

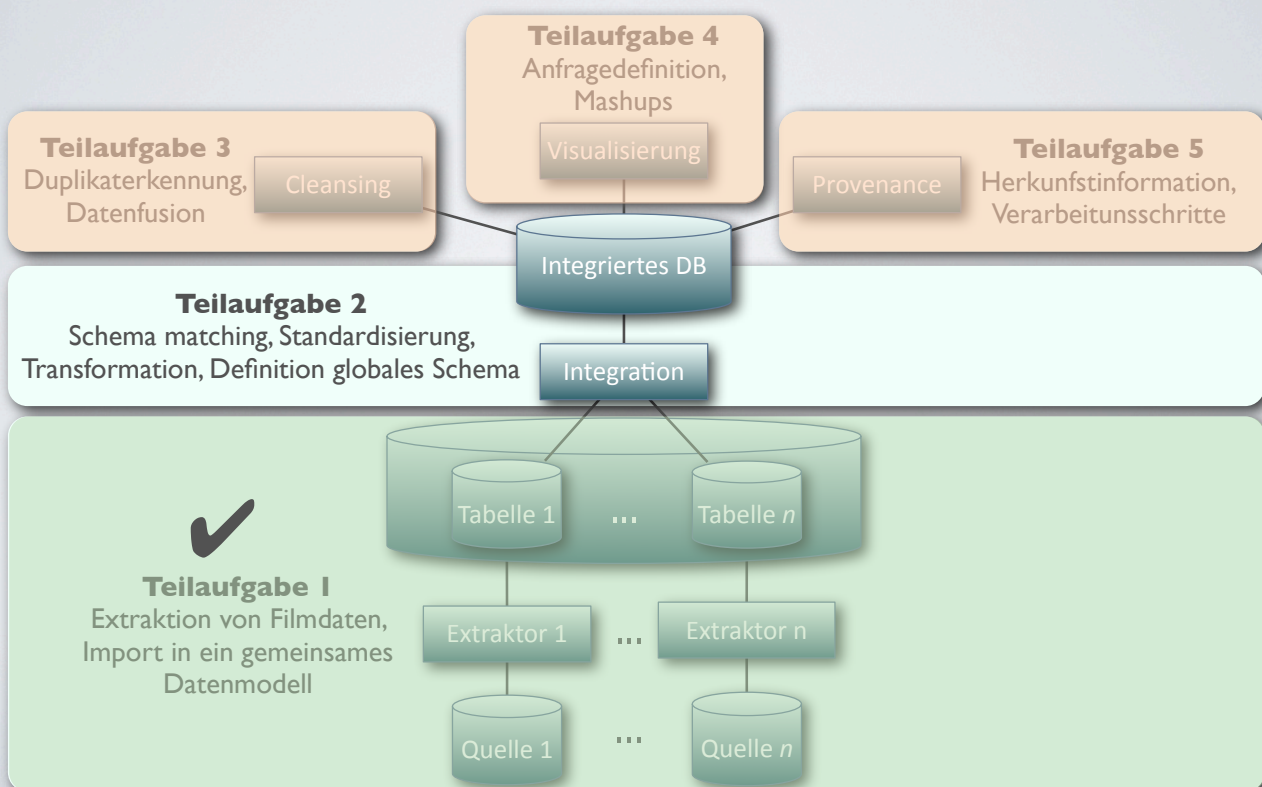


DATENINTEGRATION & DATENHERKUNFT

Übung Wintersemester 2010/11

Melanie Herschel | Lehrstuhl für Datenbanksysteme | Universität Tübingen

AKTUELLER STAND



ZEITPLAN

Termin 1	Heute	Überblick + Einführung Extraktion	
Termin 2	10.11.2010	Einführung Integration	Vorstellungen Extraktion
Termin 3	24.11.2010		Vorstellungen Integration
Termin 4	8.12.2010	Einführung Datenreinigung	
Termin 5	22.12.2010	Einführung Visualisierung	Vorstellungen Datenreinigung
Termin 6	19.01.2011	Einführung Datenherkunft	Vorstellungen Visualisierung
Termin 7	26.01.2011		Vorstellungen Gesamtprojekt

Melanie Herschel | Lehrstuhl für Datenbanksysteme | Universität Tübingen

Notenrelevant!

