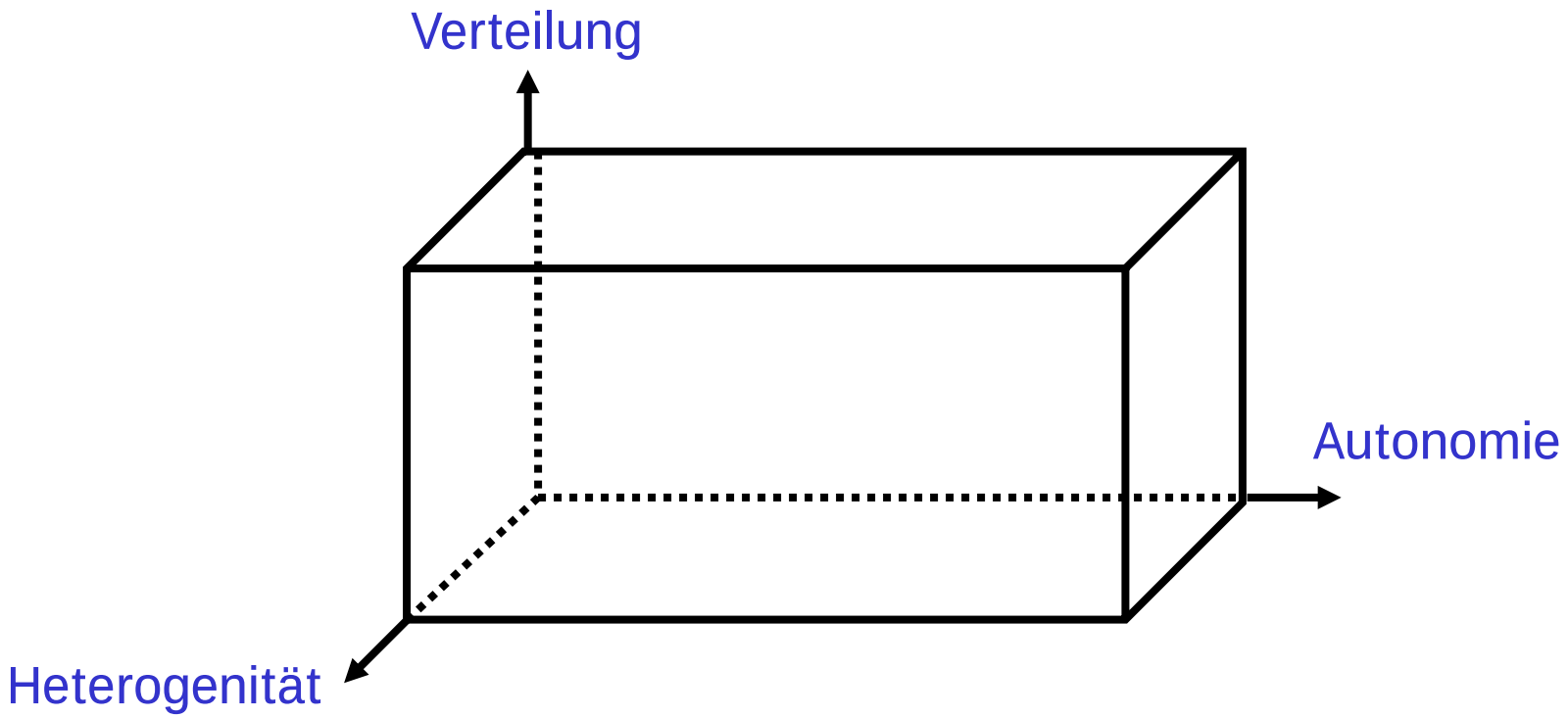
Informationsintegration

Verteilung, Autonomie, Heterogenität, Transparenz

Ulf Leser

Klassifikationsdimensionen [ÖV99]

Eigenschaften von Informationssystemen in Bezug auf deren Integration



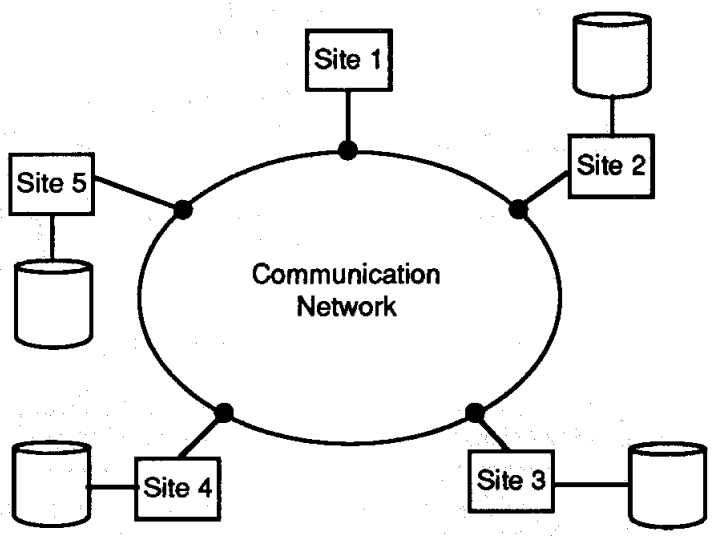
Inhalt dieser Vorlesung

- Verteilung
  - Physische Verteilung
  - Logische Verteilung
- Autonomie
- Heterogenität
- Transparenz

Verteilung

- Ein verteiltes Informationssystem ist eine Sammlung mehrerer, logisch verknüpfter Informationssysteme, die zur Erfüllung einer gemeinsamen Aufgabe untereinander kommunizieren

- Zwei Aspekte
  - Physische Verteilung
  - Logische Verteilung



Physische Verteilung

- Server stehen an unterschiedlichen Orten
  - Anderes Land, Gebäude, Raum, Schrank, Rack, ...
  - Server sind physikalisch unabhängig (hoffentlich)
- Bedeutet idR „Shared Nothing“
  - Server haben keine gemeinsamen Speicher, Disk, CPU, ...
- Gründe / Vorteile
  - Höhere Sicherheit im Katastrophenfall
  - Lastbalancierung
  - Latenz / Bandbreite: Nähe von Servern zu Clients
  - Historisch begründet
  - Physikalische Einschränkungen (Hitze, Gewicht, Energie)

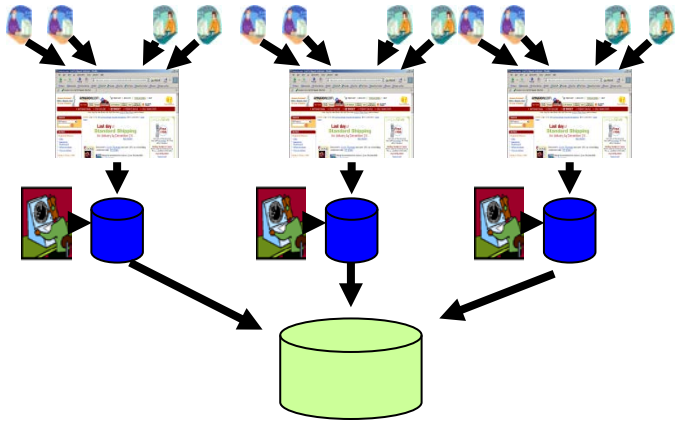
Nachteile

- Optimierung von Anfragen schwieriger
- Daten müssen transportiert werden
  - Netzwerkverkehr statt IO oder Prozesswechsel
- Schwierigeres Management
- Gefahr schleichend zunehmender Heterogenität

Logische Verteilung

- Daten stehen an verschiedenen logischen „Orten“
  - Können, müssen aber nicht physisch unterschiedlich sein
  - Verschiedene Attribute / Tabellen / Schemata
  - Auch in einem Schema (Tabellen buecher1, buecher2)
- Relationale Normalformen verlangen Verteilung
  - Verschiedene Attribute für Tabellen für verschiedene Daten
  - Verknüpfung verlangt intensionale und extensional Überlappung (Primary key – foreign key)
- Manchmal: Verteilung semantisch gleicher Daten
  - Schlechte Integration – Redundanz wird nicht entfernt
  - Historisch gewachsen, Autonomie von Datenquellen
  - Zur Beschleunigung (Partitionierung)
  - Birgt Gefahr (extensionaler) Duplikaten

Logischer Verteilung und Intension



- Logische Verteilung
  - Tabellen „de_verkauf“, „fr_verkauf“ sind kontrolliert verteilt bzgl. des Verkaufsort
  - Damit sind die Daten disjunkt
  - Versucht man aber, alle Verkäufe eines Kunden X aus Strassburg zu erhalten ...
    - X kann in Deutschland oder Frankreich eingekauft haben
    - X wird doppelt in den Kundendatenbanken stehen
  - Extensionale Überlappung bzgl. Kunden
  - Keine extensional Überlappung bzgl. Verkäufen

Vorteile und Nachteile von Verteilung

- Vorteile
  - Befriedigung anarchischer oder machtpolitischer Instinkte
  - Kosten sparen – keine Migration
  - Kontrollierte logische Verteilung ist Voraussetzung für geplante physische Verteilung
- Nachteile unkontrollierter Verteilung
  - Daten werden übersehen
  - Daten erscheinen doppelt
  - Daten haben Widersprüche (bei Duplikaten)
  - Daten sind inhomogen und schwer verständlich
  - Unkontrollierte Redundanz hat keine rationalen Vorteile

Überwindung

- Integration macht Verteilung unsichtbar
  - Einheitlicher Zugang, einheitliche Daten
- Physische Verteilung
  - Virtualisierte Integration
  - Materialisierung an einem Ort
- Logische Verteilung
  - Semantische Integration
    - Kann unter Beibehaltung physischer Verteilung erfolgen
    - Kann vom System (Mediatoren) oder vom Benutzer (Multidatenbanksprachen) erledigt werden
    - Schemaintegration, Schemamapping, Übersetzungstabellen, logische Formalismen, ...

Inhalt dieser Vorlesung

- Verteilung
- Autonomie
- Heterogenität
- Transparenz

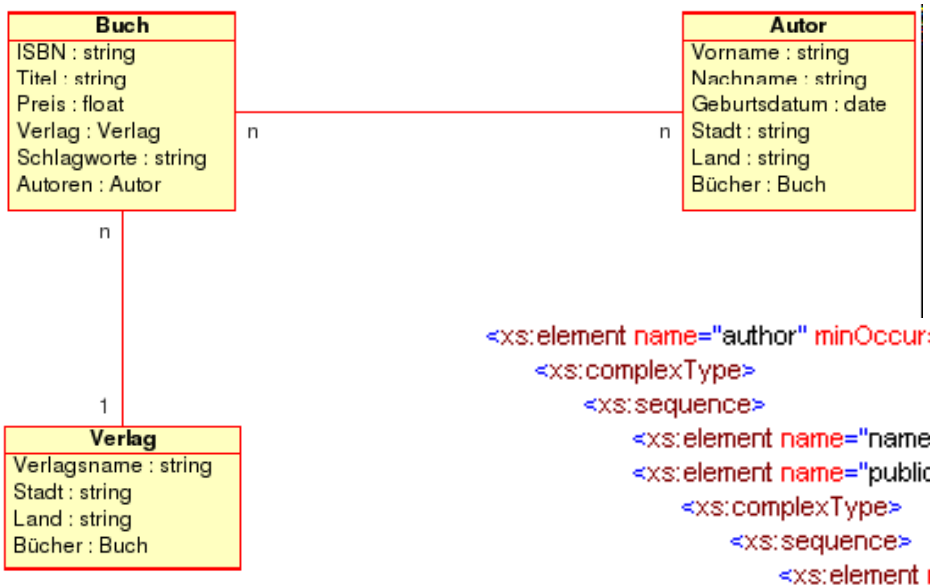
Autonomie

- Der Grad, zu dem verschiedene Quellsysteme unabhängig voneinander betrieben werden
  - Unabhängige Zugriffskontrolle und -methoden, Konfiguration, Modellierung, ...
  - Meint nicht „Unabhängigkeit“ im Sinne von Stromversorgung, Vernetzung ...
  - Beinhaltet die Freiheit, einmal getroffene Entscheidungen jederzeit zu ändern
- Arten von Autonomie (nicht disjunkt) [ÖV99]
  - Designautonomie
  - Kommunikationsautonomie
  - Ausführungsautonomie

Designautonomie (auch: Entwurfsautonomie)

- Freiheit einer Quelle bezüglich
  - Datenmodell (Relational, hierarchisch, XML, ...)
  - Schema
    - Abdeckung der Domäne (universe of discourse)
    - Benennung
    - Schemaentwurf
    - ...
  - Gestaltung der Schnittstellen (Funktionsnamen, Parameter, Ergebnistypen, Rückgabeformat, ...)

