



Datenintegration als automatisierter Prozess



Hauptseminar

Tom Schöner

tom.schoener@haw-hamburg.de

Department Informatik, Informatik Master

12.06.2018

- Einführung
- Thema und Fragestellungen
- Problembeschreibung
- Motivation
- Relationale Ansätze mit Views
- Semantische Integration
- Risiken
- Ausblick

Dissertation

Data Integration against Multiple Evolving Autonomous Schemata [1]

Christoph Koch, 2001

- Zusammenführung einer oder mehrerer Quellen zu einer Entität
- Voraussetzung für Aggregation und Analytical Processing
- Unterscheidung zwischen:
 - Source integration
 - Data integration
- Ziel: Interoperabilität unter Informationssystemen

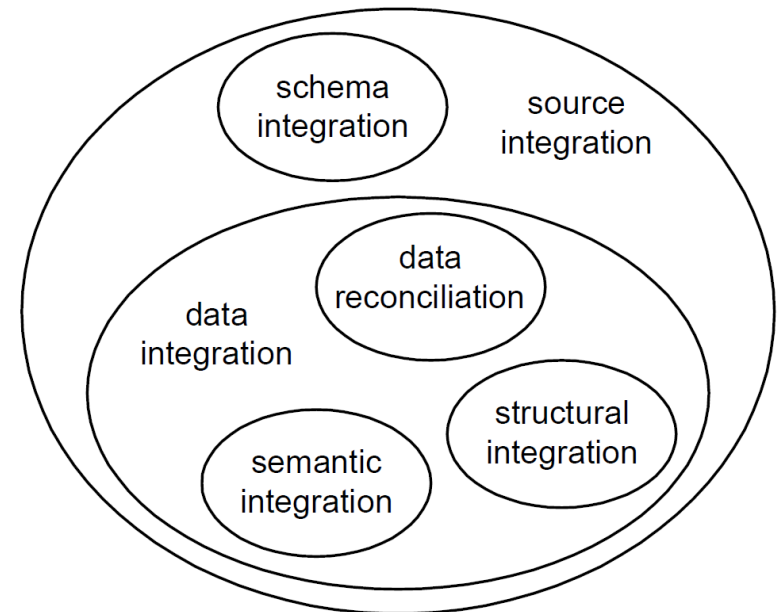
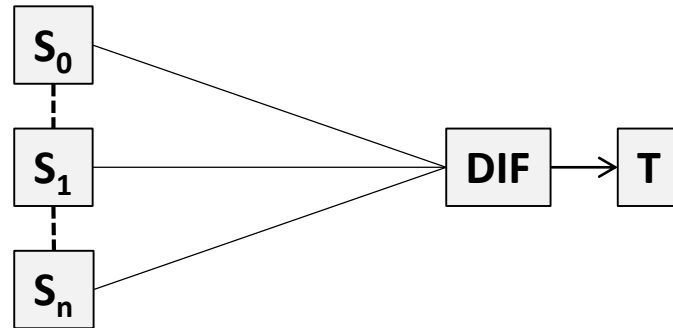


Abb. 1: Source Integration Hierarchie

- Structural integration [3]
 - Auflösung der Heterogenität von u. a. Modellen, Sprachen oder Protokollen mittels *wrapping*
 - Hilfreich bei Legacy-Applikationen
- Semantic integration [4]
 - Erkennung semantischer Unterschiede der Schemata
 - Problematik: Semantik/Bedeutung der Schemata und Daten unterschiedlicher Quellen ist nicht äquivalent
- Data reconciliation [2]
 - Validierung, Filterung, Erkennung von Duplikaten
 - Aufbereitung fehlerhafter Daten (z.B. Rechtschreibfehler)

