



Informationsintegration

Peer Data Management Systems

Wintersemester 2010/11

Armin Roth
armin.roth@hpi.uni-potsdam.de

1

Peer Data Management Systems

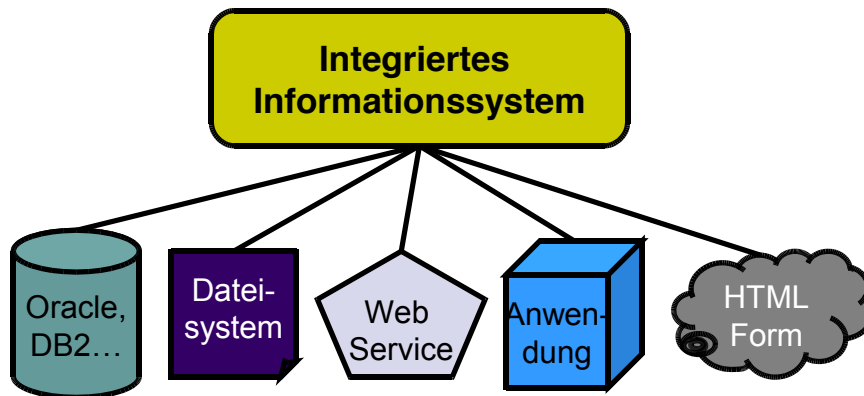
- Skalierbare und flexible Informationsintegration
- Struktur eines PDMS
- Anfragebearbeitung
- Optimierungsansätze
- Forschungssysteme



2

Rückblick: Integrierte Informationssysteme

- Globales Schema
- Direkter Zugriff auf jedes Quellsystem



Nachteile integrierter Informationssysteme

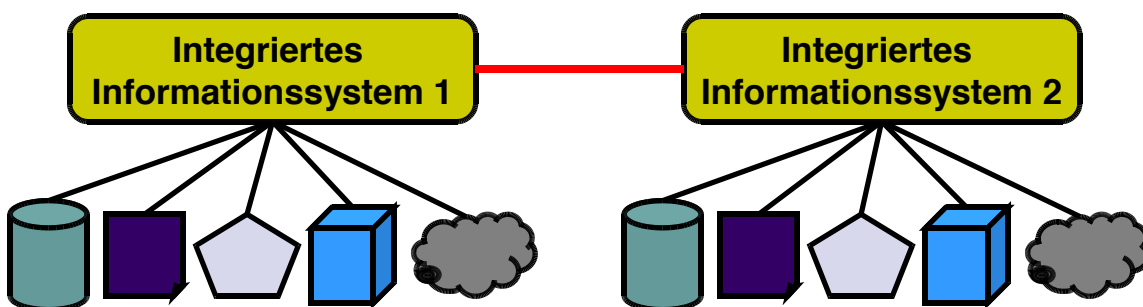
- Globales Schema:
 - ▶ Komplex
 - ▶ Einigungsprozess erforderlich
 - ▶ Wartung bedeutet (aufwändige) Schema-Evolution
- Skalierbarkeit, Flexibilität problematisch
- Mediator: Single point of ...
 - ▶ Control
 - ▶ Failure
 - ▶ Maintenance

Beispiel: Standardisiertes Schema

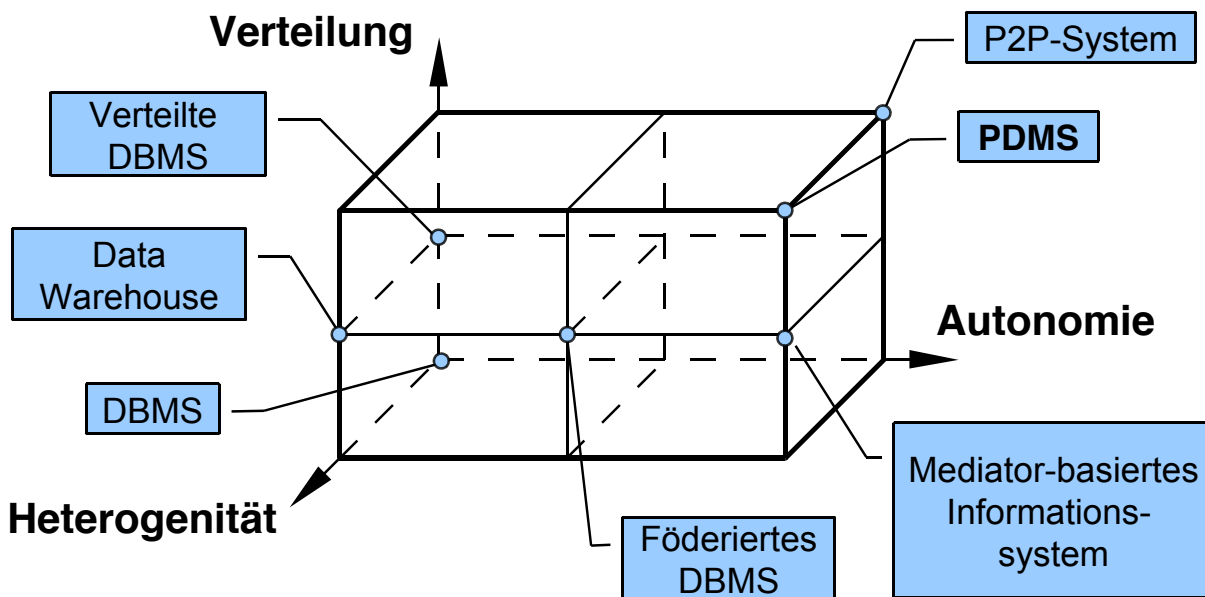
- ISO 10303 (STEP):
 Standard for Exchange of Product Data
 (Core for automotive mechanical design)
- Dokumentation: 3551 Seiten!
- Entstanden in langem Standardisierungsprozess
- Datentransformation aufwändig
- Flexibilität