Beispiel




```
<xs:element name="author" minOccurs="0" maxOccurs="unbounded">
    <xs:complexType>
        <xs:sequence>
            <xs:element name="name" type="xs:string"/>
            <xs:element name="publication" minOccurs="0" maxOccurs="unbounded">
                <xs:complexType>
                    <xs:sequence>
                        <xs:element name="title" type="xs:string"/>
                        <xs:element name="year" type="xs:string"/>
                        <xs:element name="booktitle" type="xs:string" minOccurs="0"/>
                        <xs:element name="journal" type="xs:string" minOccurs="0"/>
                    </xs:sequence>
                </xs:complexType>
            </xs:element>
        </xs:sequence>
    </xs:complexType>
</xs:element>
```

Kommunikationsautonomie

- Freiheit einer Quelle bezüglich
  - Mit welchen Systemen kommuniziert wird
    - Sperren von Clients nach 1200 Zugriffen?
    - Nur bestimmte IP-Räume zulassen?
  - Wann mit anderen Systemen kommuniziert wird
    - Jederzeit Eintritt/Austritt aus integriertem System
    - Zugriff nur in „ruhigen“ Zeiten
    - Priorisierung von Zugriffen (Quality of Service)
  - Was (welche Informationen) kommuniziert wird
  - Welche Anfragemöglichkeiten zur Verfügung gestellt werden
    - Anfragesprachen, API, Web Services, ...

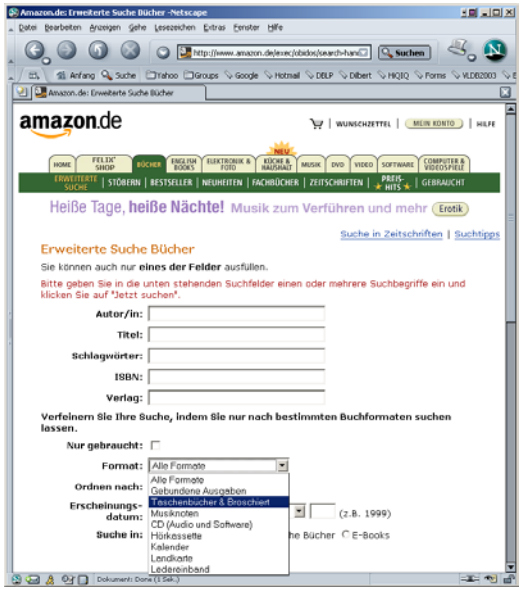
Beispiele

- Voller SQL Zugang
  - z.B. via JDBC
  - Beliebige Queries
    - SQL Injection, Ressourcenverbrauch
  - Lesend (und Schreibend?)
  - Schemaveränderungen?

```
WITH RECURSIVE TOPIC_PATHS (topic_above, topic_below, level) AS
(SELECT topic_id, topic_id, level FROM TOPIC
UNION ALL
SELECT TOPIC_PATHS.topic_above, TOPIC.topic_id, TOPIC_PATHS.level
FROM TOPIC_PATHS, TOPIC
WHERE TOPIC_PATHS.topic_below = TOPIC.topic_above)

SELECT TOPIC_PATHS.topic_above, DistinctCount (CLASSIFY.Doc_ID)
FROM TOPIC_PATHS, CLASSIFY
WHERE TOPIC_PATHS.topic_below = CLASSIFY.Topic_ID AND
TOPIC_PATHS.level = 1
GROUP BY TOPIC_PATHS.topic_above;
```

- HTML Formular
  - Nur ein (oder mehr) Suchfelder
  - Antwort als HTML Text
  - Nur vordefinierte Daten sichtbar



Ausführungsautonomie

- Freiheit einer Quelle bezüglich
  - Wann Anfragen ausgeführt werden (Scheduling)
  - Wie Anfragen ausgeführt werden (Optimierung)
  - Behandlung von Transaktionen (gibt‘s die?)
  - Beteiligung an globalen Transaktionen

Weitere Arten von Autonomie

- Schnittstellenautonomie
  - Festlegung, wie auf die Daten technisch zugegriffen werden kann
  - Web Services und SOAP, Web und HTML, SQL und Tupel, ...
- Zugriffsautonomie
  - Festlegung, wer auf die Daten zugreifen darf
  - Authentifizierungsprotokoll, Accounts, Gruppenrechte
- Juristische Einschränkung der Verwendbarkeit
  - Zugriff für bestimmte Verwendungen kann verboten werden
    - Nur akademisch, nur mit Referenz, kein Text Mining, ...
  - Bestimmte Zugriffsarten können verboten werden
    - Kein Screen-Scraping, nicht mehr als 10 Anfragen / Sekunde, ...

Probleme

- Änderungen bei vielen Quellen
  - 12 Quellen, eine Änderung pro Quelle pro Jahr
  - Ergibt eine Änderung pro Monat im Integrationssystem
- Autonomie ist oft unvermeidlich
  - Unabhängige Unternehmen / Unternehmensteile
  - Innerhalb von Unternehmen: Andere Administratoren, andere Abteilung, andere Anforderungen
    - Lokale Erweiterungen, andere Einkaufsprozesse, Besonderheiten in Abteilungen etc.
    - Soziologisch-psychologische Aspekte: Macht
- Gegenbewegung: Standardisierung
  - Vereinheitlichung im und über Unternehmen hinweg

Softwareentwickler lieben Autonomie

- Das vermeintliche Recht, alles dauernd zu ändern
- „Not invented here“ Syndrom
- Standards grenzen Autonomie ein

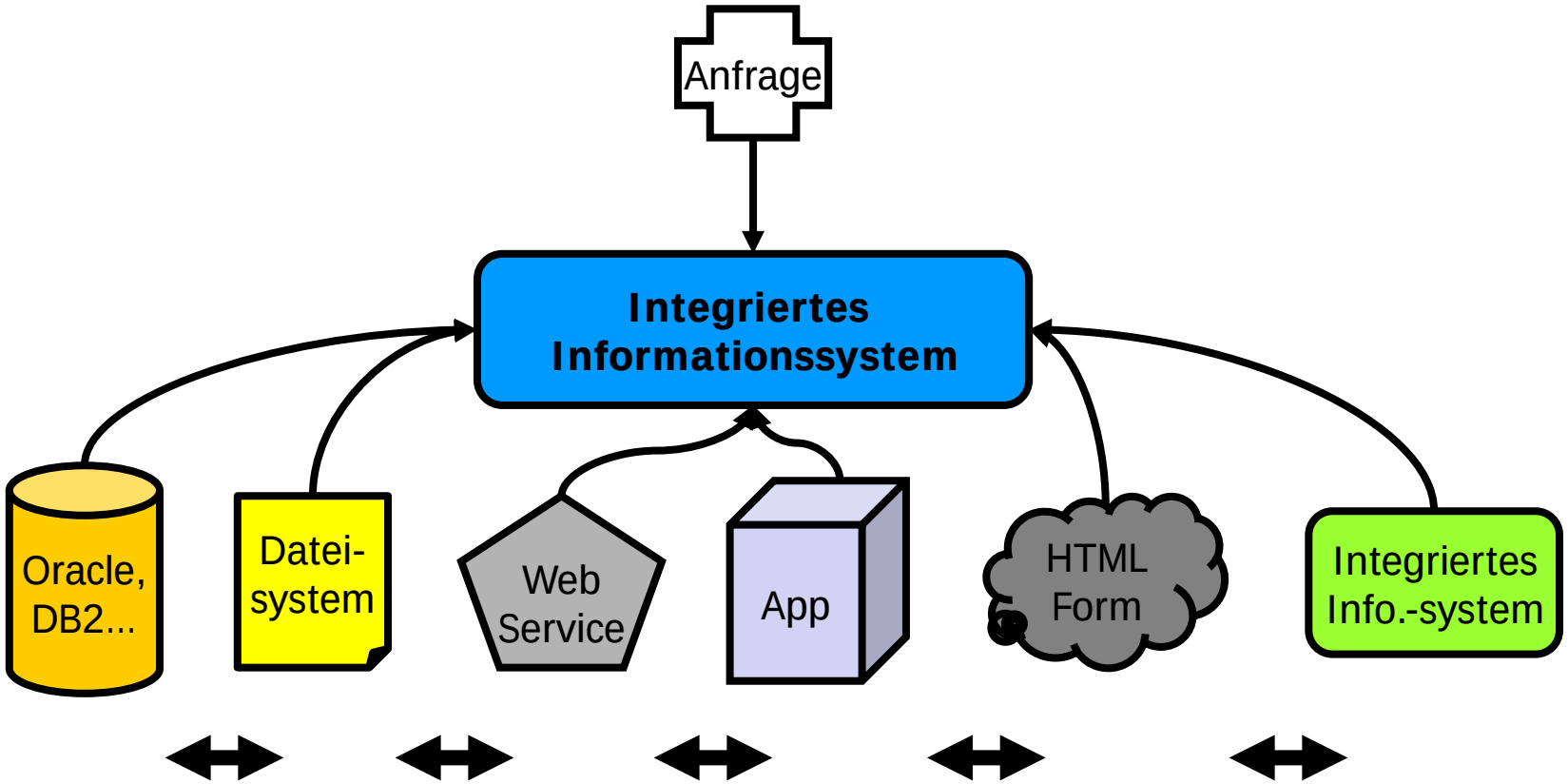
Inhalt dieser Vorlesung

- Verteilung
- Autonomie
- Heterogenität
  - Technische Heterogenität
  - Syntaktische Heterogenität
  - Datenmodellheterogenität
  - Strukturelle Heterogenität
    - Schematische Heterogenität
  - Semantische Heterogenität
- Transparenz

Heterogenität

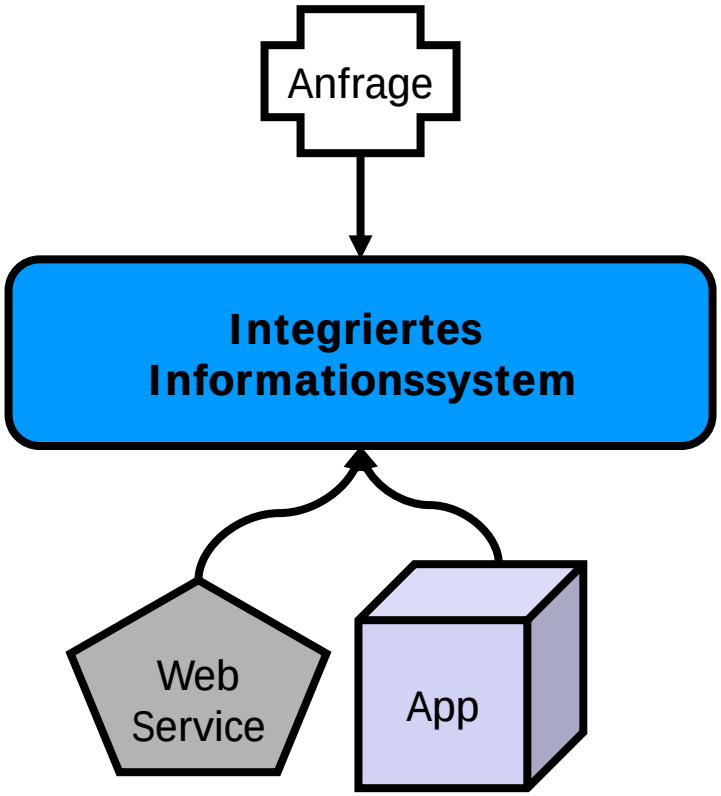
- Zwei Informationssysteme sind heterogen, wenn sie sich irgendwie unterscheiden
- Verschiedene Ausprägungen von „irgendwie“ => verschiedene Arten von Heterogenität
- Informationsintegration beinhaltet Verbergen von Heterogenität
  - Materialisierung: Erstellung eines homogenen Systems
  - Virtuelle Integration: Erweckung des Anscheins eines homogenen Systems

Zwischen Datenquellen



Heterogenität zwischen Datenquellen

Zwischen Schichten



Heterogenität zwischen Integrationsschicht und Datenquellen

Inhalt dieser Vorlesung

- Verteilung
- Autonomie
- Heterogenität
  - Technische Heterogenität
  - Syntaktische Heterogenität
  - Datenmodellheterogenität
  - Strukturelle Heterogenität
    - Schematische Heterogenität
  - Semantische Heterogenität
- Transparenz

