cleansing and the merge/purge problem. Data Mining and Knowledge Discovery 2(1), 9–37 (1998)

3. Naumann, F., Freytag, J.-C., Leser, U.: Completeness of integrated information sources. Inf. Syst. 29(7), 583–615 (2004)
4. Online Mendelian Inheritance in Man, OMIM. McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), http://www.ncbi.nlm.nih.gov/omim/ (2006)
5. Pierce, E.: Assessing data quality with control matrices. Commun. ACM 47(2), 82–86 (2004)
6. Pipino, L., Lee, Y., Wang, R.: Data quality assessment. Commun. ACM 4, 211–218 (2002)
7. Rahm, E., Do, H.-H.: Data cleaning: Problems and current approaches. IEEE Data Eng. Bull. 23(4), 3–13 (2000)
8. Redman, T.C.: Data Quality – The Field Guide. Boston: Digital Press (2001)
9. Saake, G., Sattler, K.-U., Naumann, F. (Eds.) Datenbankspektrum – Daten- und Informationsqualität, volume 14. Heidelberg: dpunkt.verlag (2005)
10. Shankaranarayanan, G., Wang, R.Y., Ziad, M.: IP-MAP: Representing the manufacture of an information product. In: Proceedings of the International Conference on Information Quality (IQ), pages 1–16 (2000)
11. Wang, R.Y., Strong, D.M.: Beyond accuracy: What data quality means to data consumers. J. Manage. Inf. Syst. 12(4), 5–34 (1996)