PDMS als Generalisierung integrierter Informationssysteme

- Integration bereits existierender integrierter Informationssysteme
- Beispielsweise bei
 - ▶ Firmenzusammenschlüssen
 - ▶ Reorganisationen



Verteilte Informationssysteme



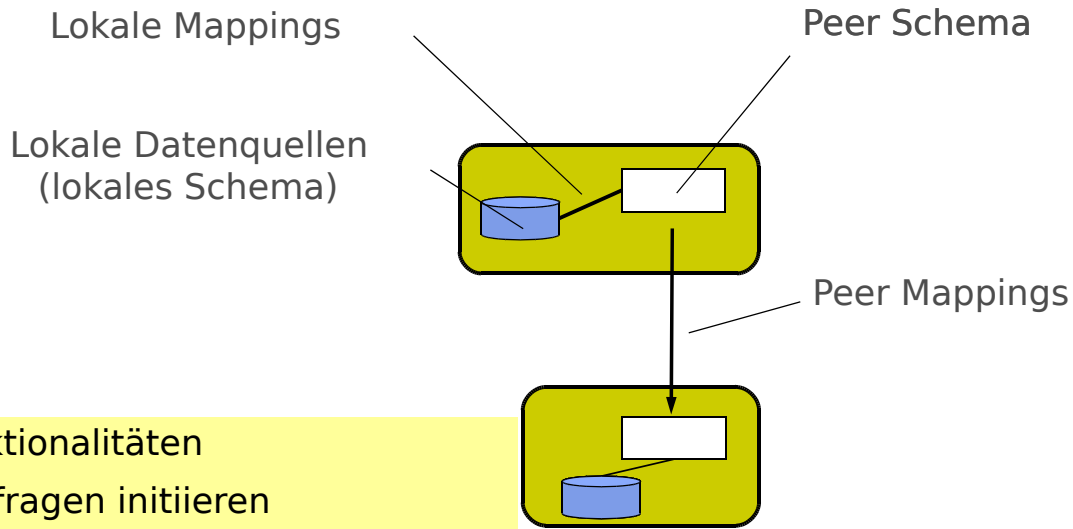
[Özsu, Valduriez 1999]

Peer Data Management Systems

- Skalierbare und flexible Informationsintegration
- Struktur eines PDMS
- Anfragebearbeitung
- Optimierungsansätze
- Forschungssysteme



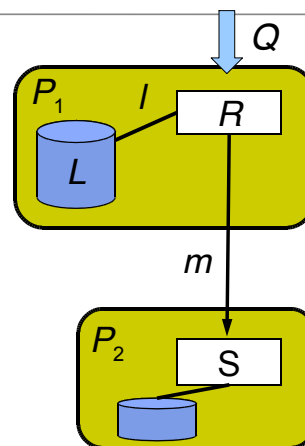
Peers: Integrierte Informationssysteme



Funktionalitäten

- Anfragen initiieren
- Anfragen auswerten
- Anfragevermittlung (Mediation)

