

# Aktuelle Techniken der Webdatenintegration

Prof. Dr. E. Rahm und Mitarbeiter  
Seminar WS 2010/11

Google Produktsuche

Canon HF100 Camcorder - 3.31 MP - 12 x opt. :



von €532 bis €873 bei 76 Hand

Flash card, SD Memory Card, SDHC-Speicherk  
Der attraktive HF100 speichert ganz bequem 1.5 dabei einfach atemberaubend brillante HD Video  
Der HF100 speichert HD-Videos direkt auf SDH teilen. Die Karten lassen sich nämlich prima ver  
Der HF100 ist ein Camcorder im superkompakt Speicherung ohne mechanisch bewegliche Korr  
Der HF100 zeichnet in 1.920 x... [mehr >](#)  
[Zur Einkaufsliste hinzufügen](#)

[Preise vergleichen](#) [Technische Spezifikationen](#) [Verw](#)

[Allgemein](#) - [Hauptmerkmale](#) - [Speicher / Speicher](#) - [Objektivsystem](#) - [Zusat](#)  
[Batterie](#) - [Universal Product Identifiers](#)

Allgemein	
Produkttyp	Camcorder
Breite	7.3 cm
Tiefe	12.9 cm
Höhe	6.4 cm
Gewicht	380 g
Hauptmerkmale	
Sensoraufösung Camcorder	3.31 Mpix
Effektive Videoaufösung Camcorder	2.07 Mpix
Effektive Fotoaufösung Camcorder	2.76 Mpix
Widescreen-Videoaufzeichnung	Videoaufnahme im Breitbildformat
Medientyp	Flash card
Farbunterstützung	Farb
Typ des optischen Sensors	CMOS
Größe des optischen Sensors	1/3.2"
Mindestbeleuchtung	0.2 Lux
Digitales Zoom	200 x
Aufnahmegeschwindigkeit	

**Canon HF100** Produkt vergleichen

Camcorder

- Video-System: HD-Video
- Zoom: 12x optisch, 200x digital
- Brennweite: 4,80 mm - 57,60 mm
- Bild-Sensor 1/3,2"
- 3.310.000 Pixel Bild-Sensor
- 2,70" LCD-Monitor
- 211.000 Pixel LCD-Monitor Auflösung
- Standbildaufnahme
- Effektiv 2.760.000 MegaPixel
- [+Beschreibung erweitern]

Meinung 8: ★★★★☆ (15 Meinungen)  
Testnote 8: 1,8 (16 Testberichte)

Hersteller: **Canon**

Alle Angaben ohne Gewähr

**PREIS**  
SUCHMASCHINE



[Preisvergleich](#) [Meinungen](#) [Preisverlauf](#)

Meinungen zu Canon HF100

Klasse Cam 09.09.2008

Hier nach Vergleich mit einer Panasonic 3CCD (HDC-SD9) muss ich sagen, das im Low-Light-Bereich die Canon mehrere Nasen Vorsprung vor der 3CCD-Panasonic hat. Die Aufnahmen sind selbst bei wenig Licht sehr gut. Die Panasonic hat hier eine wesentlich schlechtere Bildqualität. Bei ausreichend Licht sind beide Cams in etwa gleichwertig. Wenn man eine 3-Chip-Cam mit 1-Chip-Cam vergleicht, so sollte die 3-Chip-Variante deutlich teurer sein. Bei Preisgleichstand gewinnt meist die 1-Chip-Cam. Insbesondere weil die Canon einen empfindlichen C-MOS Chip hat. Positiv bei der Canon: Größe + Gewicht Haptik Schärfe/Bildqualität Tonqualität Bildqualität von Fotos Negativ bei der Canon: Automatischer Weißabgleich Akkulaufzeit mitgelieferte Software Bei der Canon sollte man die vordefinierten Weisabgleichseinstellungen, oder den manuellen Weissabgleich bevorzugen (wenn möglich). Ich werde die Canon behalten und die Panasonic zurückgeben. Für das Geld gibts z.Zt. nichts besseres.