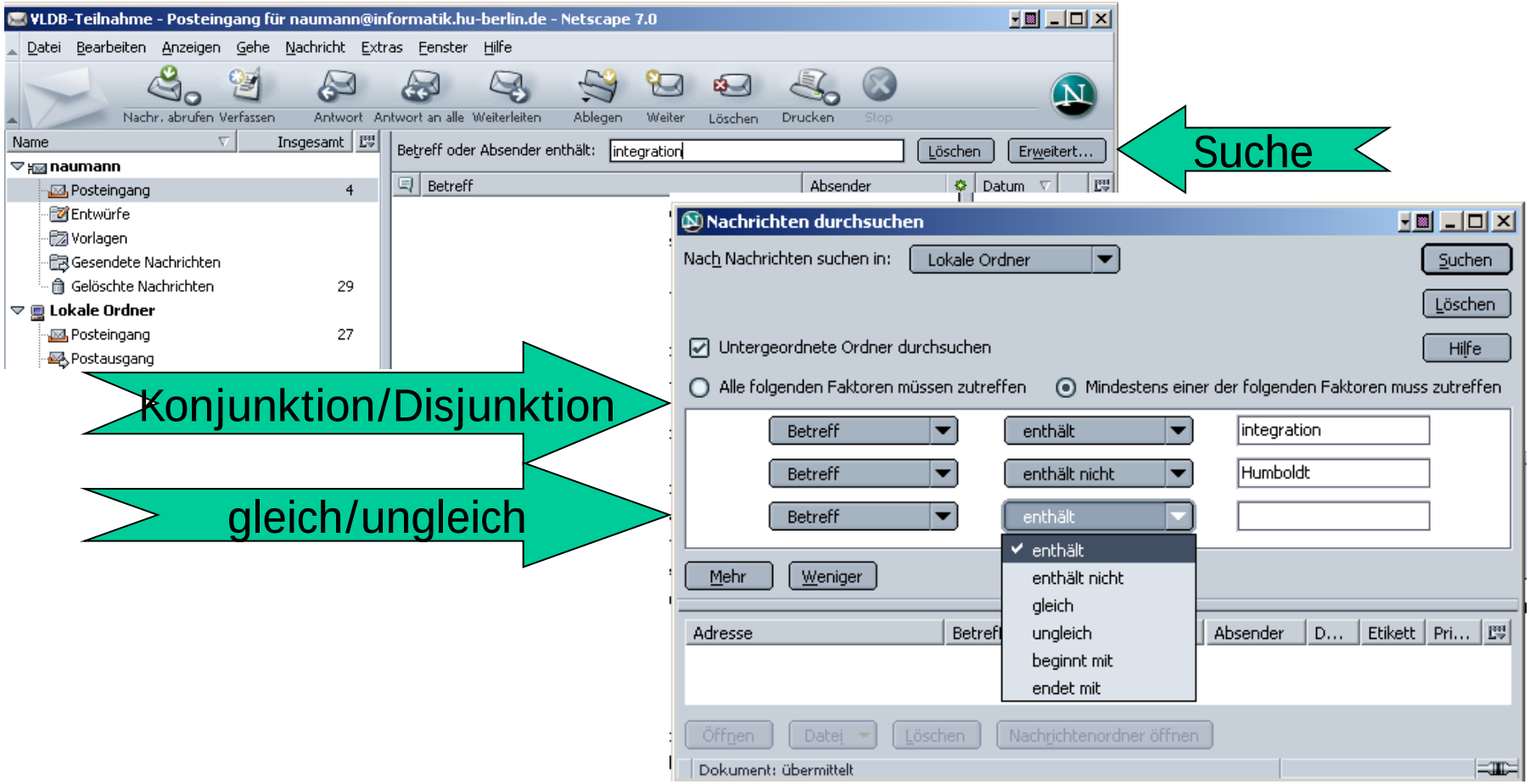
Technische Heterogenität

Ebene	Mögliche Ausprägungen
Anfragemöglichkeit	Anfragesprache, parametrisierte Funktionen, Formulare (engl. *canned queries*)
Anfragesprache	SQL, XQuery, Volltextsuche
Austauschformat	Binärdaten, XML, HTML, tabellarisch
Kommunikationsprotokoll	HTTP, JDBC, SOAP

Gründe für Einschränkungen

- Komplexität
  - Negation / Ungleichheit: teuer
  - „=“ oder auch „>,<, ≥, ≤“
  - Konjunktion / Disjunktion (eher teuer)
- Einfache Schnittstellen sind leicht verständlich und einfacher zu entwickeln und zu unterhalten
- Technische Gründe
  - Maximale Länge bei GET Kommandos
  - Maximale Länge von Queries
- Stabilität / Sicherheit
  - Teure Queries vermeiden
  - SQL-Injektion vermeiden

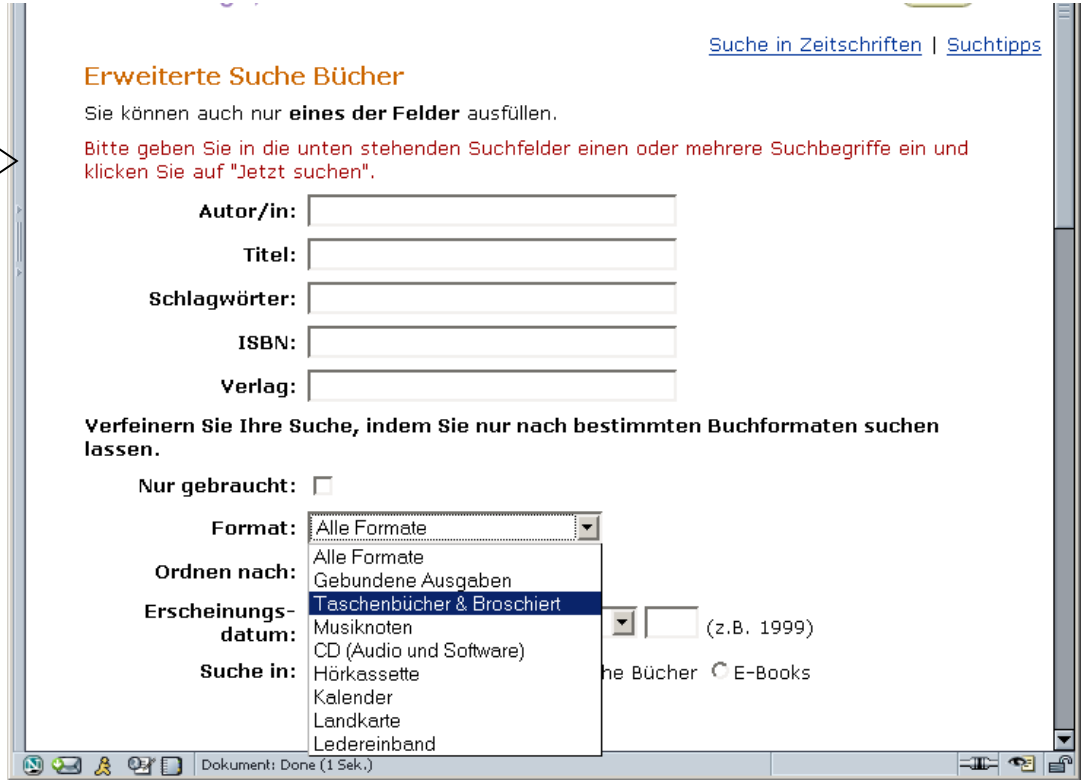
Beispiel


VLDB-Teilnahme - Posteingang für naumann@informatik.hu-berlin.de - Netscape 7.0
Datei Bearbeiten Anzeigen Gehe Nachricht Extras Fenster Hilfe
Nachr. abrufen Verfassen Antwort Antwort an alle Weiterleiten Ablegen Weiter Löschen Drucken Stop
Name Insgesamt
naumann
Posteingang 4
Entwürfe
Vorlagen
Gesendete Nachrichten
Gelöschte Nachrichten 29
Lokale Ordner
Posteingang 27
Postausgang
Betreff oder Absender enthält: integration
Löschen
Erweitert...
Betreff
Absender
Datum
Suche
Konjunktion/Disjunktion
gleich/ungleich
Nachrichten durchsuchen
Nach Nachrichten suchen in: Lokale Ordner
Suchen
Löschen
Hilfe
Untergeordnete Ordner durchsuchen
Alle folgenden Faktoren müssen zutreffen
Mindestens einer der folgenden Faktoren muss zutreffen
Betreff enthält integration
Betreff enthält nicht Humboldt
Betreff enthält
enthält
enthält nicht
gleich
ungleich
beginnt mit
endet mit
Mehr
Weniger
Adresse
Absender
D...
Etikett
Pri...
Öffnen
Datei
Löschen
Nachrichtenordner öffnen
Dokument: übermittelt

Beispiel

Gebundene Variablen, vorgegebene Disjunktion

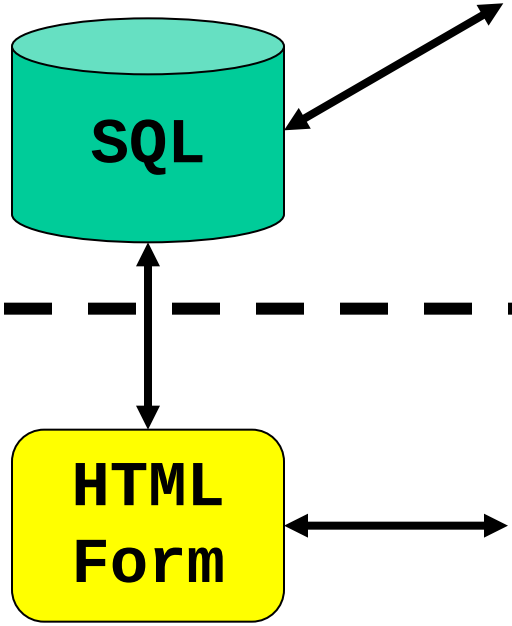
Feste Auswahl von Werten, vorgegebene Konjunktion



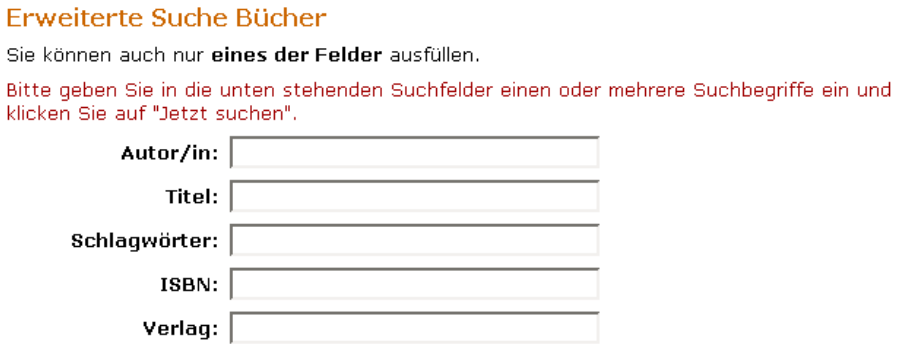
Typische Probleme

- Globale Anfragesprache ist mächtiger als lokale Anfragesprache
  - Anfragen nicht ausführbar oder
  - globales System muss kompensieren
- Lokale Anfragesprache ist mächtiger als globale Anfragesprache
  - Verpasste Chance, lokale (effizientere) Ausführung auszunutzen
- Einschränkungen bzgl. Variablenbindung sind inkompatibel
  - Anfragen eventuell nicht ausführbar oder kompensieren
- Übersetzung von Anfragesprachen notwendig
  - SQL – XQuery, SQL – HTTP/GET, Web-Service – SQL, etc.
  - Oft nicht einfach semantikerhaltend möglich
    - Behandlung von NULL; outer-joins; Substringsuche; ...