Mappings eines Peers



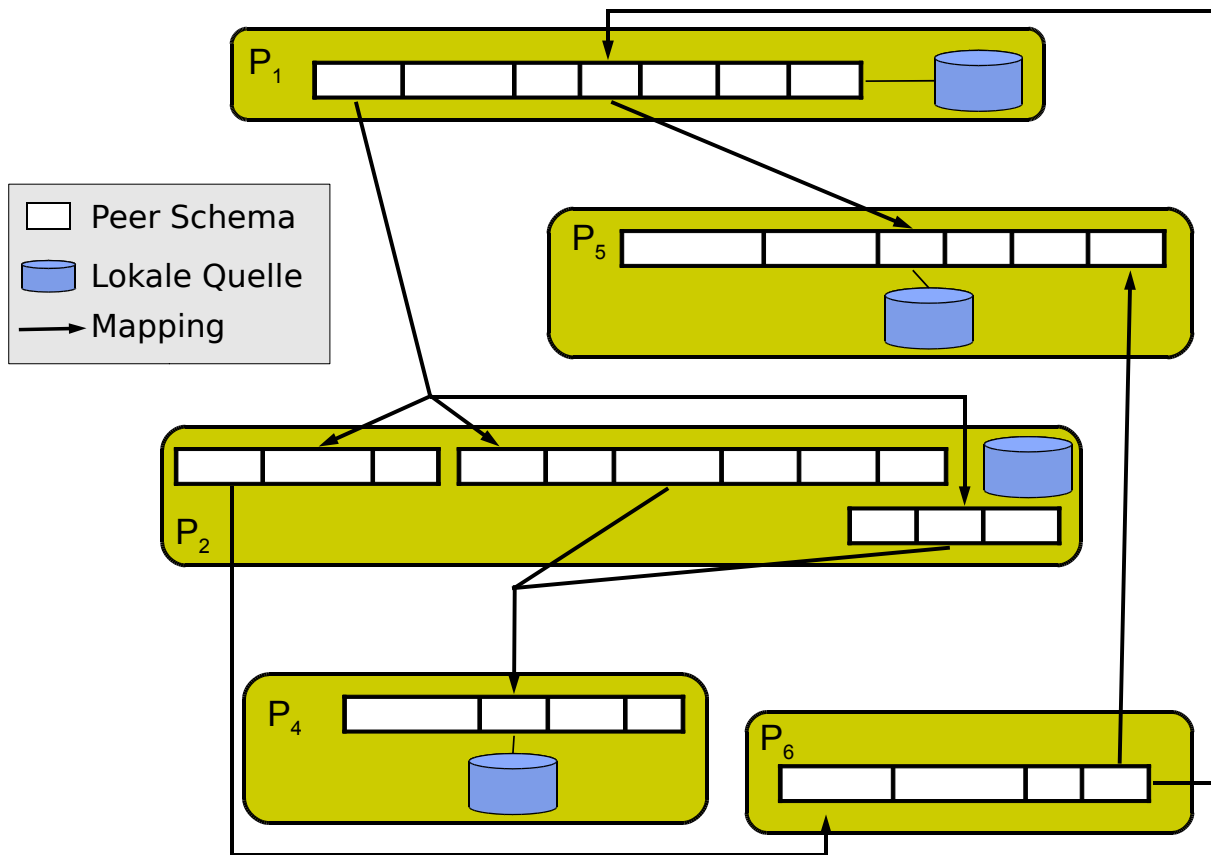
- Equality und Inclusion Mappings
- Inclusion Mappings sind gerichtet
 - ▶ Quelle: globales Schema
 - ▶ Ziel: lokales Schema

Lokales Mapping:

$$I: R(x, y) \supseteq L(x, y)$$

Peer Mapping:

$$m: R(x, y) \supseteq S(x, y)$$



Anwendungen

- Zusammenschlüsse von Organisationen
- Semantic Web
[Halevy et al. WWW 2003], [Heese et al. BTW 2005]
- Krisenmanagement
[Halevy et al. ICDE 2003]
- Gesundheit und Kliniken
- Groupware
[Albrecht et al. NetDB 2007]
- Generell:
große, lose gekoppelte integrierte Informationssysteme