Meine Wertung: ★★★★★ Author: skeptiker

**Canon HF100 Test** (Camcorder)

[Merken und Vergleichen](#)

**GUT** 1,6

Google-Anzeige: **Canon HF100**  
Canon Camcorder im Vergleich! Testberichte lesen & sparen.  
www.ciao.de/Canon

Produktart: Typ: Speicherkarten - Camcorder, Sonderfunktionen: Digitalkamerafunktion, Optischer Zoom: 12x, Gewicht: 424g, Monitorgröße: 2,7", Digitaler Zoom: 200x, Aufzeichnungsformat: AVCHD; Speichermedien: SDHC ... [mehr Infos](#)



**Testberichte (14)** Preise (ca. 587 €) Meinungen (2) Datenblatt

**videofilmen**  
Ausgabe 2/2009  
Platz 3/14 im Test

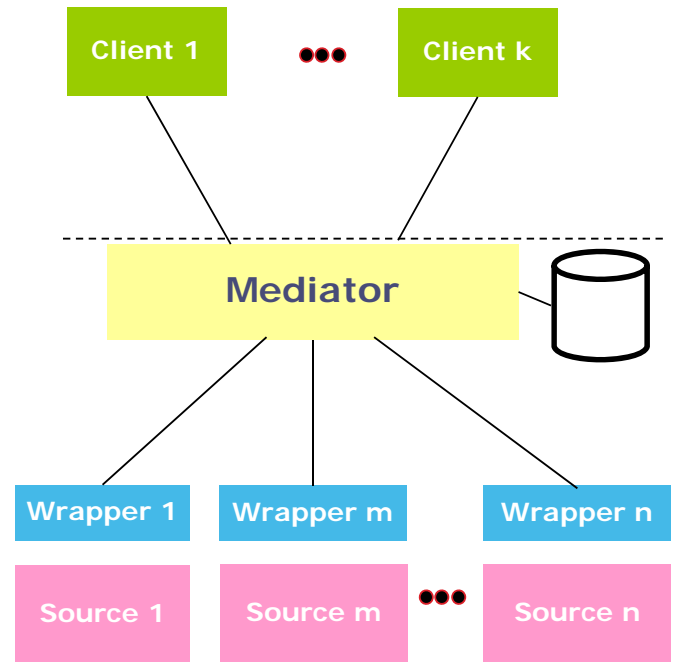
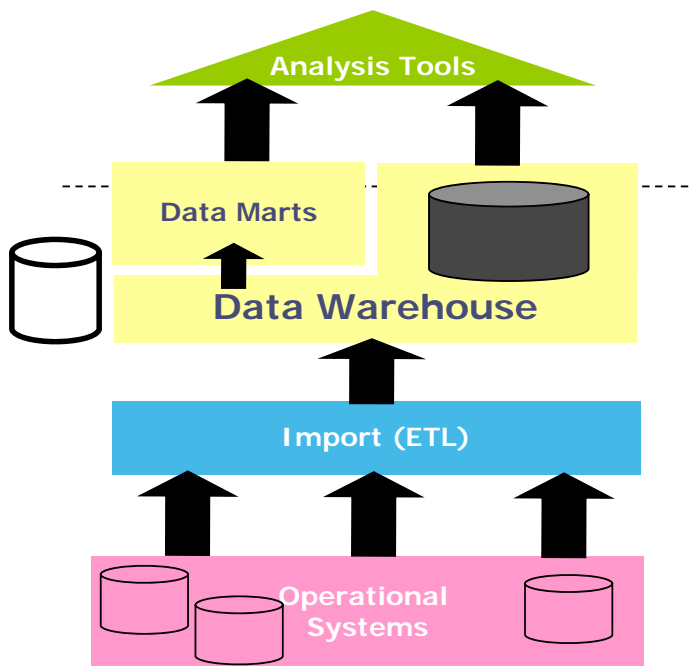
„sehr gut“ (146 von 200 Punkten)  
Bei der Bildschärfe liegen die HF-Modelle fast an der Grenze des technisch Machbaren bei der HD-Auflösung. Das schafft Canon auch ohne Tricks wie die elektronische Kantenerhebung, bei der auffällige Schärfungssäume auftreten. ...  
[Im Testbericht lesen](#)  
[Download \(2,00 €\)](#)  
[586,95 bei Amazon.de](#) [Camcorder suchen bei Otto.de](#)