Mächtigere globale Anfragesprache




```
SELECT      *
FROM        Books
WHERE       Author = „Defoe“
AND         PubYear = 1979
```



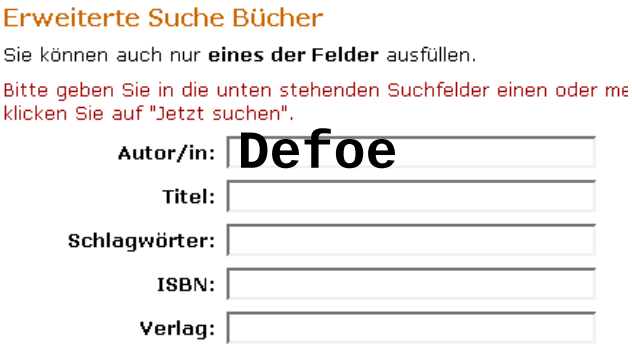
Kompensation

```
SELECT *
FROM   Books
WHERE  Author = „Defoe“
AND    PubYear = 1979
```

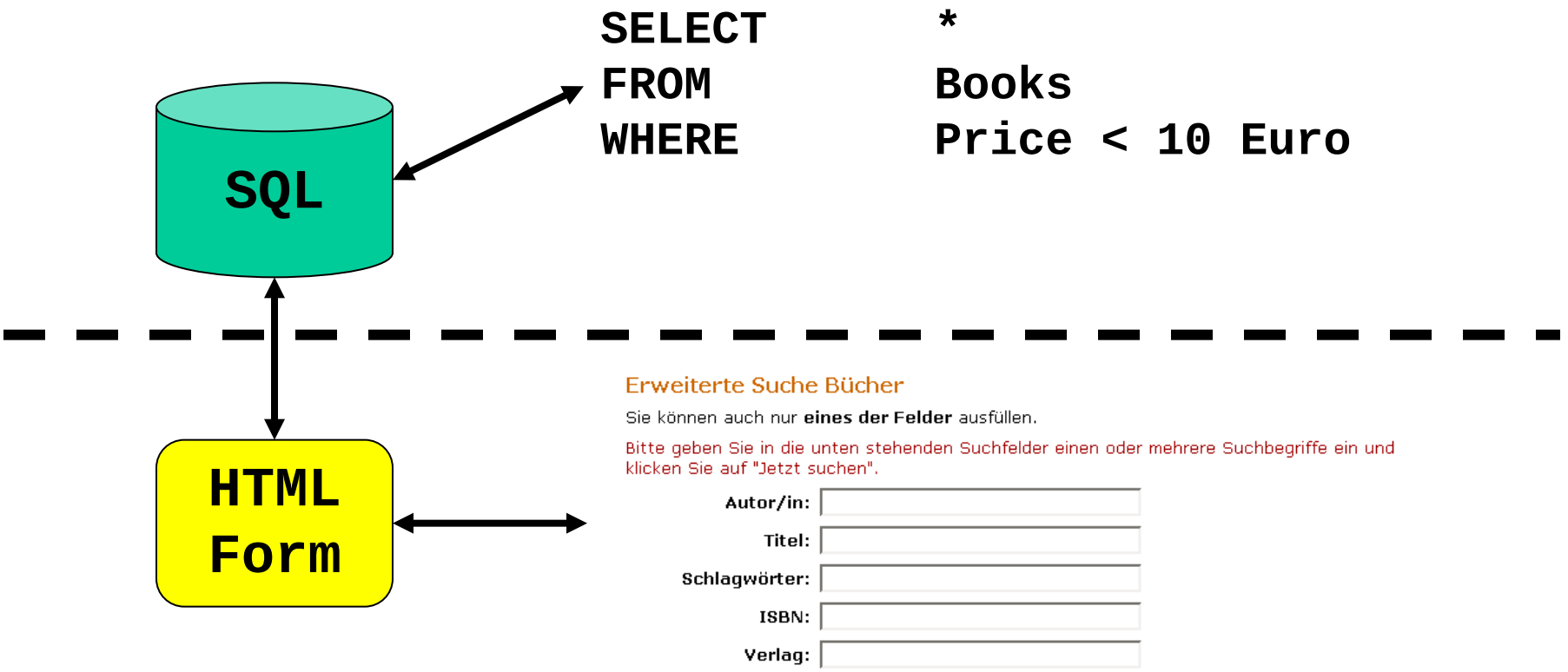
```
Daniel Defoe, Robinson Crusoe, 1979
```

PubYear = 1979

```
Daniel Defoe, Robinson Crusoe, 1986
Daniel Defoe, Robinson Crusoe, 1979
Daniel Defoe, Moll Flanders,   1933
```



Kompensation nicht möglich



Inhalt dieser Vorlesung

- Verteilung
- Autonomie
- Heterogenität
  - Technische Heterogenität
  - Syntaktische Heterogenität
  - Datenmodellheterogenität
  - Strukturelle Heterogenität
    - Schematische Heterogenität
  - Semantische Heterogenität
- Transparenz

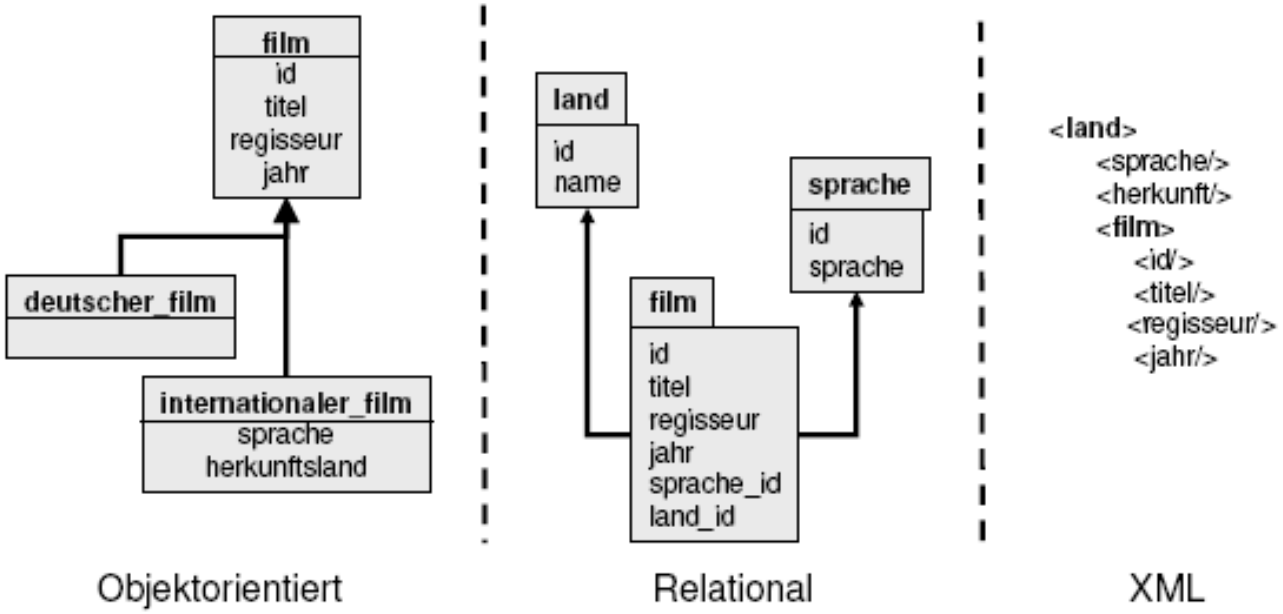
Syntaktische Heterogenität

- Unterschiedliche Darstellung desselben Sachverhalts
  - Dezimalpunkt oder –komma
  - Euro oder €
  - Comma-separated oder tab-separated
  - HTML oder ASCII oder Unicode oder ...
  - Notenskala 1-6 oder „sehr gut“, „gut“, ...
  - Binärcodierung oder Zeichen
  - Datumsformate (12. September 2006, 12.9.2006, 9/12/2006, ...)
- Überwindung in der Regel nicht problematisch
  - Umrechnung, Übersetzungstabellen, ...

Datenmodellheterogenität

- Typische Datenmodelle
  - Flach: CSV, EXCEL, ...
  - Relational: Tupel
  - Hierarchisch: XML, HTML, ...
  - Domänenspezifisch: EXPRESS, OPEN-GIS, ...
- Modell zum Austausch oder zur persistenten Datenhaltung
  - Black-Box-Sicht: Entscheidend ist, was die Quelle liefert (also Austauschformat)
- Konvertierung z.T. nicht semantikerhaltend möglich
  - m:n Relationen in XML?
  - Verschiedene Tabellen in eine CSV Datei?

Beispiel



„Fast“ dasselbe in verschiedenen Datenmodellen - Unterschiede?

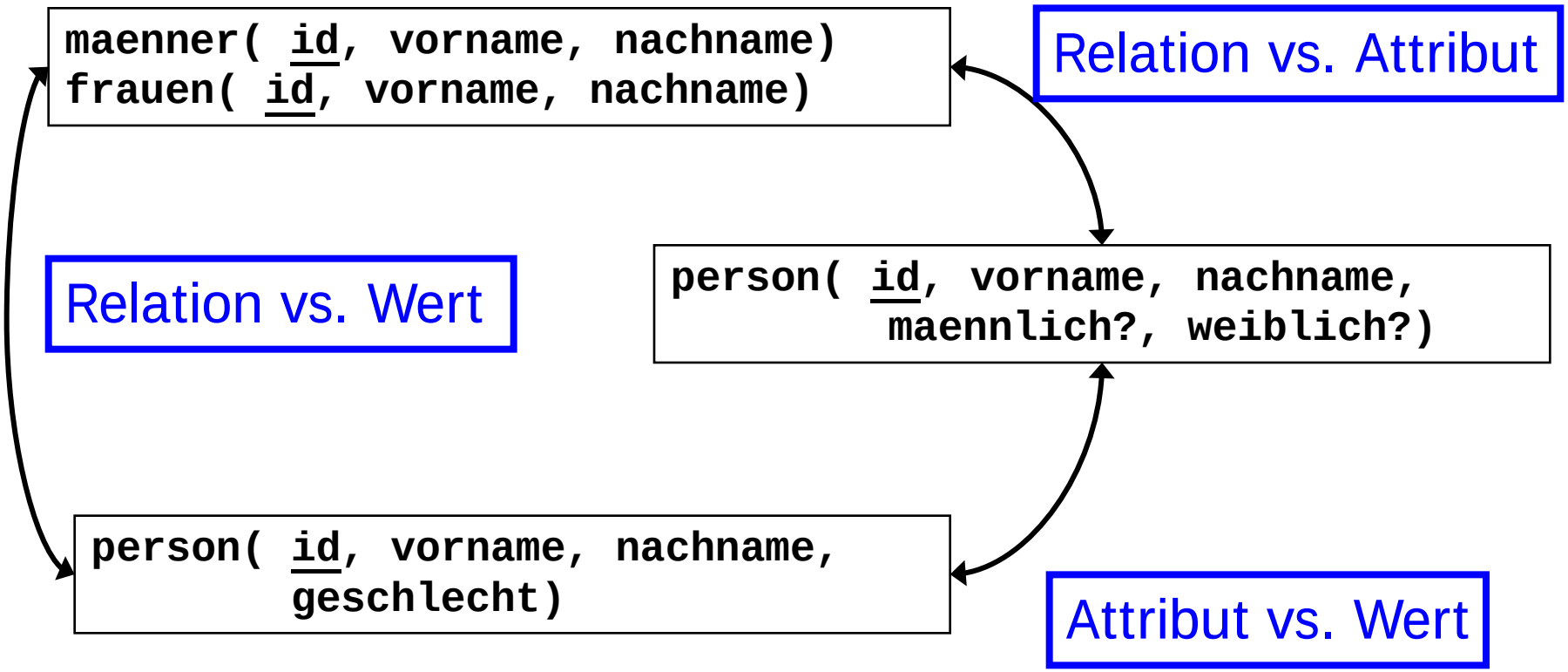
Inhalt dieser Vorlesung

- Verteilung
- Autonomie
- Heterogenität
  - Technische Heterogenität
  - Syntaktische Heterogenität
  - Datenmodellheterogenität
  - Strukturelle Heterogenität
    - Schematische Heterogenität
  - Semantische Heterogenität
- Transparenz

Strukturelle Heterogenität

- Gleiche Dinge in unterschiedlichen Schemata ausdrücken
  - Andere Aufteilung von Attributen auf Tabellen
  - Fehlende / neue Attribute (wenn Intension nicht betroffen ist)
- Inhärent bei Datenmodellheterogenität
- Sehr oft mit semantischer Heterogenität verbunden
- Spezialfall: Schematische Heterogenität
  - Verwendung anderer Elemente desselben Datenmodells
  - Besonderheit: Kann durch Anfragen idR nicht überwunden werden

Spezialfall: Schematische Heterogenität



Schematische Konflikte sind schwierig

Modellierung als Relationen

```
spielfilm      ( id, titel, laenge)
dokumentarfilm( id, titel, laenge)
```

```
SELECT 'dpielfilm', AVG(laenge)
FROM   spielfilm
UNION
SELECT 'doku', AVG(laenge)
FROM   dokumentarfilm
UNION
...
```