Peer Data Management Systems

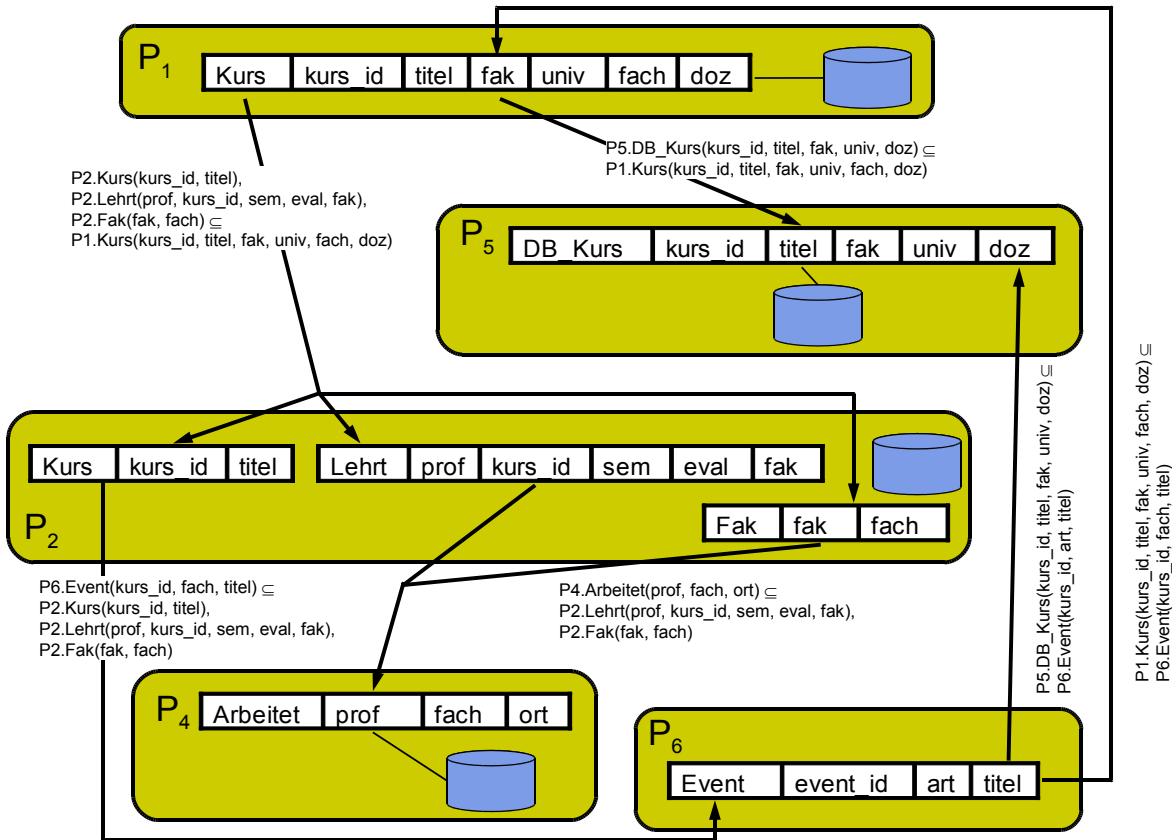
- Skalierbare und flexible Informationsintegration
- Struktur eines PDMS
- **Anfragebearbeitung**
- Optimierungsansätze
- Forschungssysteme



1

Ablauf Anfragebearbeitung

1. **Anfrageplanung** (durch -umformulierung)
2. **Anfrageoptimierung** (lokal vs. global)
3. **Anfrageausführung** (lokale Quelle/Nachbar-Peer)
4. **Kombination der Ergebnisse** (JOIN, UNION)

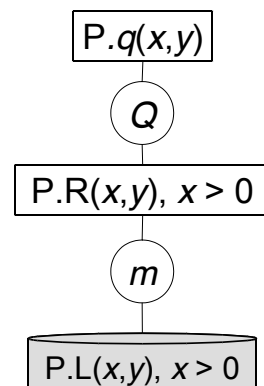
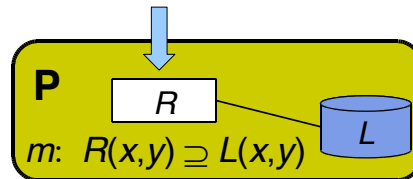


Anfragebearbeitung: Rule-Goal Tree

[Halevy ICDE 2003]

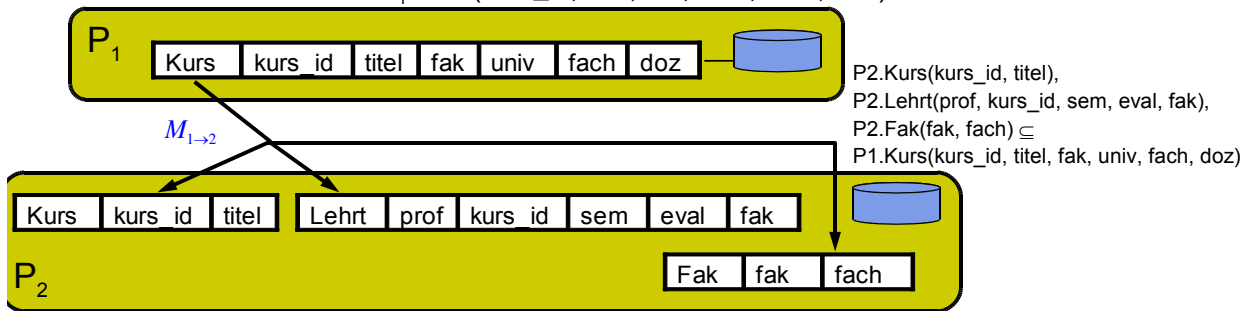
- Goal-Knoten: Prädikate (umformulierter) Anfrage(n) + zugehörige Selektionsprädikate
- Rule-Knoten: entsprechen Peer-Mappings