Datenintegration als eigenständiger, automatisierter Prozess

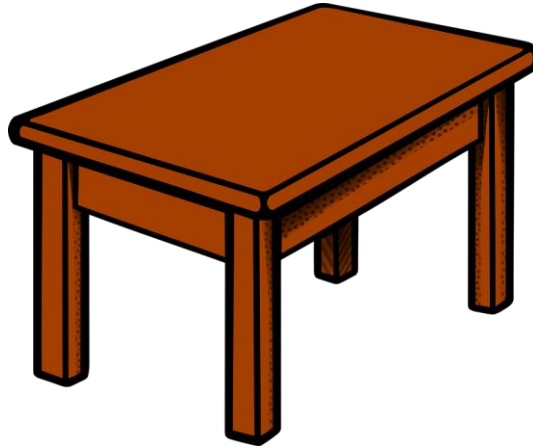
- Wie sieht ein Datenintegrationsprozess aus?
- Welche Bereiche der Datenintegration können automatisiert werden – Welche Parameter werden benötigt?
- Auf welchen Modellen kann aufgebaut werden?
- Lassen sich relationale und nicht-relationale Ansätze verbinden?



$S_0 - S_n$	Quellen
DIF	Data-Integration Framework
T	Ziel-System

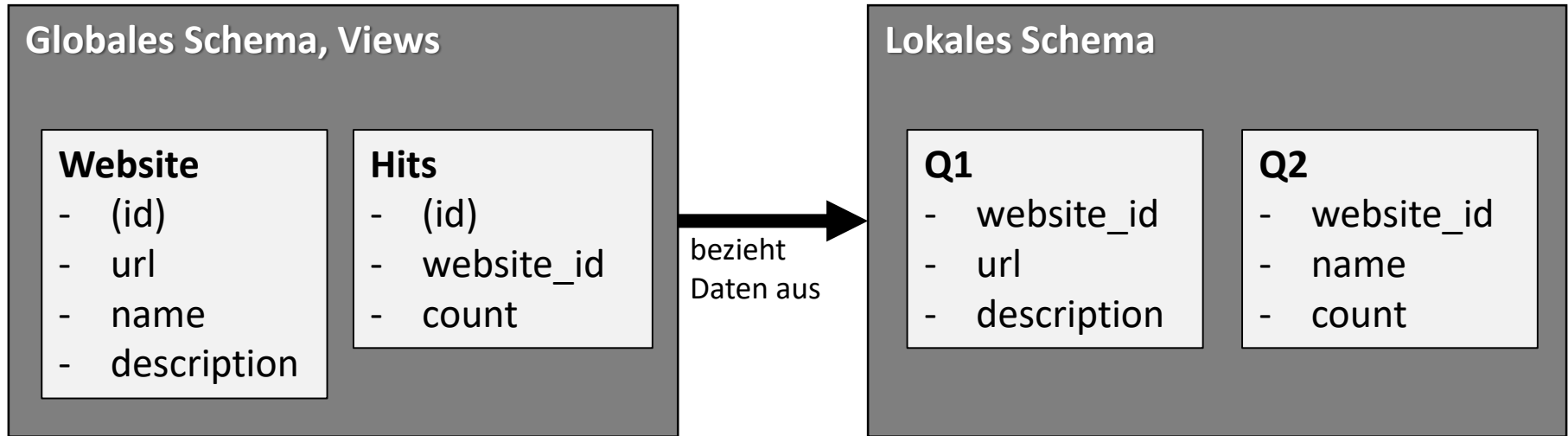
- Problem:
 - **T** kann nicht direkt auf die Daten der Quellen **S** zugreifen.
 - Die Daten liegen im falschen Format vor, sind unvollständig, ...
- Data-Integration Framework **DIF** als Kommunikationsschicht

- Aggregation und Überführung relevanter Daten in ein anderes System ist komplex
- Relevanz: Integration von Daten(-Systemen) ist als essentieller Teil der Softwareentwicklung anzusehen
- MARS: Datenintegration für Multiagenten-simulationen
- Verarbeitung großer Datenmengen
- Datenintegrität