Modellierung als Attribute

```
film( id, titel, laenge, spielfilm, doku)
```

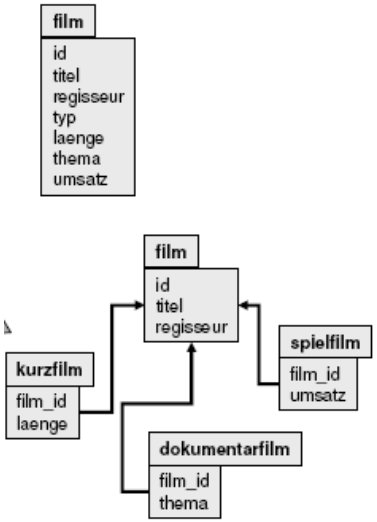
```
SELECT 'dpielfilm', AVG(laenge)
FROM   film
WHERE  spielfilm = TRUE
UNION
```

Modellierung als Attributwerte

```
film( id, titel, laenge, typ)
```

```
SELECT   typ, AVG(laenge)
FROM     film
GROUP BY typ;
```

Integrierte Sicht

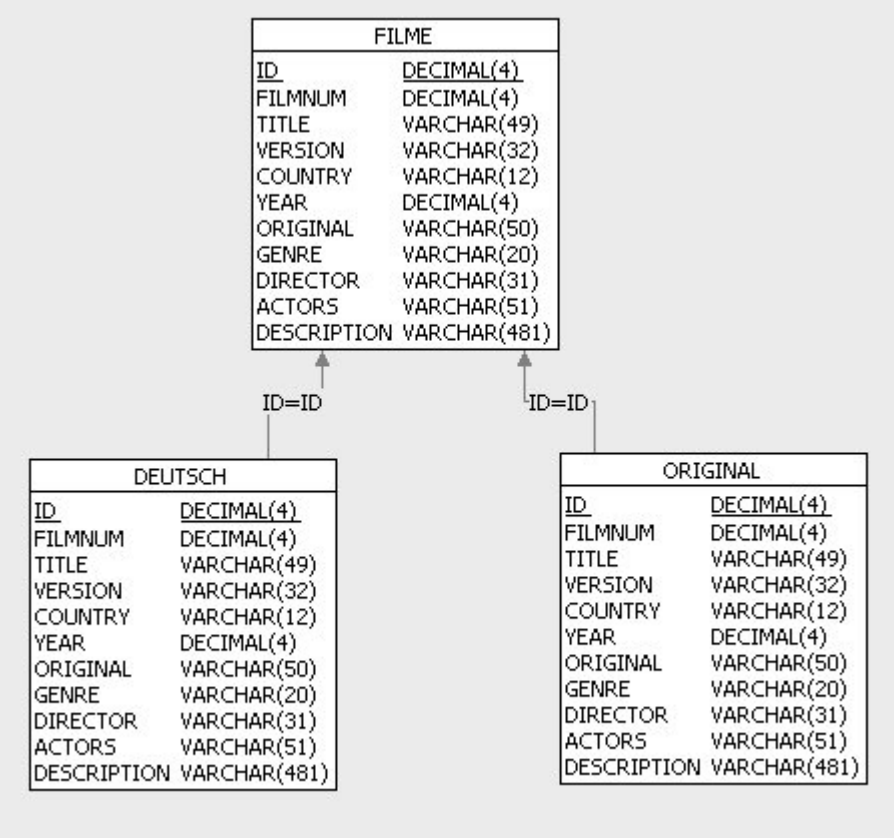



```
CREATE VIEW q1_q2
SELECT id, titel, regisseur, typ
FROM   q1.film
UNION
SELECT id, titel, regisseur, 'spielfilm' AS typ
FROM   q2.spielfilm
UNION
SELECT id, titel, regisseur, 'kurzfilm' AS typ
FROM   q2.kurzfilm
UNION
SELECT id, titel, regisseur, 'doku' AS typ
FROM   q2.dokumentarfilm;
```

- Sichtanpassung nötig, wenn neue Filmtypen vorliegen
  - Datenänderung bedingt Query-/Schemaänderung
  - Das will man vermeiden
  - Entscheidung, was Daten und was Metadaten sind

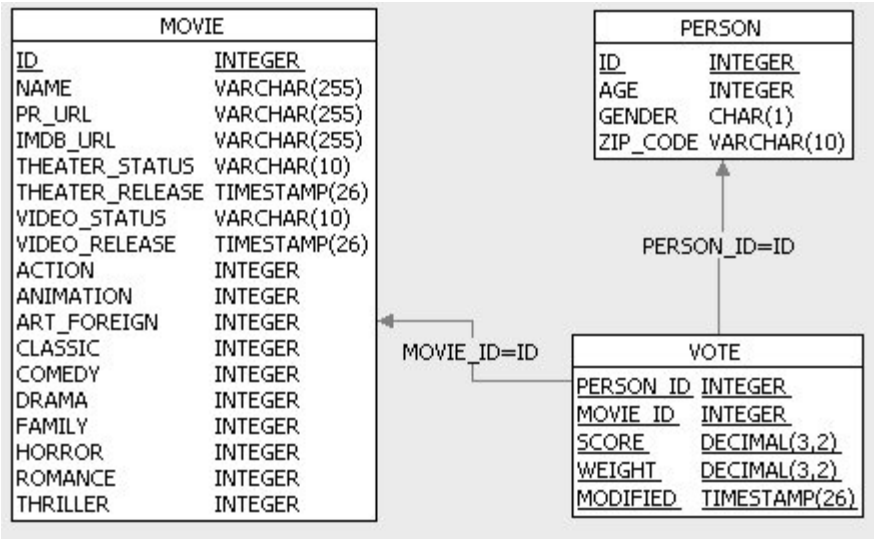
Exotische Probleme?

- Schema zur Speicherung von Filmen des Verleihers „XYZ“
  - **ACTORS** als **VARCHAR**
  - **ORIGINAL** – bedeutet was?
  - **TITLE**, **YEAR**, ... an drei Stellen
  - ID-Räume **DEUTSCH** und **ORIGINAL** getrennt?
  - **GENRE** ist **VARCHAR**
  - ...



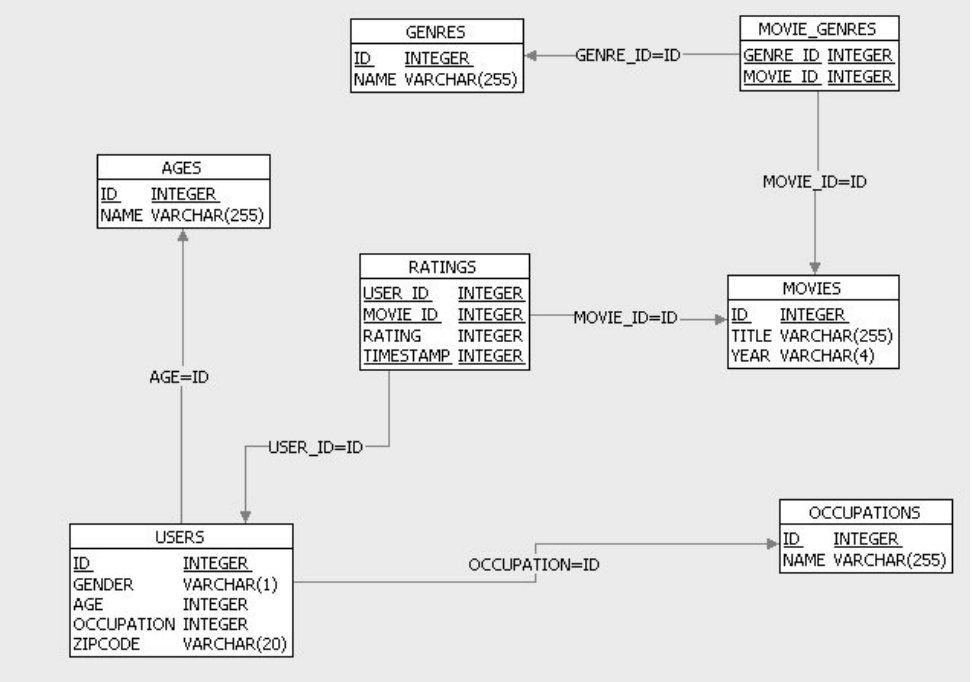
Exotische Probleme?

- Schema von eachmovie (HP)
  - Eine einzige Tabelle für Filme
  - Zusätzliche Informationen über Benutzer des Web-Systems
  - Wenig Infos über Filme, aber Links zu weiteren Quellen
  - **GENRE** sind boolsche Attribute
  - ...



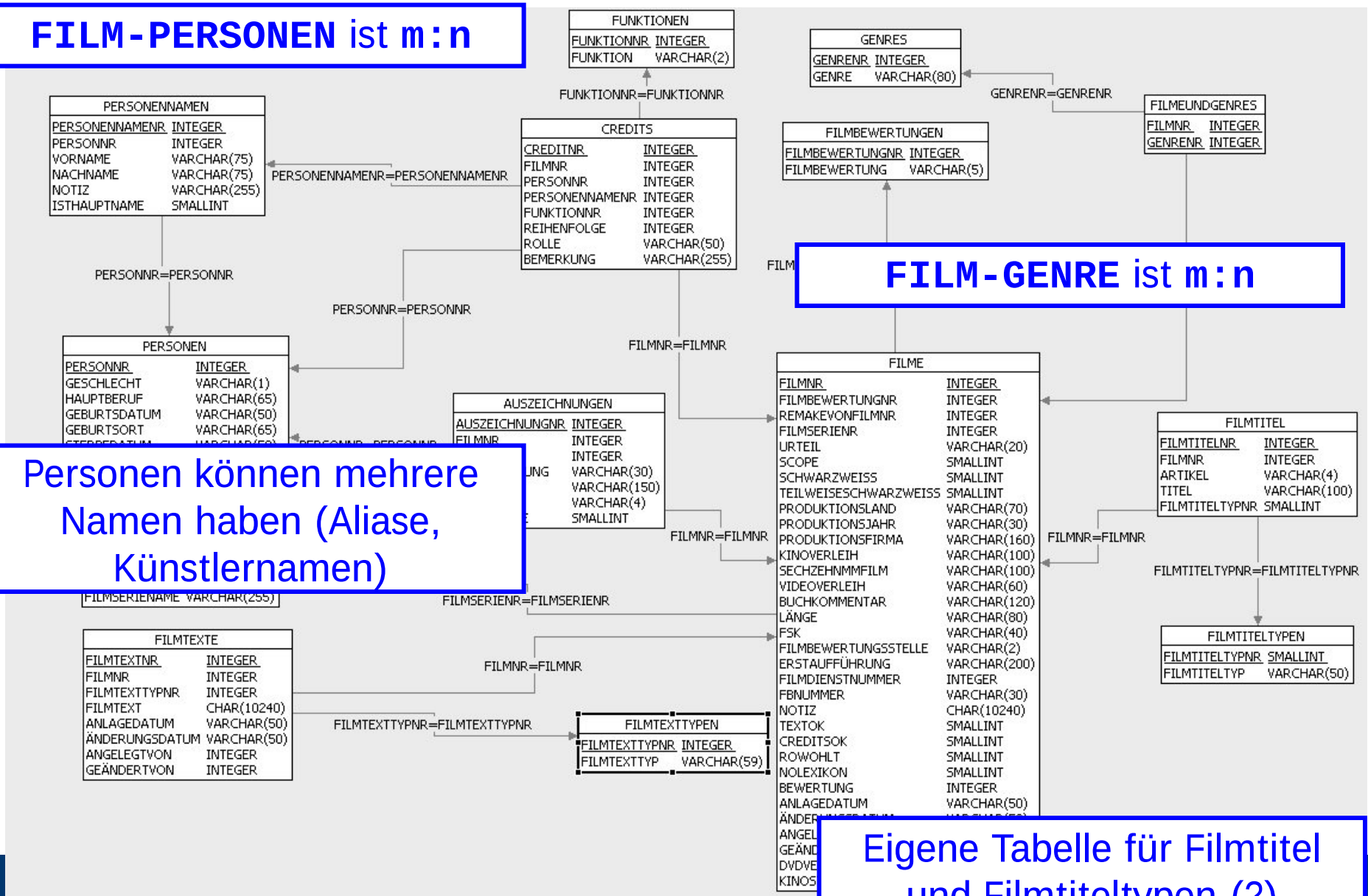
Exotische Probleme?

- Schema von movielens.umn.edu
  - Praktisch keine Informationen über Filme
  - Eigene Tabelle für **GENRE**
  - **FILM-GENRE** ist **m:n**
  - Mehr Informationen über Benutzer
  - Keine externen Links
  - ...