$$Q: q(x,y) :- P.R(x,y), x > 0$$



Global-as-View- Anfrageumformulierung

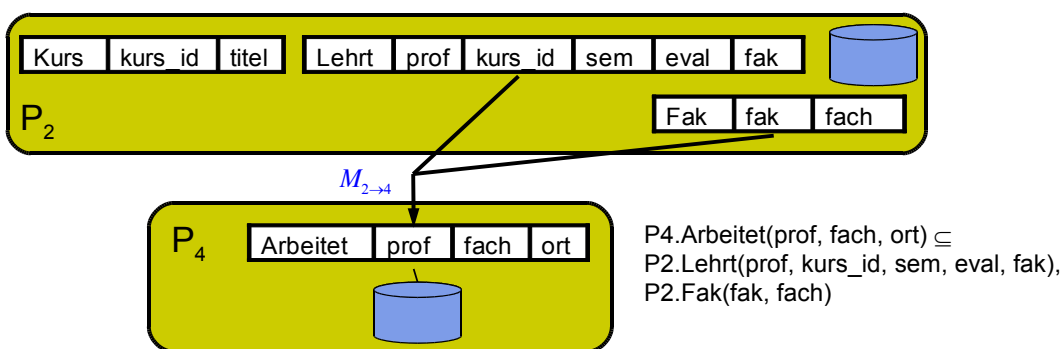
↓ $Q: P_1.q(kurs_id, titel, fak, univ, fach, doz) :-$
 $P_1.Kurs(kurs_id, titel, fak, univ, fach, doz)$



```

[] Peer001.q(kurs_id, titel, fak, univ, fach, doz)
  () Q
    [] Peer001.Kurs(kurs_id, titel, fak, univ, fach, doz)
      () M1-2
        [] Peer002.Kurs(kurs_id, titel)
        [] Peer002.Lehrt(prof_1, kurs_id, sem_2, eval_3, fak)
        [] Peer002.Fak(fak, fach)
    
```

Local-as-View- Anfrageumformulierung

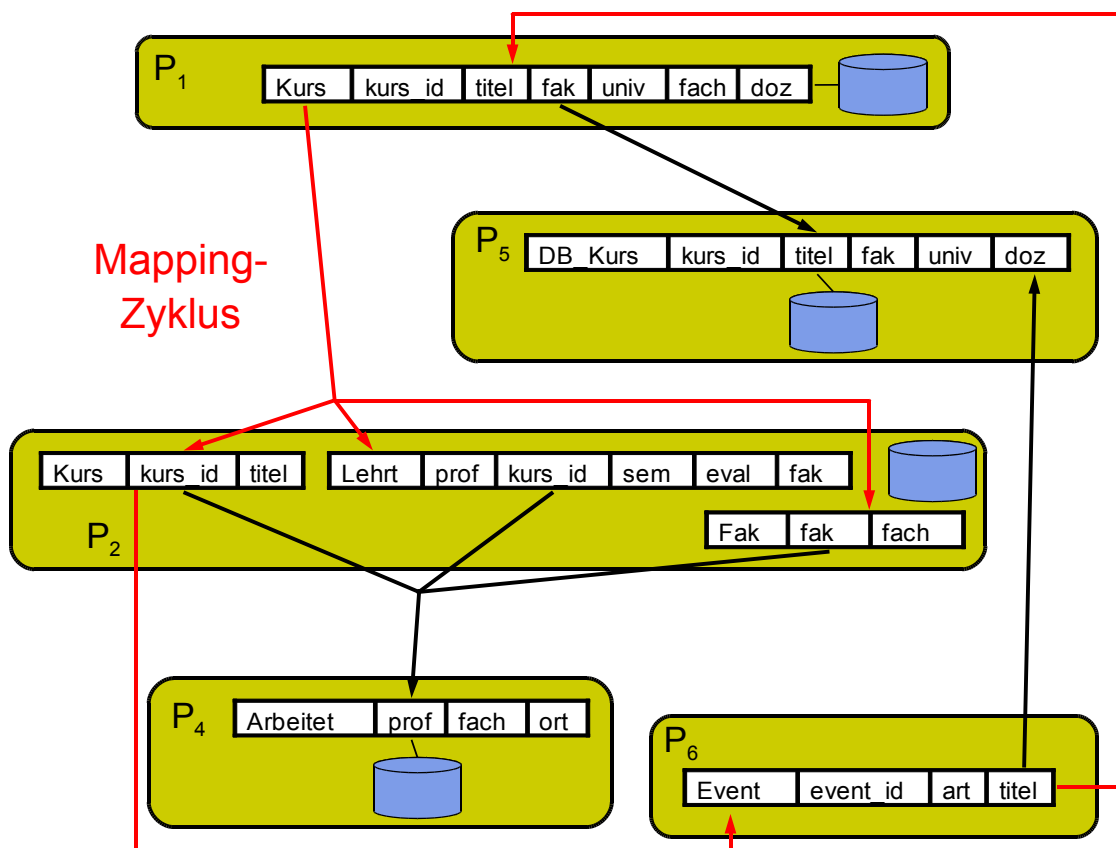
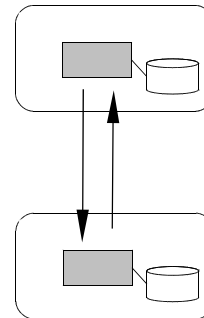


```

[] Peer001.q(kurs_id, titel, fak, univ, fach, doz)
  () Q
    [] Peer001.Kurs(kurs_id, titel, fak, univ, fach, doz)
      () M1-2
        [] Peer002.Kurs(kurs_id, titel)
        [] Peer002.Lehrt(prof_1, kurs_id, sem_2, eval_3, fak)
          () M2-4
            [] Peer004.Arbeitet(prof_1, fach, ort_6)
            [unc] Peer002.Fak(fak, fach)
        [] Peer002.Fak(fak, fach)
    
```