**video**  
Ausgabe 8/2008  
Platz 1/3 im Test

„sehr gut“ (75 von 100 Punkten)  
Preis-Leistung: „hervorragend“ „Teststieger“  
„Plus: scharfstes Bild im Test, manuelle Tonaussteuerung, Mikro- und Kopfhörerbuchse; optischer Bildstabilisator. Minus: Zubehörschuh nur für Canon-Equipment“  
[Im Testbericht lesen](#)  
[Download \(1,20 €\)](#)  
[586,95 bei Amazon.de](#) [Camcorder suchen bei Otto.de](#)

**Informationsfusion**

# Physische vs. virtuelle Datenintegration



3

## Probleme derzeitiger DI-Lösungen

- ▶ Unzureichende Integration von Webdaten
- ▶ **Hoher Vorbereitungsaufwand:** Globales Schema, Schemaintegration, ...
- ▶ Datenqualität für heterogene/unsaubere Webdaten
- ▶ Dynamische Integration von Webdaten zur Laufzeit
- ▶ Unzureichende Unterstützung komplexerer Integrationsaufgaben
  - Welches zentral gelegene Hotel in Leipzig (ab 4 Sterne) unter 100 Euro/Nacht hat die besten Gastbewertungen

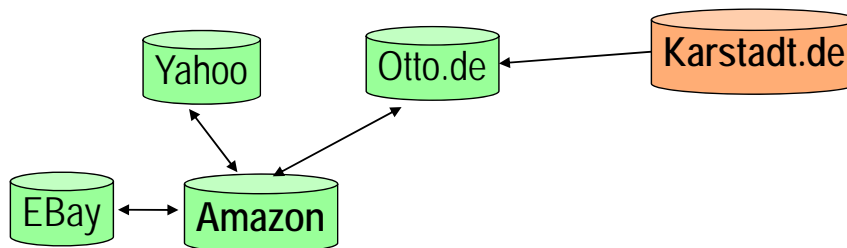
4

# Neuere Datenintegrationsansätze

- ▶ Peer-to-Peer-Ansätze
- ▶ Dataspaces
- ▶ Mashup-Ansätze
- ▶ Linked Data

5

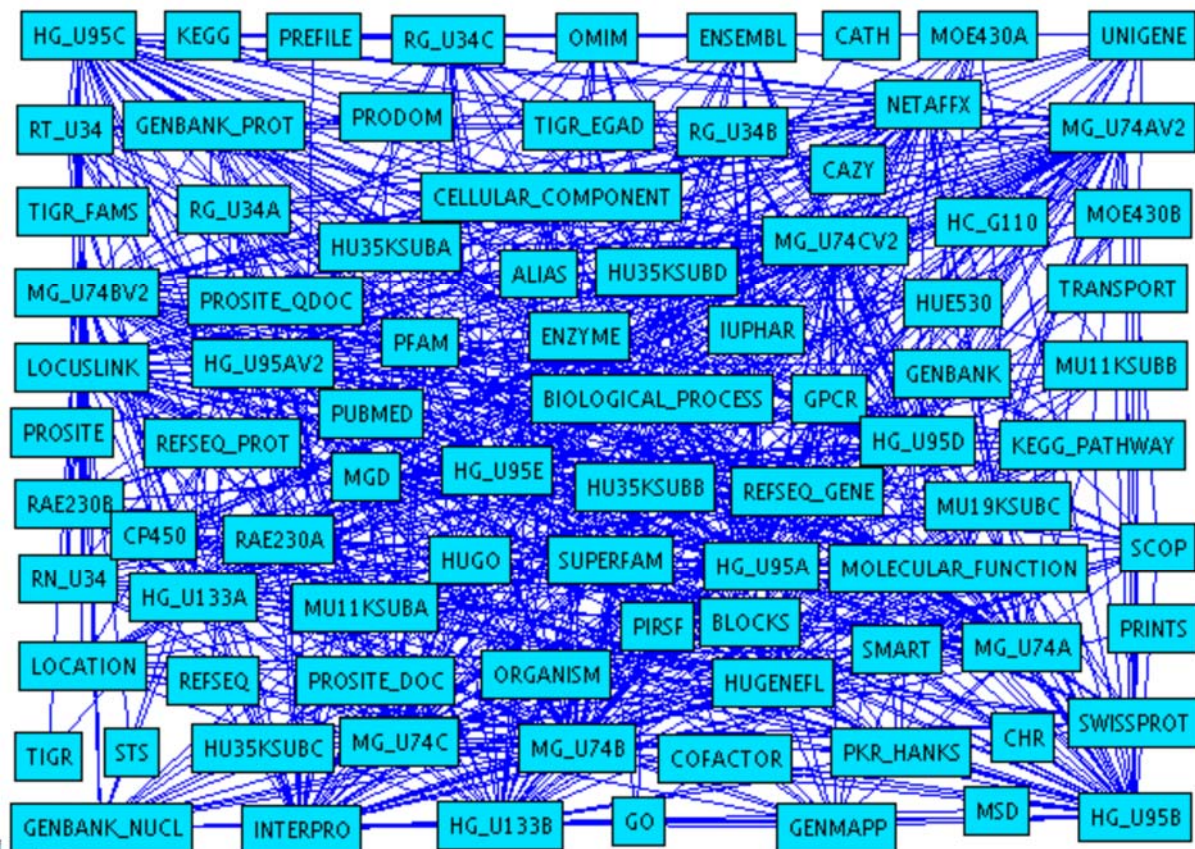
## Peer-to-Peer-Integration



- **Bidirektionale Verknüpfung von Datenquellen anstelle einem globalem Schema**
- **Propagieren von Anfragen zu Nachbarknoten**
- **Einfachere Erweiterung um neue Quellen**