Schema des Filmdienst

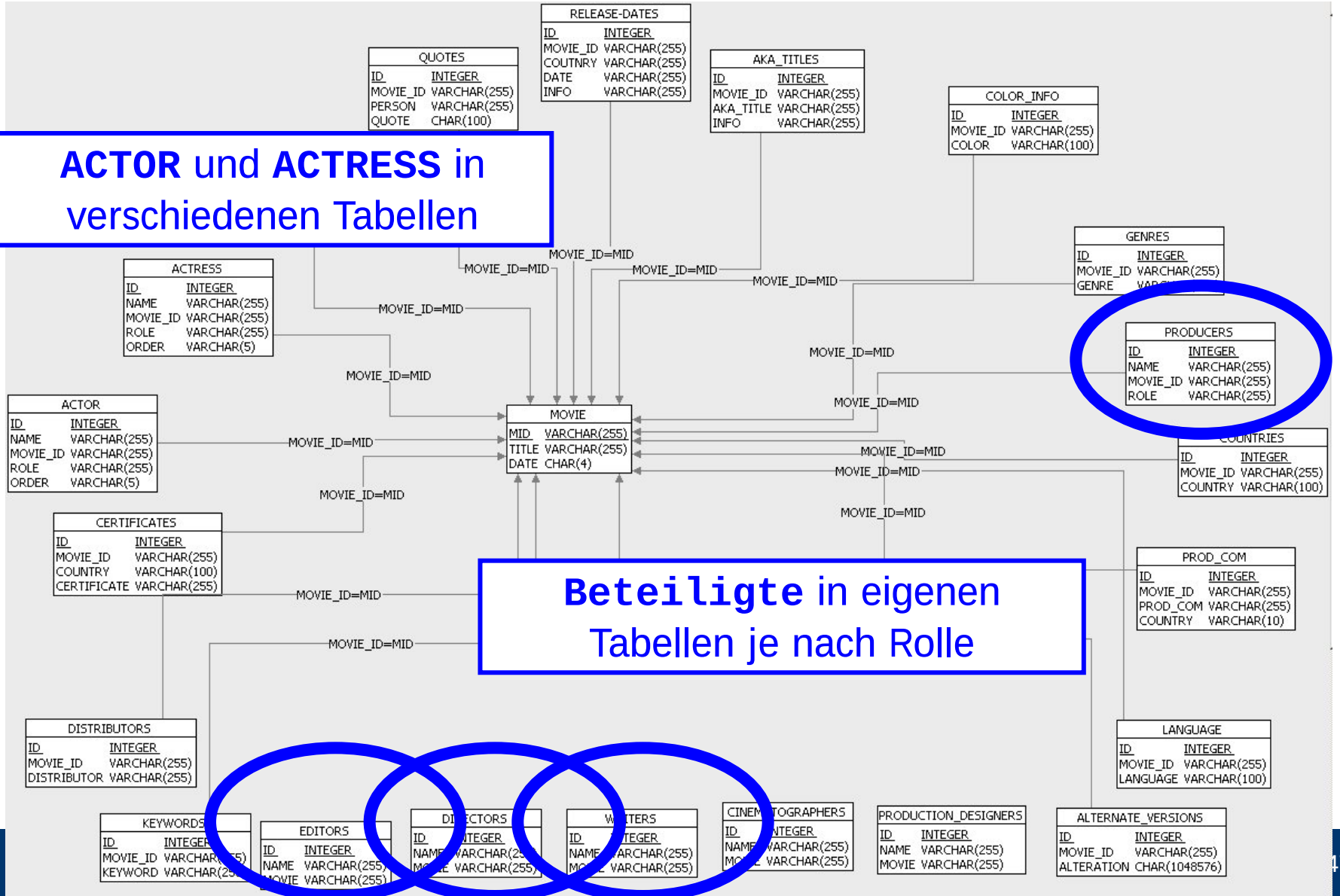
FILM-PERSONEN ist **m:n**



Personen können mehrere Namen haben (Aliase, Künstlernamen)

Eigene Tabelle für Filmtitel und Filmtiteltypen (?)

Schema der IMDB



Inhalt dieser Vorlesung

- Verteilung
- Autonomie
- Heterogenität
  - Technische Heterogenität
  - Syntaktische Heterogenität
  - Datenmodellheterogenität
  - Strukturelle Heterogenität
    - Schematische Heterogenität
  - Semantische Heterogenität
- Transparenz

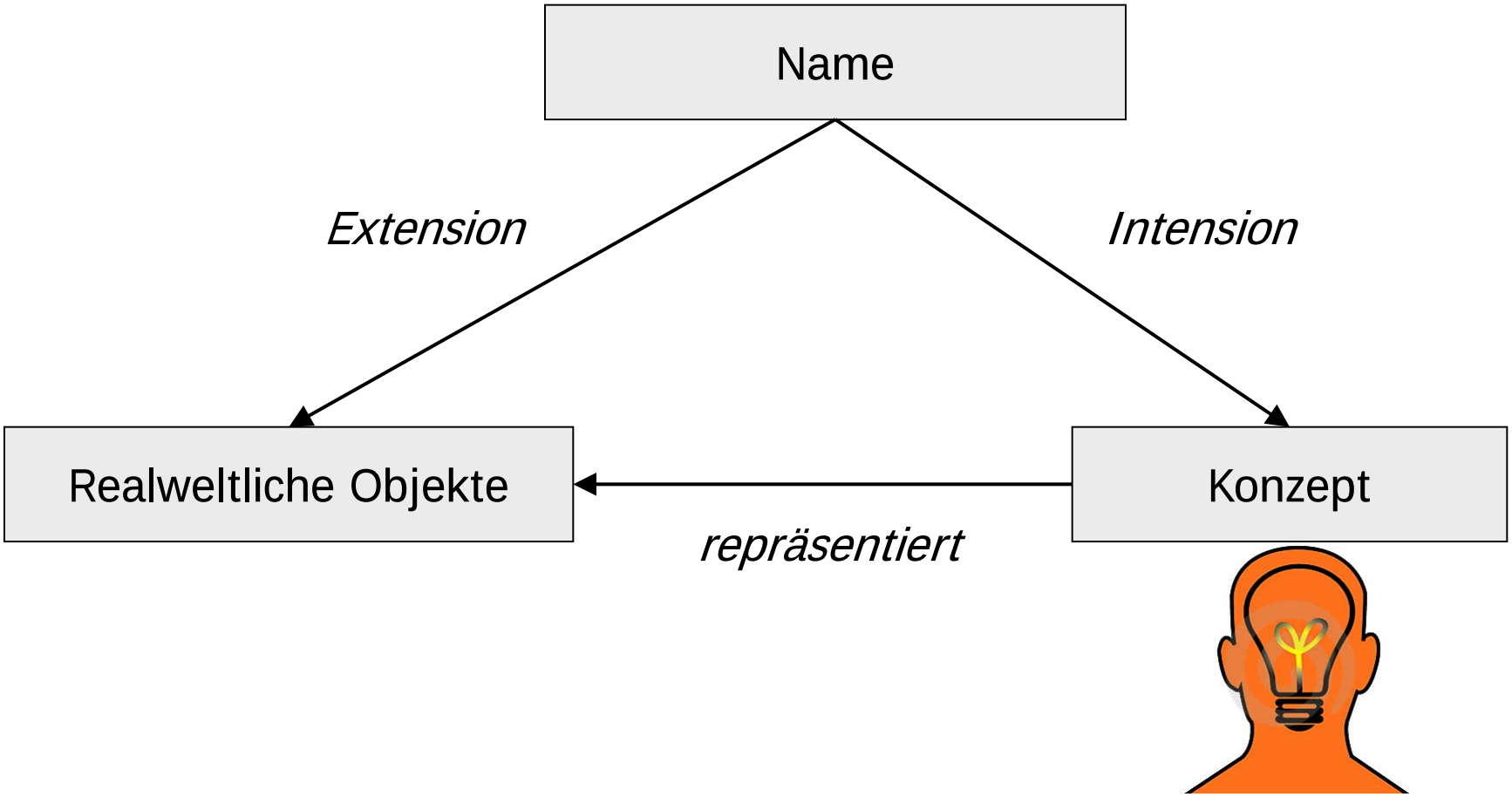
Semantik

- „Teilgebiet der Linguistik, das sich mit den Bedeutungen sprachlicher Zeichen und Zeichenfolgen befasst“
- „Bedeutung, Inhalt eines Wortes, Satzes oder Textes“
  - [Beides Fremdwörterduden]
- Programmiersprachen
  - Syntax: EBNF, Grammatiken
  - Semantik: Wirkungen der Ausführung; operationale Semantik, Fixpunktsemantik, ...
- Sprache
  - Syntaktisch falsch: „Ich esse Butterbrot ein“
  - Semantisch falsch: „Ich esse einen Schrank“

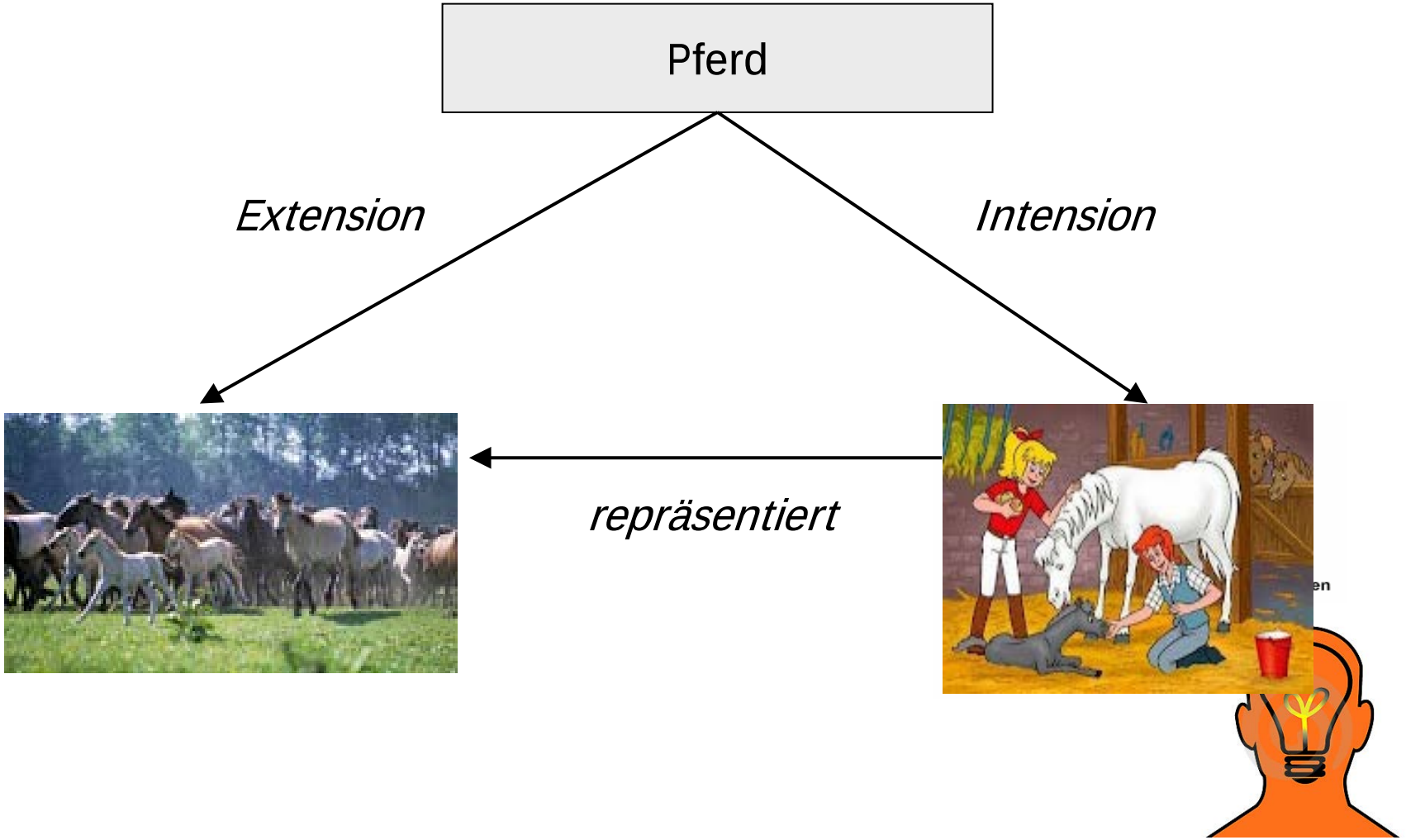
Semantik in dieser Vorlesung

- „Semantische Heterogenität ist ein überladener Begriff ohne klare Definition. Er bezeichnet die Unterschiede in Bedeutung, Interpretation und Art der Nutzung.“ [ÖV91]
- Wir meinen intuitive Semantik: Bedeutung (?) von
  - Modellelementen (selten - Metamodellierung)
  - Schemaelementen (meistens)
  - Daten / Werten

Semantik von was?