Behandlung von Zyklen

- Equality-Mappings bedeuten Zyklen:
 $Q_1(P_1) = Q_2(P_2)$ äquivalent zu
 $Q_1(P_1) \subseteq Q_2(P_2)$ und $Q_1(P_1) \supseteq Q_2(P_2)$
- First-order Logic-Semantik:
 Anfragebearbeitung bei Zyklen unentscheidbar
 [Halevy et al. ICDE 2003, Calvanese et al. PODS 2004]
- Abbruchkriterien
 (z.B. mehrfache Verwendung eines Mappings):
 verliert u.U. Antworten [Schweigert 2006]



```

[] Peer001.q(kurs_id, titel, fak, univ, fach, doz)
  () Q
    [] Peer001.Kurs(kurs_id, titel, fak, univ, fach, doz)
      () M1,1
        [] LS001_1.Kurs(kurs_id, titel, fak, univ, fach, doz)
      () M1,2
        [] Peer002.Kurs(kurs_id, titel)
          () M2,3-6
            [] Peer006.Event(event_id_4, art_5, titel)
              () M6,5
                [] Peer005.DB_Kurs(kurs_id_8, titel, fak_9, univ_10, doz_11)
                  () M1,5
                    [] LS005_1.DB_Kurs(kurs_id_8, titel, fak_9, univ_10, doz_11)
                  () M6,1
                    [] Peer001.Kurs(event_id_4, titel, fak_12, univ_13, art_5, doz_14)
                      () M1,1
                        [] LS001_1.Kurs(event_id_4, titel, fak_12, univ_13, art_5, doz_14)
                      () M1,5
                        [] Peer005.DB_Kurs(event_id_4, titel, fak_12, univ_13, doz_14)
                          () M1,5
                            [] LS005_1.DB_Kurs(event_id_4, titel, fak_12, univ_13, doz_14)
                          [unc] Peer002.Lehrt(prof_1, kurs_id, sem_2, eval_3, fak)
                          [unc] Peer002.Fak(fak, fach)
                    () M1,2
                      [] LS002_1.Kurs(kurs_id, titel)
                    [] Peer002.Lehrt(prof_1, kurs_id, sem_2, eval_3, fak)
                      () M1,2
                        [] LS002_1.Lehrt(prof_1, kurs_id, sem_2, eval_3, fak)
                      () M2,3-4
                        [] Peer004.Arbeitet(prof_1, fach, ort_6)
                          () M1,4
                            [] LS004_1.Arbeitet(prof_1, fach, ort_6)
                          [unc] Peer002.Fak(fak, fach)
                    [] Peer002.Fak(fak, fach)
                      () M1,2
                        [] LS002_1.Fak(fak, fach)
          () M1,5
            [] Peer005.DB_Kurs(kurs_id, titel, fak, univ, doz)
              () M1,5
                [] LS005_1.DB_Kurs(kurs_id, titel, fak, univ, doz)

```

$M_{1 \rightarrow 2}$ wird nicht
mehr genutzt:
Abbruch des
Zyklus



2

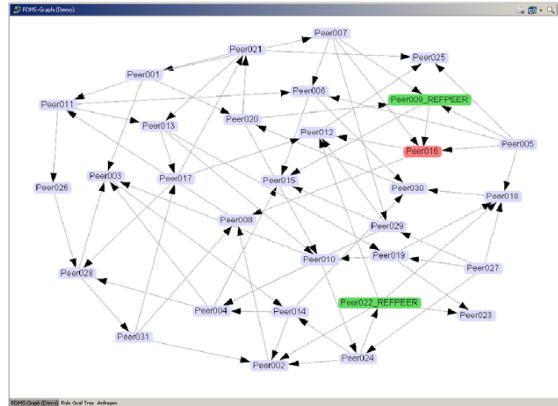
Anfrageumformulierung und Selektionsprädikate