- **Ähnlichkeiten zu Linked-Data-Ansätzen**

6

# Datenintegration in der Bioinformatik



7

H. Do, E. Rahm: Flexible Integration of Molecular-biological Annotation Data: The GenMapper Approach. Proc. EDBT Conf., 2004

## iFuice-Ansatz zur Datenintegration

- ▶ **Information Fusion** utilizing **Instance Correspondences** and **Peer Mappings**
  - **Generischer** P2P-Ansatz zur Informationsfusion
- ▶ **Mappings**
  - Referenzen auf Instanzebene
  - dynamische Berechnung über Queries, Web-Services
- ▶ mengenorientierte **Operatoren** auf Daten und Mappings
- ▶ **ontologisches) Domänenmodell** zur Kategorisierung von Quellen und Mappings
- ▶ **Mapping Mediator**: verwaltet Mappings und führt Operatoren aus

\*Rahm, E., et al.: iFuice - Information Fusion utilizing Instance Correspondences and Peer Mappings, Proc. 8th WebDB, Baltimore, June 2005

8

# Dataspaces


- ▶ Kompromiß zwischen Datenintegration („schema first“) und Suchmaschine („no schema“)
- ▶ Koexistenzansatz für unabhängige Datenquellen
- ▶ Dataspace = Menge von Participants + Mappings
  - Participants: heterogene und unabhängige datenquellen (DB, XML, Dateien, Web services, ...)
  - Schema Mappings + Instance Mappings
- ▶ Pay-as-you-go: inkrementeller Integrationsaufwand
- ▶ Heterogene Services
  - Catalog & Browse
  - Search & Query, ...