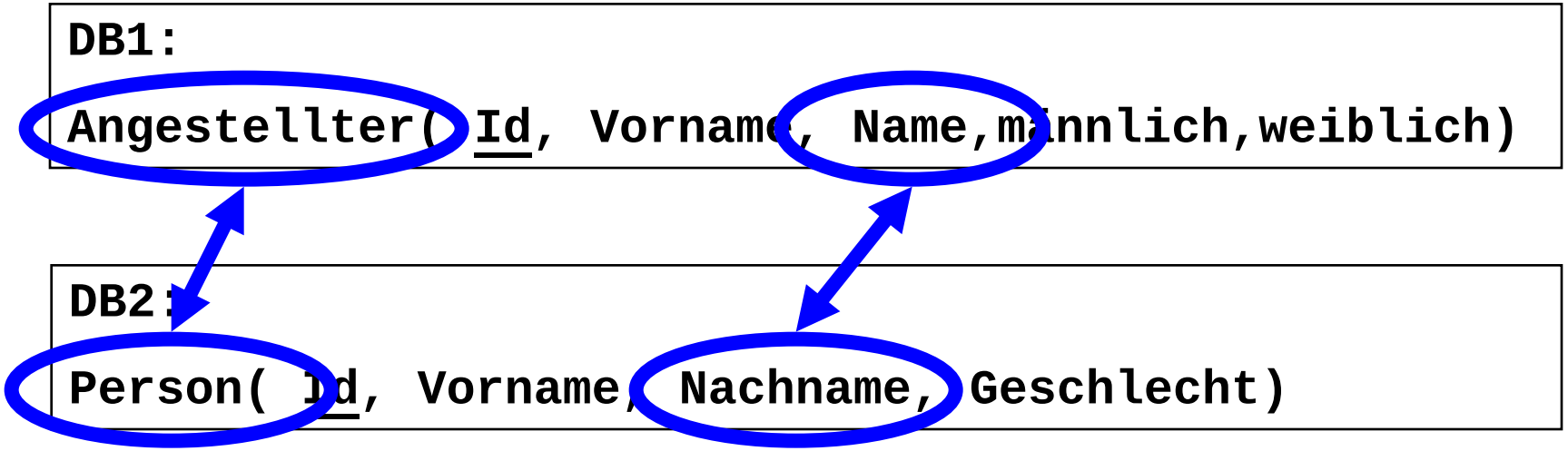


Beispiel



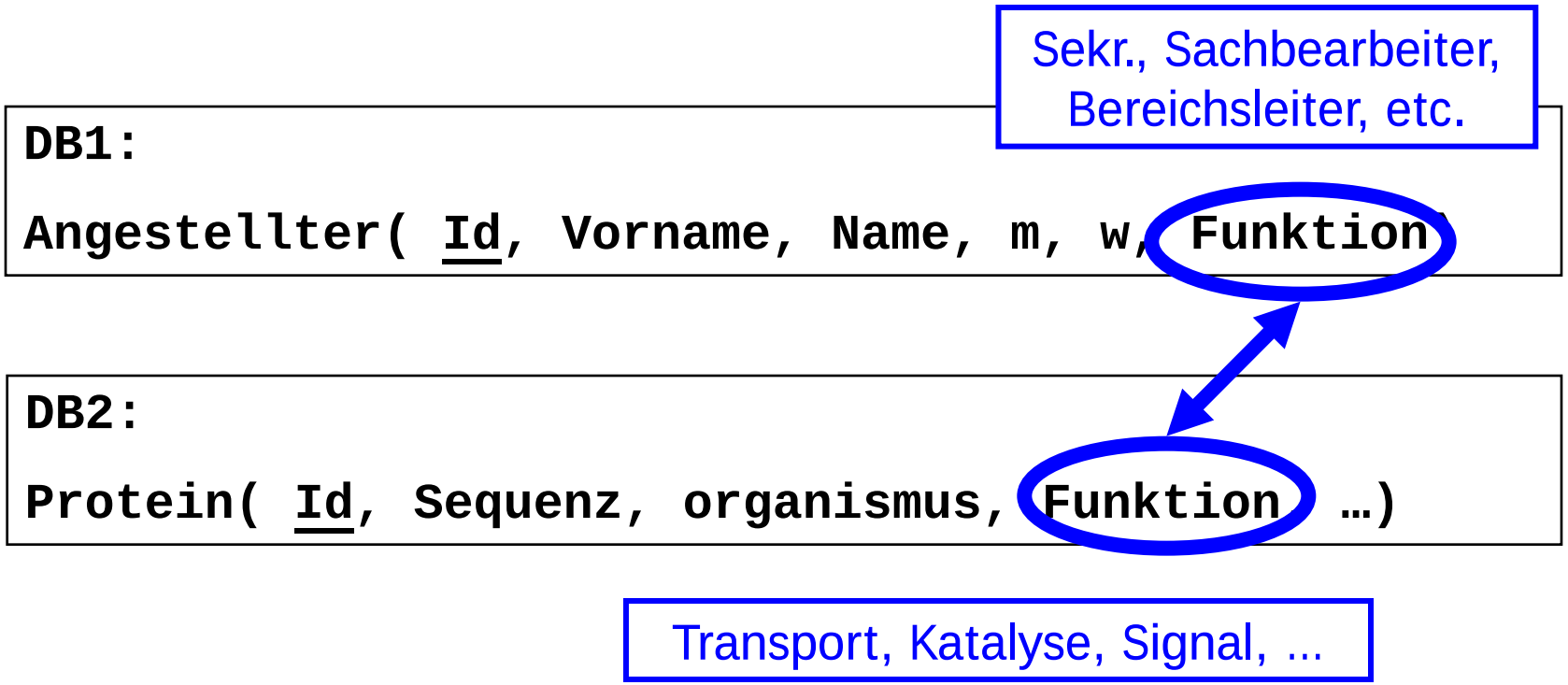
Synonyme

- Verschiedene Namen für dieselbe Menge
  - Immer im Kontext der Anwendung



Homonyme

- Gleiche Namen für verschiedene Mengen
  - Treten oft bei Überschreitung von Domänengrenzen auf



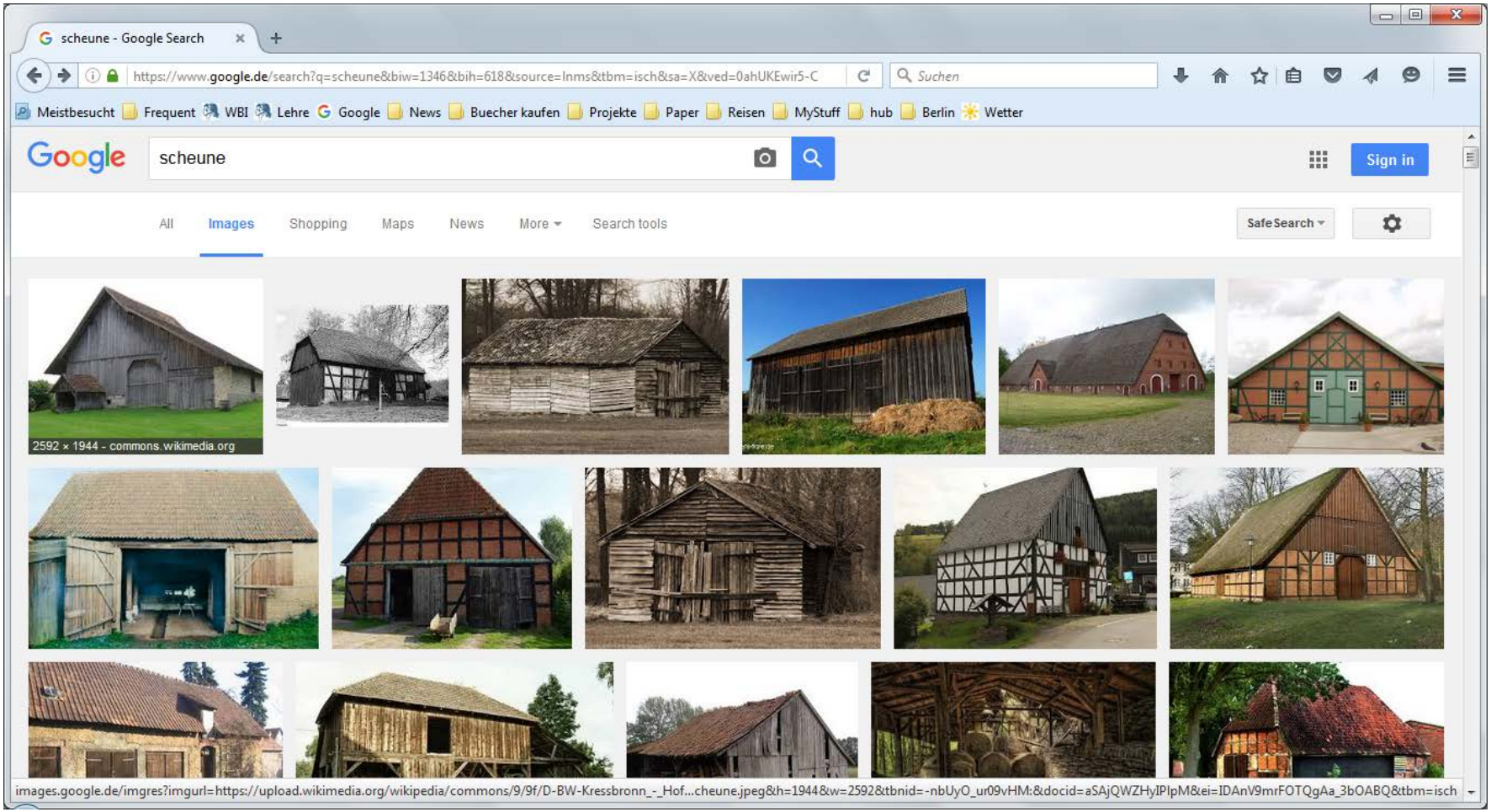
Konstellationen

- Seien a,b zwei Namen, a≠b, für die Mengen A, B
- A=B: a und b sind Synonyme
  - Kreditinstitut-Bank?, Schuppen-Scheune?
  - Gibt es echte Synonyme?
- A⊆B: b ist Hyperonym (Oberbegriff); a ist Hyponym zu b
  - Tochter ⊆ Kind
- $A \cap B \neq \emptyset \ \wedge A \neq B$ (Überlappung): Schwierigster Fall
  - Küche-Kochnische; Haus-Gebäude; Regisseur-Schauspieler
- $A \cap B = \emptyset$ (Disjunktion): a und b sind nicht verwandt
  - Schrank - Arbeitnehmer