- Überlappung Selektionsprädikate
 - ▶ von Nutzer-Anfrage und
 - ▶ von Anfragen in Mappings
- Entlang von Mapping-Pfaden akkumulieren sich Selektionen:
Verringern Kardinalität des Anfrageergebnisses

2

Effizienzprobleme durch Redundanzen

- Redundante Mapping-Pfade führen zu stark verzweigten Rule-Goal Trees
- Beispiel [Schweigert 2006]:
 - ▶ 31 Peers
 - ▶ Rang ca. 5 (Durchschnittliche Anzahl Nachbarpeers)
 - ▶ 34378 Union- und 17035 Join-Operationen



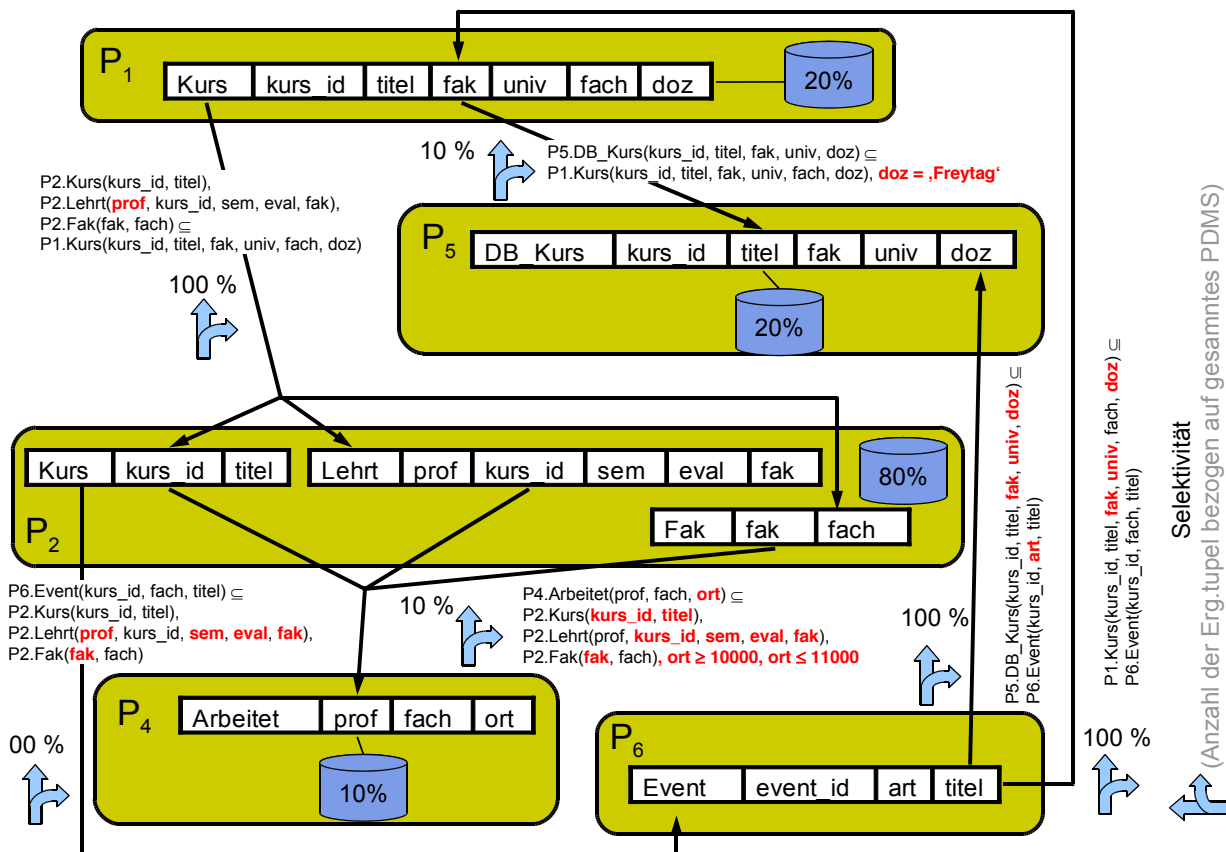
Peer Data Management Systems

- Skalierbare und flexible Informationsintegration
- Struktur eines PDMS
- Anfragebearbeitung
- Optimierungsansätze
- Forschungssysteme



