



Datenqualität: allgemeiner Überblick

Waldemar Braun

Seminar Datenqualität

OvGU Magdeburg

03.12.2009

Gliederung

1. Einleitung
2. Motivation
3. Definition
4. DQ-Probleme
5. DQ-Dimensionen
6. DQ-Modelle
7. Messen der Datenqualität
8. Verbesserung der Datenqualität
9. Zusammenfassung
10. Quellen

Einleitung

- Daten von niedriger Qualität sind in DB allgegenwärtig
 - Daten in unterschiedlichen Einheiten
 - Fehlerhaft eingetippte Daten
 - ➔ Garbage-In-Garbage-Out
- Niedrige Datenqualität
 - Mangelndes Vertrauen in die Daten
 - Fehlentscheidungen

Motivation

- Customer Relationship Management (CRM)
- Mögliche Fehler:
 - Tippfehler
 - Verständigungsfehler (z.B. über Telefon)
 - Fehlende angaben (z.B. E-Mail-Adresse)
 - Duplikate (auch „Dublette“)
- fehlerhafte Daten verursachen Kosten
 - Nichterreichen eines Kunden
 - Mehrfaches versenden von Unterlagen
 - Überschreitung des Kreditlimits

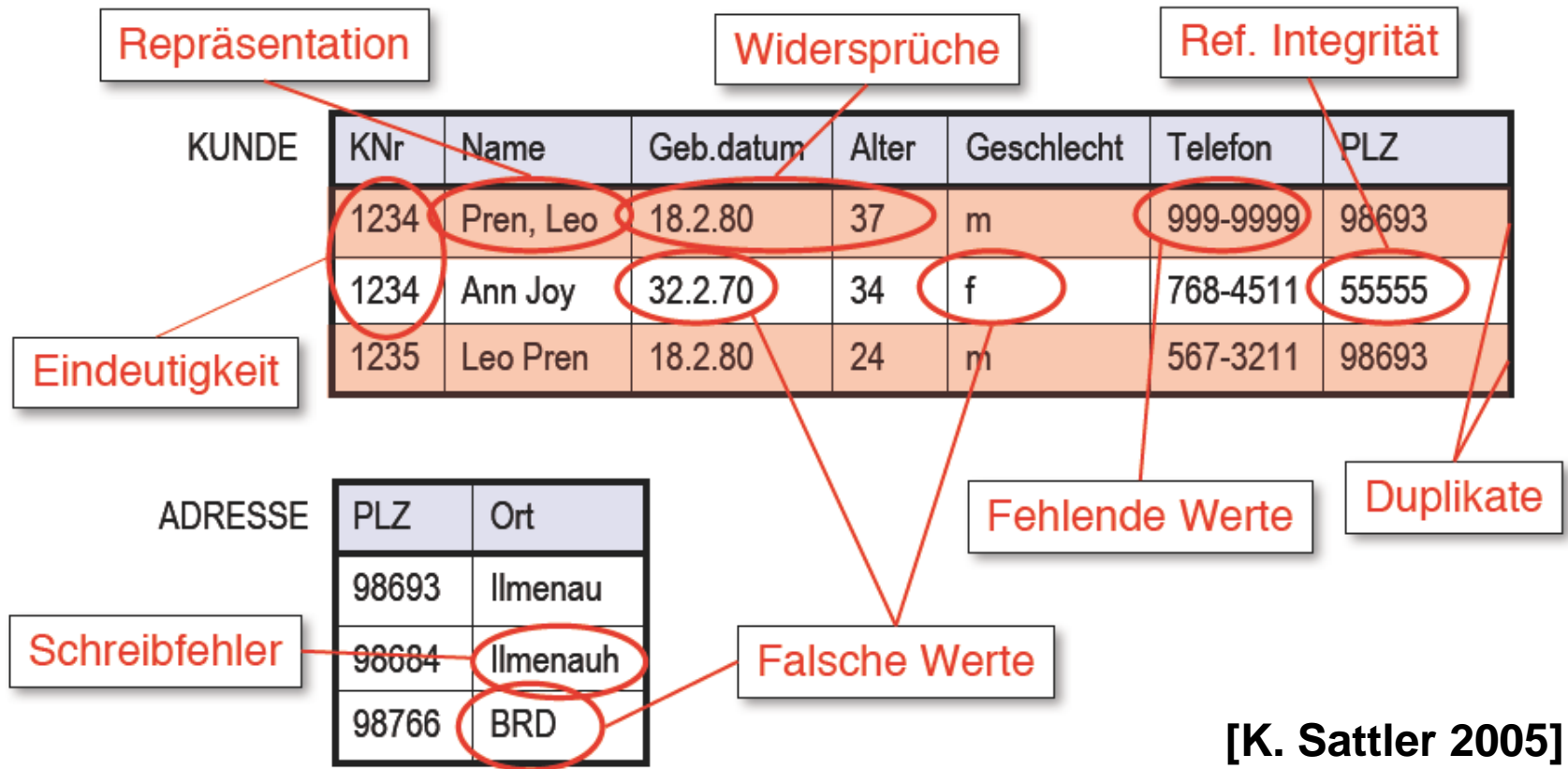
Definition

- „Daten“: logisch gruppierte Informationseinheiten
- „Qualität“: Grad der Übereinstimmung zw. Soll und Ist