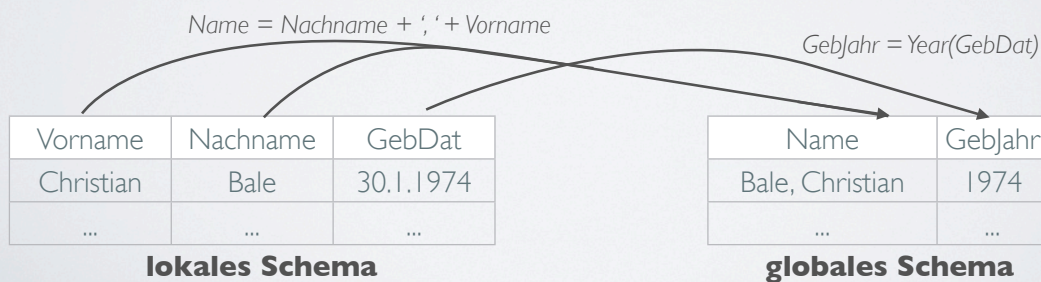
EINFÜHRUNG INTEGRATION

- Definieren Sie ein **globales Schema**
 - Datenmodell frei wählbar (z.B. relational, XML)
 - Redundanz vermeiden
 - im Schema (z.B. nur eine Film-Tabelle)
 - in den Daten (Normalisierung)
 - Nur Informationen übernehmen, die Ihre Anwendung braucht (falls nicht schon bei der Extraktion geschehen).

Melanie Herschel | Lehrstuhl für Datenbanksysteme | Universität Tübingen

EINFÜHRUNG INTEGRATION

- Definieren Sie ein **Schema Matching** zwischen den Quellschemata und dem globalen Zielschema
 - Automatisierung, soweit möglich
 - Manuelle Lösung als Fallback für vom automatischen Matching nicht gefundene oder falsch gefundene Matches



Melanie Herschel | Lehrstuhl für Datenbanksysteme | Universität Tübingen

EINFÜHRUNG INTEGRATION

- Nützliche **Links** für **automatisiertes Schema Matching**
 - Frei verfügbares Tool: COMA++
<http://dbs.uni-leipzig.de/Research/coma.html>
 - Zur eigenen Implementierung nützliche Bibliothek: SecondString (für Stringvergleiche)
<http://secondstring.sourceforge.net/>
 - Techniken zum automatisierten Schema Matching in der Vorlesung

Melanie Herschel | Lehrstuhl für Datenbanksysteme | Universität Tübingen

EINFÜHRUNG INTEGRATION

- **Standardisierung** der Quelldaten
 - Alle Werte eines Attributs sollten dem selben Format entsprechen
 - Beispiele:
 - amerikanisches (MM/DD/YYYY) vs deutsches (DD.MM.YYY) Datumsformat
 - Namen: "Brad Pitt" vs. "Pitt, Brad" vs. "Mr. Pitt" vs. "B. Pitt" ...
 - Frühzeitige Standardisierung erleichtert z.B. späteres Data Cleaning

Melanie Herschel | Lehrstuhl für Datenbanksysteme | Universität Tübingen

EINFÜHRUNG INTEGRATION

- **Transformation** der Quelldaten in die Zielrepräsentation
 - Laden Sie Ihre Quelldaten in das Zielschema
 - Führen Sie dabei die Standardisierung durch
 - Beachten Sie dabei das von Ihnen ermittelte Schema Matching
 - Somit implementieren Sie hier materialisierte Integration

Melanie Herschel | Lehrstuhl für Datenbanksysteme | Universität Tübingen

ABGABE

- Erstellen Sie eine Präsentation Ihrer bisherigen Lösung.
- Schicken Sie mir Ihre Präsentation (PDF, Keynote, oder PPT(X)) mit Angabe Ihrer Namen und Matrikelnummern bis zum 23.11., 18 Uhr
- Am 24.11. stellt ihr Team seine Ergebnisse vor. Zusätzlich zu den Folien können Sie anhand Ihres Prototypen Ihre Lösung veranschaulichen.