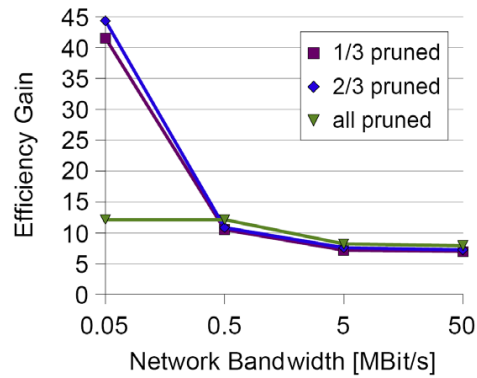
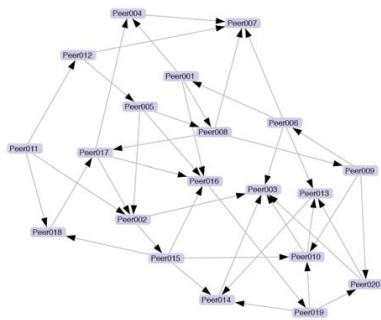
Qualitätsgesteuerte Anfragebearbeitung

- Wichtiges Qualitätskriterium in Informationsintegration: Vollständigkeit
- Extensionale Vollständigkeit: Tupelanteil
- Intensionale Vollständigkeit: Dichte von Datenwerten (Nicht-NULL)
- Projektionen und Selektionen in Peer-Mappings führen zu Informationsverlust
- Konzessionen an Vollständigkeit



Vollständigkeitsbasierte Anfrageplanung

- Bewertung von Mappings
Vollständigkeitsdifferenz ΔC bzgl. voll expandiertem lokalem Anfrageplan
- Pruning von Mapping mit ΔC unterhalb Grenzwert



Peer Data Management Systems

- Skalierbare und flexible Informationsintegration
- Struktur eines PDMS
- Anfragebearbeitung
- Optimierungsansätze
- Forschungssysteme



PDMS Piazza

[Halevy et al. ICDE 2003, Tatarinov et al. SIGMOD 2004]

- Semantik: First order Logic (FOL)
gesamtes PDMS hat *eine* Semantik
- Anfragen:
 - ▶ Subset von XQuery
 - ▶ Punktanfragen mit Negation
- Containment-basiertes Pruning von geschachtelten XQuery-Anfragen
- Minimalisierung umformulierter Anfragen

PDMS Hyper

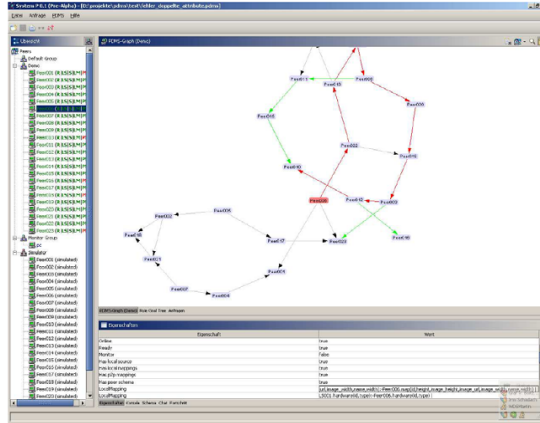
[Calvanese et al. PODS 2004]

- Epistemische Logik:
 - ▶ Jeder Peer hat eigene Semantik
 - ▶ Peers geben nur weiter, was sie sicher wissen
- Vorteile
 - ▶ Peers sind wirklich modulare Einheiten
 - ▶ Ermittlung aller Antworten ist in polynomieller Datenkomplexität (FOL-Semantik: unentscheidbar)

PDMS Humboldt Peers

[Roth VLDB PhD Workshop 2007]

- Relationales Datenmodell mit Punkt- und Range-Anfragen
- GaV- und LaV-Umformulierung
- Vollständigkeits-gesteuerte Anfragebearbeitung
- Visualisierung der Anfragebearbeitung
- Aktuelle Entwicklung: Selektivitätsschätzung mit selbstadaptiven Histogrammen
- Ausblick: Kostenmodell, Parallele Anfragebearbeitung



Zusammenfassung + Ausblick

- **Zusammenfassung**
 - ▶ Dezentral organisiert (kein globales Schema)
 - ▶ Hohe Flexibilität und Dynamik
 - ▶ Ineffizienz durch Redundanzen
 - ▶ Informationsverluste entlang Mapping-Pfade
 - ▶ Vollständigkeits-basierte Anfragebearbeitung
- **Ausblick**
 - ▶ Übergang zwischen unterschiedl. Datenmodellen (Relational/XML/unstrukturiert, pay-as-you go)
 - ▶ Management von Schemata und Mapping-Netzen
 - ▶ Approximative Anfragebearbeitung