„Datenqualität“ (auch “Informationsqualität”):

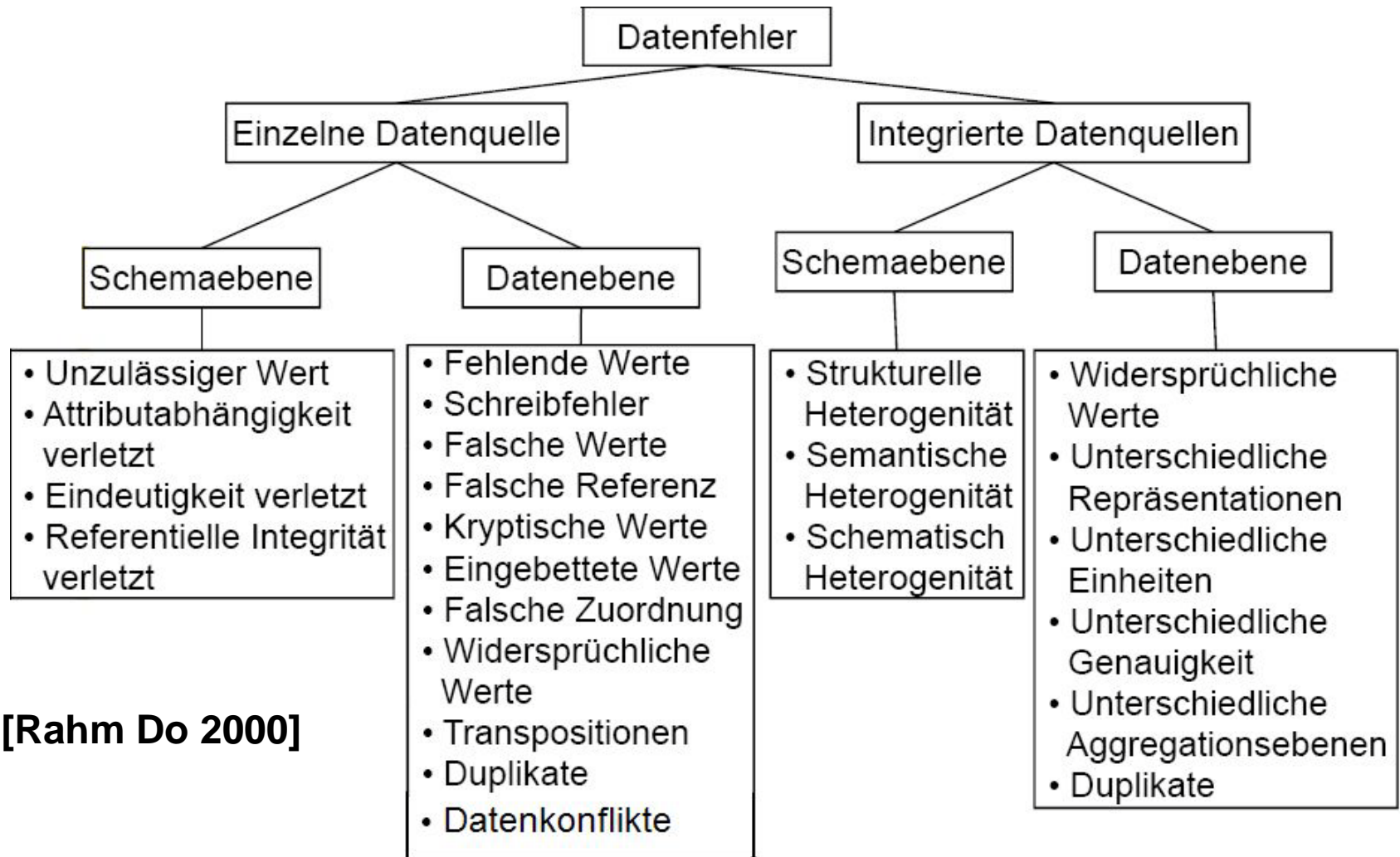
- Eignung der Daten für die jeweilige datenverarbeitende Anwendung
- „Fitness for use“ – Gebrauchstauglichkeit der Daten in den Augen des Kunden
- Menge von Qualitätsmerkmalen für einen bestimmten Verwendungszweck