9






**DBLife**   Help | The Cimple Project

---

**Erhard Rahm**  
[Bing](#) [Citeseer](#) [DBLP](#) [Google](#) [Google Scholar](#) [Kosmix](#) [Wikipedia](#) [Yahoo!](#)



from Google Images

[more](#)

**Recent News**

- [Data Warehouse Scenarios for Model Management](#) cited 2 times - [details](#)
- [A survey of approaches to automatic schema matching](#) cited 23 times - [details](#)
- [Data Cleaning: Problems and Current Approaches](#) cited 12 times - [details](#)
- [Schema and ontology matching with COMA++](#) cited 12 times - [details](#)
- [MOMA - A Mapping-based Object Matching System](#) cited 3 times - [details](#)
- [Evaluating Instance-based Matching of Web Directories](#) cited 2 times - [details](#)

[News Archive](#)

Sorted by [Year/Conf](#), [Year/Citation](#), [Citation](#) [Community Statistics](#)

1	<a href="#">A survey of approaches to automatic schema matching</a> . Erhard Rahm, Philip A. Bernstein. VLDB J. (10): 334-350 (2001). <a href="#">Cited by 1890</a> <a href="#">Web Search</a> <a href="#">BibTeX</a> <a href="#">Download</a>
2	<a href="#">Generic Schema Matching with Cupid</a> . Javant Madhavan, Philip A. Bernstein, Erhard Rahm. VLDB 2001, 49-58. <a href="#">Cited by 1006</a> <a href="#">Web Search</a> <a href="#">BibTeX</a> <a href="#">Download</a>
3	<a href="#">Similarity Flooding: A Versatile Graph Matching Algorithm and Its Application to Schema Matching</a> . Sergej Melnik, Hector Garcia-Molina, Erhard Rahm. ICDE 2002, 117-128. <a href="#">Cited by 689</a> <a href="#">Web Search</a> <a href="#">BibTeX</a> <a href="#">Download</a>
4	<a href="#">COMA - A System for Flexible Combination of Schema Matching Approaches</a> . Hong Hai Do, Erhard Rahm. VLDB 2002, 610-621. <a href="#">Cited by 578</a> <a href="#">Web Search</a> <a href="#">BibTeX</a> <a href="#">Download</a>
5	<a href="#">Data Cleaning: Problems and Current Approaches</a> . Erhard Rahm, Hong Hai Do. IEEE Data Eng. Bull. (23): 3-13 (2000). <a href="#">Cited by 416</a> <a href="#">Web Search</a> <a href="#">BibTeX</a> <a href="#">Download</a>
6	<a href="#">Comparison of Schema Matching Evaluations</a> . Hong Hai Do, Sergej Melnik, Erhard Rahm. Web, Web-Services, and Database Systems 2002, 221-237. <a href="#">Cited by 249</a> <a href="#">Web Search</a> <a href="#">BibTeX</a> <a href="#">Download</a>
7	<a href="#">Rondo: A Programming Platform for Generic Model Management</a> . Sergej Melnik, Erhard Rahm, Philip A. Bernstein. SIGMOD Conference 2003, 193-204. <a href="#">Cited by 224</a> <a href="#">Web Search</a> <a href="#">BibTeX</a> <a href="#">Download</a>
8	<a href="#">Schema and ontology matching with COMA++</a> . David Aumüller, Hong Hai Do, Sabine Massmann, Erhard

**Professor**  
<http://dbs.uni-leipzig.de/personen/rahm.html>  
 Database  
 University of Leipzig  
 Germany  
 Papers cited 6,844 times  
[H-Index](#) of 27

**Related People**

- [Javant R. Hanitsa](#)
- [Philip A. Bernstein](#)
- [Alon Y. Halevy](#)
- [Javant Madhavan](#)

[more](#)

**Related Topics**

- [ontology](#)
- [location](#)
- [similarity](#)
- [metadata](#)

[more](#)

**Services**

- [VLDB 2010](#) (Session Chair) <sup>[1]</sup>
- [SIGMOD 2009](#) (Committee Members) <sup>[2]</sup>
- [ICDE 2008](#) (PC) <sup>[3]</sup>
- [SIGMOD 2006](#) (PC) <sup>[4]</sup>

[more](#)

**Related Organizations**

- [Microsoft Research](#)
- [National University of Singapore](#)
- [University of Leipzig](#)
- [Microsoft](#)

[more](#)