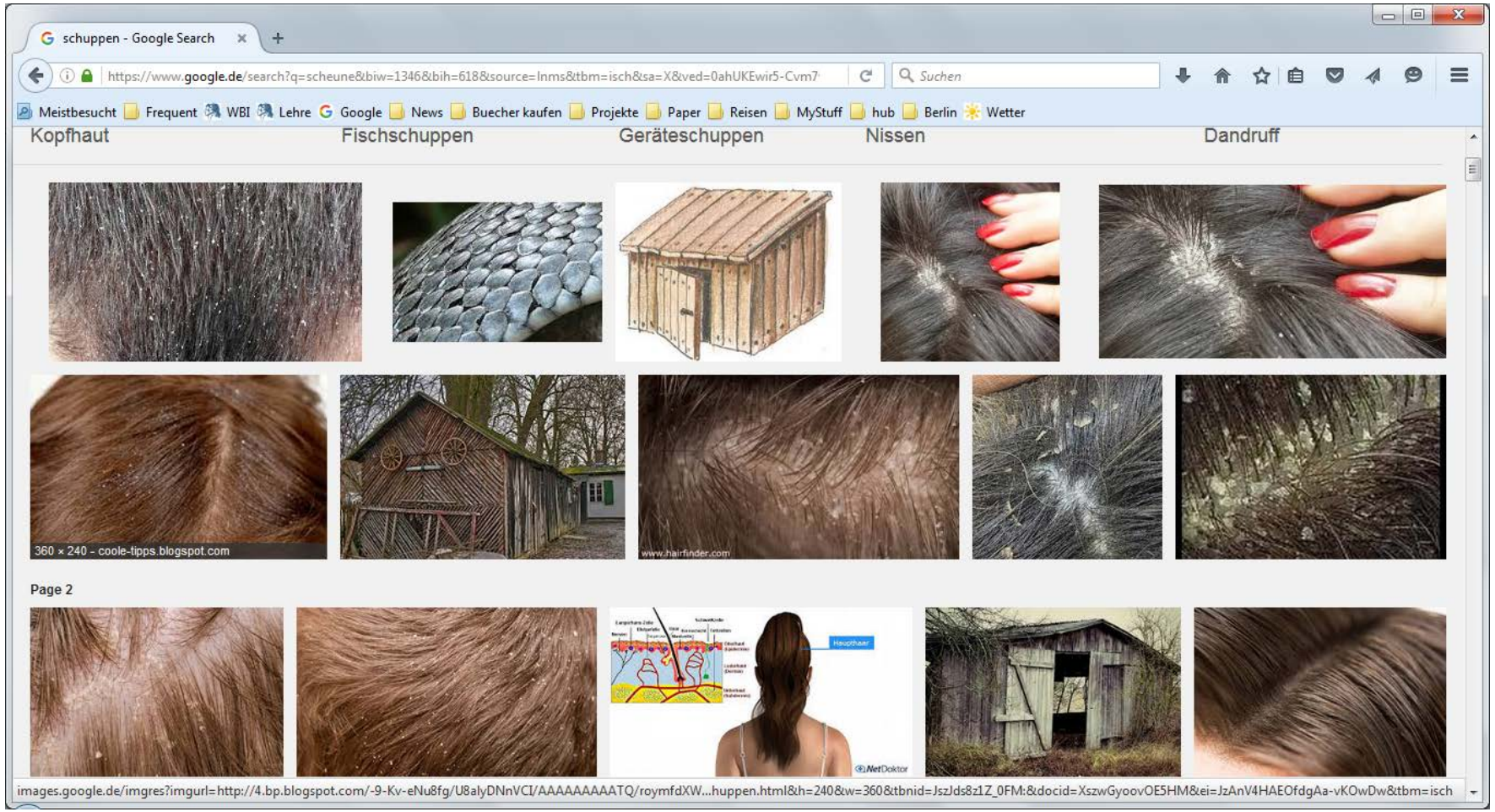
Weitere -nym Wörter

- Antonym: Verschiedene Namen, gegenteilige Semantik
  - Hell-dunkel, billig-teuer, ...
- Auto-Antonym: Gleiche Namen, gegenteilige Semantik
  - Transparenz
  - Left (bleiben, weggehen), clip (anhängen, ausschneiden)
- Heteronym: Gleiche Schreibung, verschiedene Aussprache, verschiedene Semantik
  - „It's the referee's job to record the new world record.“
  - Moderne Kunst – modernde Baumstämme
- Pseudonym
- ...
- http://www.fun-with-words.com/nym_words.html

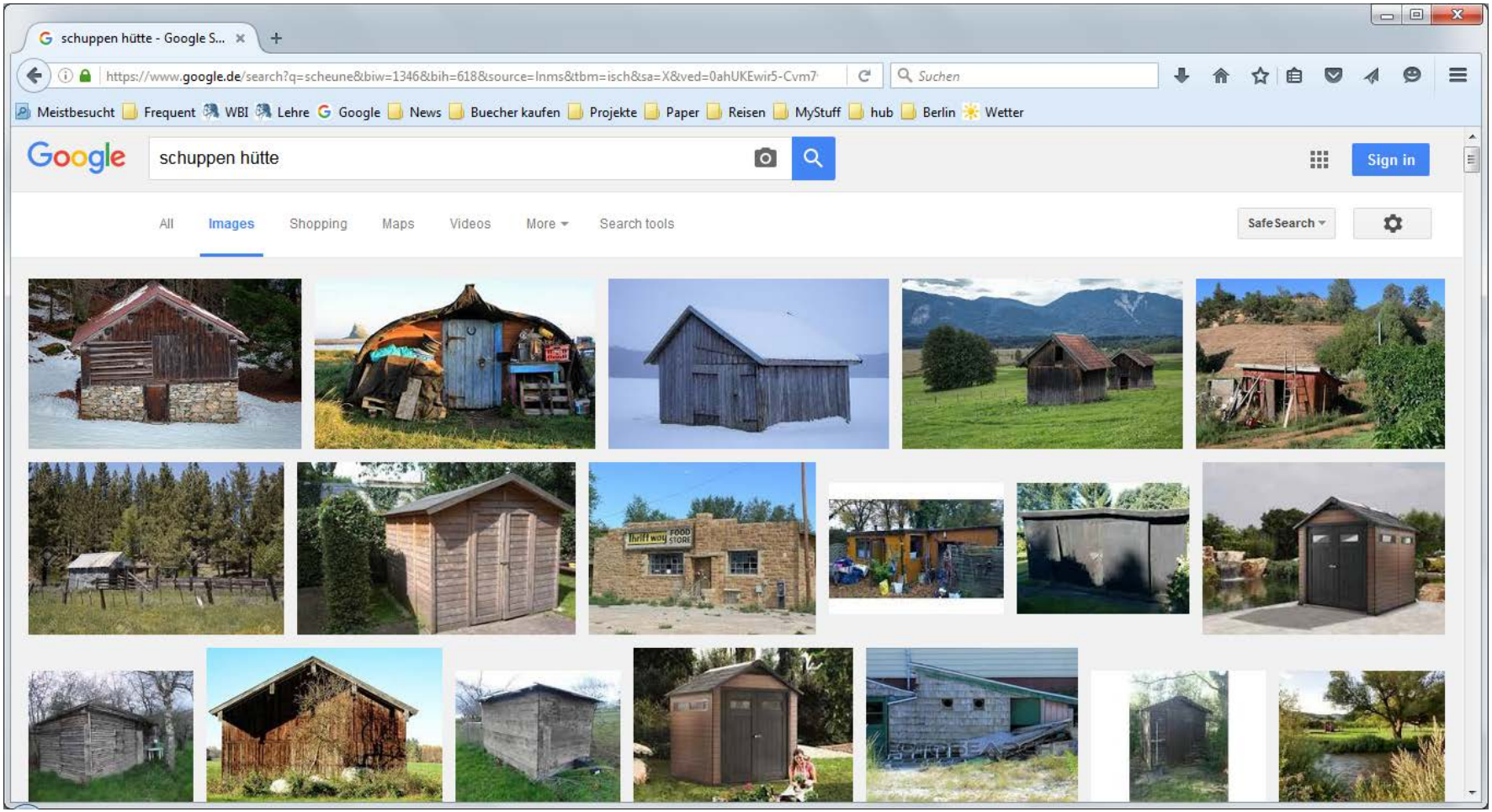
Scheune


scheune - Google Search
https://www.google.de/search?q=scheune&biw=1346&bih=618&source=lnms&tbm=isch&sa=X&ved=0ahUKEwir5-C
Suchen
Meistbesucht Frequent WBI Lehre Google News Buecher kaufen Projekte Paper Reisen MyStuff hub Berlin Wetter
Google
scheune
Sign in
All Images Shopping Maps News More Search tools
Safe Search
2592 × 1944 - commons.wikimedia.org
images.google.de/imgres?imgurl=https://upload.wikimedia.org/wikipedia/commons/9/9f/D-BW-Kressbronn_-_Hof...cheune.jpeg&h=1944&w=2592&tbnid=-nbUyO_ur09vHM:&docid=aSAjQWZHyIPlpM&ei=IDAnV9mrFOTQgAa_3bOABQ&tbm=isch

Schuppen


schuppen - Google Search
https://www.google.de/search?q=scheune&biw=1346&bih=618&source=lnms&tbm=isch&sa=X&ved=0ahUKEwir5-Cvm7
Suchen
Meistbesucht
Frequent
WBI
Lehre
Google
News
Buecher kaufen
Projekte
Paper
Reisen
MyStuff
hub
Berlin
Wetter
Kopfhaut
Fischschuppen
Geräteschuppen
Nissen
Dandruff
360 × 240 - coole-tipps.blogspot.com
www.hairfinder.com
Page 2
Haupthaar
images.google.de/imgres?imgurl=http://4.bp.blogspot.com/-9-Kv-eNu8fg/U8alyDNnVCI/AAAAAAAAATQ/roymfdXW...huppen.html&h=240&w=360&tbnid=JszJds8z1Z_0FM:&docid=XszwGyoovOE5HM&ei=JzAnV4HAEOfdgAa-vKOwDw&tbm=isch

Schuppen Hütte



Bedeutung: Woher?

- Schemaelemente sind nur Namen
- Was bestimmt die Semantik (Intension) eines Namens?
- Für Attributnamen
  - Datentyp
  - Constraints (Schlüssel, FK, unique, CHECK, ...)
  - Zugehörigkeit zu einer Relation
  - Andere Attribute dieser Relation
  - Beziehung der Relation zu anderen Relationen
  - Dokumentation
  - Vorhandene Werte
  - Wissen über den Anwendungsbereich
  - Zusammen: Kontext

Bedeutung kann sich ändern

- Synonyme?
  - Prince, „The artist formerly known as prince“
  - Ja – dieselbe Person
  - Nein – dieselbe Person zu verschiedenen Zeiten
  - Temporale Abhängigkeit
- Synonyme?
  - England, Großbritannien
  - Ja – für uns
  - Nein – für Schotten
  - Anhängig vom kulturellen Hintergrund

Kontext

- Semantik eines Namens hängt vom Kontext ab
- Beispiel
  - Unternehmen A: angestellte(...)
  - Unternehmen B: mitarbeiter(...)
  - Mitarbeiter und Angestellte kann man als Synonyme betrachten
  - Aber: A.angestellte ∩ B.mitarbeiter = ∅
    - Wenn Personen nicht in zwei Unternehmen beschäftigt sind
  - Erst bei einem Merger von A und B werden A.angestellte und B.mitarbeiter zu Synonymen
  - Sollten dann zu einer Tabelle integriert werden

Semantik ist individuell

- Konzepte existieren nur im Kopf
- Man kann sie beschreiben, aber meint man auch dasselbe?
  - Individuelle Kenntnisse und Erfahrungen
- Also: Reden, reden, reden
- Dann: So gut wie möglich definieren
- Wie definiert man die Bedeutung eines Namens?
  - Formale Wissensrepräsentation (Ontologien, OWL – später)
  - Dokumentieren (mit lauter Namen)
- Standards definieren Bedeutung
  - Deshalb sind sie so schwierig und langweilig

Andere Klassifikationen

- Nach [BKLW99]
  - Syntaktische Heterogenität (= technische Heterogenität)
  - Datenmodellheterogenität (= Datenmodellheterogenität)
  - Logische Heterogenität (= semantische & strukturelle Heterogenität)
- Nach [Con97]
  - Semantische Konflikte (= semantische Heterogenität)
  - Beschreibungskonflikte (= strukturelle Heterogenität bei geringfügig unterschiedlicher Intension)
  - Heterogenitätskonflikte (= Datenmodellheterogenität)
  - Strukturelle Konflikte (= strukturelle Heterogenität bei gleicher Intension, schematische Konflikte)
- In der Realität immer schwer entwirrbare Kombinationen

Inhalt dieser Vorlesung

- Verteilung
- Autonomie
- Heterogenität
- Transparenz

Transparenz

- Verteilung, Autonomie, Heterogenität können in unterschiedlichem Maße überwunden werden
  - Ortstransparenz
    - Benutzer müssen den Ort der integrierten Systeme nicht kennen
    - Keine URLs, Datenbankpräfixe, ...
  - Quellentransparenz, Verteilungstransparenz
    - Benutzer weiß nicht, welche Quelle für eine Anfrage benutzt wurde oder benutzt werden kann (Datenherkunft)
    - Setzt ein globales Schema voraus
  - Schnittstellentransparenz
    - Benutzer kennt verschiedene Quellen gleich ansprechen
    - Keine Kenntnis lokaler Anfragesprachen, Protokolle, ...
  - Schematransparenz (Spezialfall von Verteilungstransparenz)
    - Benutzer kennt die Schemata lokaler Quellen nicht
    - Anfragen richten sich nur an das (homogene) globale Schema

Transparenz ist nicht immer erwünscht

- Oft strebt man maximale Transparenz an
- Oft ist aber kontraproduktiv
  - Benutzer kennen und vertrauen bestimmte(n) Datenquellen
  - Datenherkunft ist wichtigstes Kriterium für Einschätzung der Qualität der Informationen
  - Globales Schemas bedeutet, dass Benutzer neues Schema lernen müssen
  - Globale Schemata können sehr kompliziert werden
- Transparenz bedingt auch Informationsverlust