DQ-Probleme



[K. Sattler 2005]

DQ-Problemen: Klassifikation



[Rahm Do 2000]

DQ-Probleme: Ursachen

- „Datenproduktion“:
 - Verschiedene Quellen
 - Probleme bei der Datenerfassung
- Speicherung:
 - Unterschiedliche Formate
 - Ungeeignete Formate
- Nutzung:
 - Veränderung der Nutzerbedürfnisse
 - Sicherheits- und Zugriffsprobleme

DQ-Dimensionen

- viele Dimensionen bzw. Qualitätskriterien zur Definition und Messung der DQ

- Vollständigkeit: Welche Daten fehlen?
- Genauigkeit: Welche Daten sind unkorrekt?
- Zeitnähe: Wie alt sind die Daten?
- Relevanz: Wie nützlich sind die Daten?
- ...

 Eigener Vortrag „DQ-Dimensionen“

DQ-Modelle

■ Datenmodelle:

- Anreicherung traditioneller Datenmodelle um Strukturen zur Repräsentation und Analyse von DQ
z.B. Quality ER-Modell

■ Prozessmodelle:

- Modellierung des „Datenproduktionsprozesses“
z.B. IP-MAP

 Eigener Vortrag „DQ-Modelle“

Messen der Datenqualität I

- **Notwendigkeit des Messens**
 - Abschätzung der Güte / Aussagekraft
 - Verbesserung notwendig?
 - Kosten-Nutzen-Verhältnis

- **DQ ist subjektiv → wichtiges Instrument zur Qualitätsbestimmung:**
 - Fragebögen - z.B.: Kontrollmatrizen → Gesamtqualität

Messen der Datenqualität II

■ Konkrete Metriken:

- Vollständigkeit = Anz. der Datensätze / alle mögl. Datensätze
- Genauigkeit = fehlerhafte Datensätze / alle Datensätzen
 - Anwendung von Sampling-Methoden (Stichproben)
- Konsistenz (Widerspruchsfreiheit): Einhaltung von Regeln
 - z.B. Alter = AktDatum - GebDatum
- Zeitnähe = aktueller Zeitpunkt - Beobachtungszeitpunkt

■ Mehrere Merkmale:

- Wichtung und Aufsummierung zum Gesamtqualitätswert
 - Wichtig für Vergleich verschiedener Datenquellen

Verbesserung der Datenqualität I

- Data-Profiling-Tools: Software zur Qualitätsbestimmung
 - Überprüfung der Einhaltung von Regeln
- Bei Qualitätsmängeln:
 - bewusster Umgang mit Daten minderer Qualität
 - aktive Verbesserung der Datenqualität
- Data Cleaning (auch „Cleansing“ oder „Scrubbing“)
 - Beseitigung von Widersprüchen, Fehlern usw.
 - Bis zu 80% des Aufwandes in DW-Projekten
- Verbesserung durch:
 - Normalisierung: TT/MM/YYYY
 - Ergänzung fehlender Werte
 - Ausreißererkennung und -behandlung

Verbesserung der Datenqualität II

Erkennung von Duplikaten:

- Einzelähnlichkeiten → Gesamtähnlichkeit
- Markierung als Duplikat und Vorlage an Experten

■ Probleme:

- geeignetes Maß für Ähnlichkeit zweier Datensätze
- geeigneter Algorithmus zum Vergleich
 - Paarweises Vergleichen → quadratischer Aufwand

■ Lösungsansätze:

- Partitionierung der Daten
- Daten geeignet sortieren

■ Verbesserung durch Fusion der Duplikate:

- Löschung
- Kombination

Zusammenfassung

- Datenqualität - Menge von Qualitätsmerkmalen
 - Auswahl und genaue Definition durch Experten
- großer Bedeutung für DB-Systemen
- Steigender Stellenwert in Unternehmen
- Mangelnde Datenqualität → großes Problemfeld der Informatik
 - Datenfehler überall wo Daten entstehen
 - Methoden: Fehler entdecken und korrigieren
 - WICHTIG: Bekämpfung der Fehlerursache

Quellen

- „Datenqualität“ von Felix Naumann, Informatik Spektrum 2007, Springer-Verlag Berlin Heidelberg
- Seminar „Duplikaterkennung“ 2008 von Felix Naumann, Uni Potsdam
- „Datenqualität – eine Datenbankorientierte Sichtweise“ DB-Tutorientage 2005, Kai-Uwe Sattler, TU Ilmenau