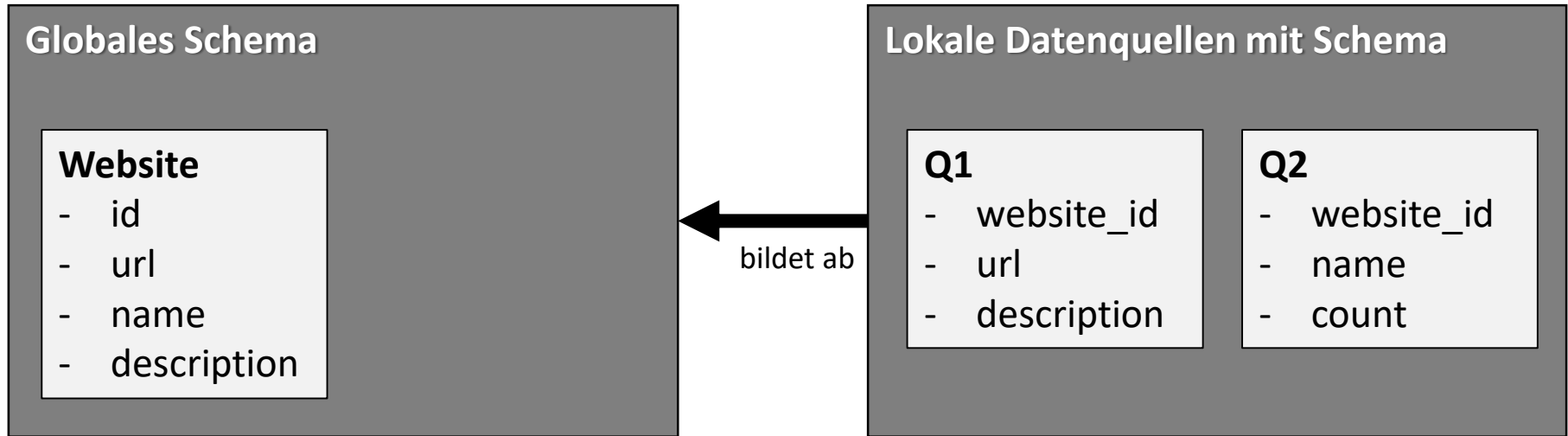
Ein relationaler Ansatz mit
Views

$$v_i(S_1, \dots, S_n) \subseteq G_i$$



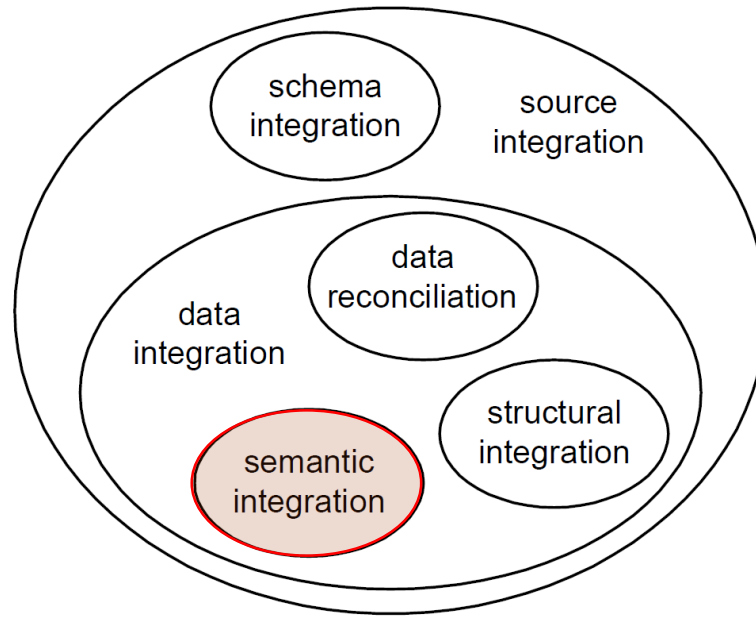
```
CREATE VIEW Website AS
  SELECT Q1.url, Q2.name, Q1.description
  FROM Q1, Q2
  WHERE Q1.website_id = Q2.website_id
```

$$S_i \subseteq v_i(G_1, \dots, G_m)$$



```
CREATE VIEW S1 AS
  SELECT website_id AS id, url, NULL, description FROM Website
CREATE VIEW S2 AS
  SELECT website_id AS id, NULL, name, NULL FROM Website
```

- Verschachtelte Views erhöhen die Komplexität
- GaV: Assoziationen gehen verloren
- LaV: Anfragen sind komplexer als bei GaV



Semantische Integration

1. What are the **sites or systems** that may contain some or all of the desired information?
2. What are the **relevant data** related to the **expressed needs**?
3. How to **transform, adapt** and **match** these data to obtain a homogeneous and **understandable result** by the user?

TABLE I. SCHEMA CONFLICTS DEFINITION TABLE

Schema conflicts	Definition
Name conflicts	Confusing labels of database schema elements to another (feature classes, attributes and relationships) due to the existence of synonyms, homonyms and hyponyms.
Identifier entities conflicts	Assigning different identifiers (primary key) on the same concept in different databases.
Generalization conflicts	Different conception for modelling related classes (inheritance between classes).
Aggregation conflicts	The use of aggregation in a database to identify a set of entities in other database. Therefore, the properties and their values in a database can aggregate the priorities and the corresponding values of entities in another database.
Isomorph schema conflicts	The logical structure of a set of attributes and their values belonging to an entity class in a database is organized to form a different structure in another database.

Abb. 2

1. Ansatz: mapping-based
 - Globale und lokale Schemata (vgl. GaV)
 - Schema- und Applikationsabhängig
2. Ansatz: query-oriented
 - Logische deklarative Sprachen oder SQL
3. Ansatz: intermediary-based
 - Mediators, Agents
 - Vermittler sind jeweils für die Koordination der heterogenen Quellen zuständig
 - Vermittler einigen sich auf standardisierte Ontologie
 - Schema- und Applikationsunabhängig

- Heterogenität von Datenquellen erschwert eine Klassifizierung
 - Mögliche Lösung: Einschränkung auf Data reconciliation, Semantic integration oder Structural integration
- Bekanntes Problem: Viele Lösungen sind bereits implementiert und erforscht (Beispiel: Mediator Schema)
- Framework lässt sich nur beschränkt einsetzen/ hat strenge Rahmenbedingungen

- Aktuellen Stand der Forschung weiter verfolgen
 - Vergleichbare Frameworks und Prozesse analysieren
 - Genauere Eingrenzung des Themengebiets
 - Prototyp des data integration Frameworks (DIFs)
- Grund- und Hauptprojekt sind noch offen



- [1] Christoph Koch, 2001. "Data Integration against Multiple Evolving Autonomous Schemata", Institut für medizinische Kybernetik und Artificial Intelligence Universität Wien
- [2] Matthias Jarke, Maurizio Lenzerini, Yannis Vassiliou, and Panos Vassiliadis, 2000. Fundamentals of Data Warehouses. Springer-Verlag
- [3] Mary Tork Roth and Peter Schwarz. "Don't Scrap It, Wrap It! A Wrapper Architecture for Legacy Data Sources". In Proceedings of the 1997 International Conference on Very Large Data Bases (VLDB'97), 1997.
- [4] Hector Garcia-Molina, Yannis Papakonstantinou, Dallen Quass, Anand Rajaraman, Yehoshua Sagiv, Jeffrey D. Ullman, Vasilis Vassalos, and Jennifer Widom. "The TSIMMIS Approach to Mediation: Data Models and Languages". Journal of Intelligent Information Systems, 8(2):117–132, 1997.
- [5] Halevy, A. The VLDB Journal (2001) 10: 270. <https://doi.org/10.1007/s007780100054>
- [6] O. Banouar and S. Raghay, "Comparative study of the systems of semantic integration of information: A survey," 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA), Marrakech, 2015, pp. 1-8., doi: 10.1109/AICCSA.2015.7507235, URL: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7507235&isnumber=7507087>

- Abb. 1: Christoph Koch, 2001. "Data Integration against Multiple Evolving Autonomous Schemata", Institut für medizinische Kybernetik und Artificial Intelligence Universität Wien
- Abb. 2: O. Banouar and S. Raghay, "Comparative study of the systems of semantic integration of information: A survey," 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA), Marrakech, 2015, pp. 1-8., doi: 10.1109/AICCSA.2015.7507235, URL: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7507235&isnumber=7507087>
- Questionmarks, Jan 29, 2016: <https://blog.knowyourcompany.com/the-four-questions-you-should-ask-every-employee-c781b175d0b4>