**Tutorials**

- [VLDB 2002](#) <sup>[5]</sup>

10

Woody Allen

Gefällt mir

- Overview
- Reference
- Video
- Images
- News & Updates
- Opinions & Discussions
- Gossip & Blogging
- Shopping
- Articles & Guides

woody allen

Get great deals on woody allen. Shop eBay!  
www.eBay.com

MyLife.com

Get Up To Date Information On Anyone In The US. Start Your Search Now.  
MyLife.com

Articles

The Kosmix Community

Contribute to Kosmix

AUTHOR: Video

Woody Allen centuries: Video from Fancast



Josh Brolin Woody Allen  
01:31



Woody Allen On Radio Days  
01:48

Love Life

Celebrity Lovelife from Whosdatedwho



Woody Allen  
Sagittarius

Allen's distinctive films, which run the gamut from dramas to screwball sex comedies, have made him a notable American director. He is also distinguished by his rapid rate of production and his very large body of work.[1] Allen writes and directs his movies and has also acted in the majori...

Who has Woody Allen dated?



Harlene Rosen Konigsberg



Louise Lasser



Diane Keaton



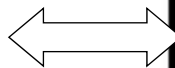
Mia Farrow



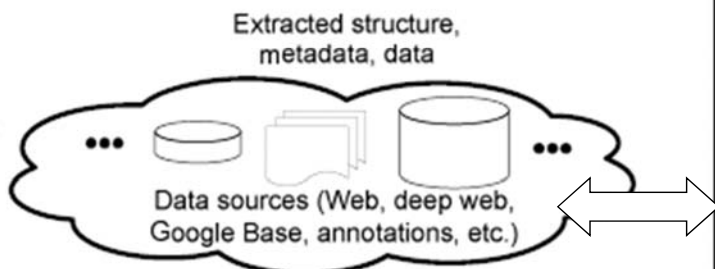
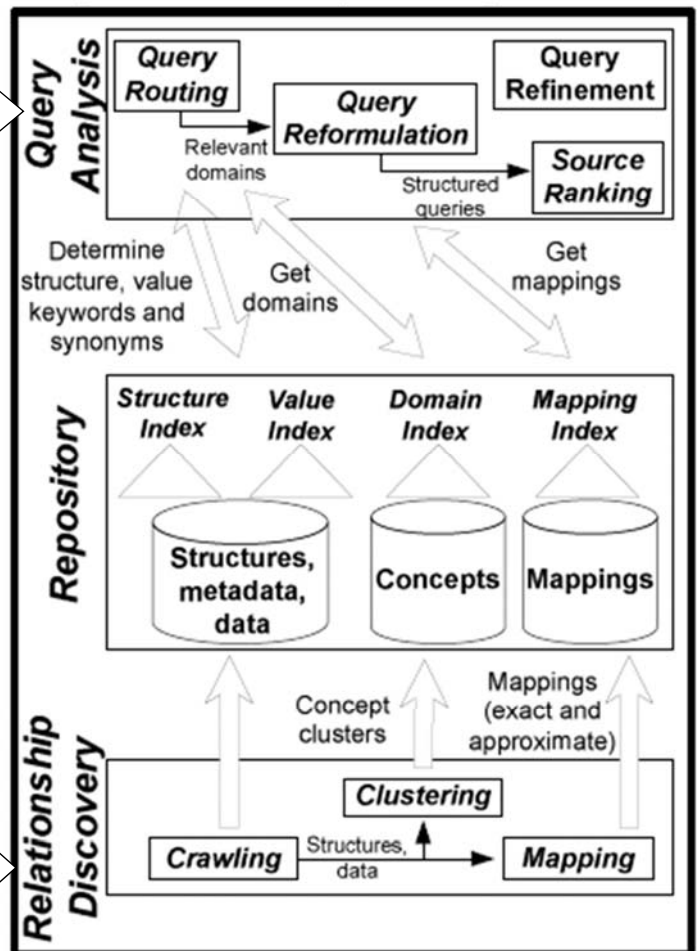
Soon-Yi Previn

PayGo: Architecture

[Madhavan et.al., CIDR'07]



- . Web-scale
- . Multiple domains
- . Exploit structures & mappings for query answering
- . Clustering
- . Approximated Mappings



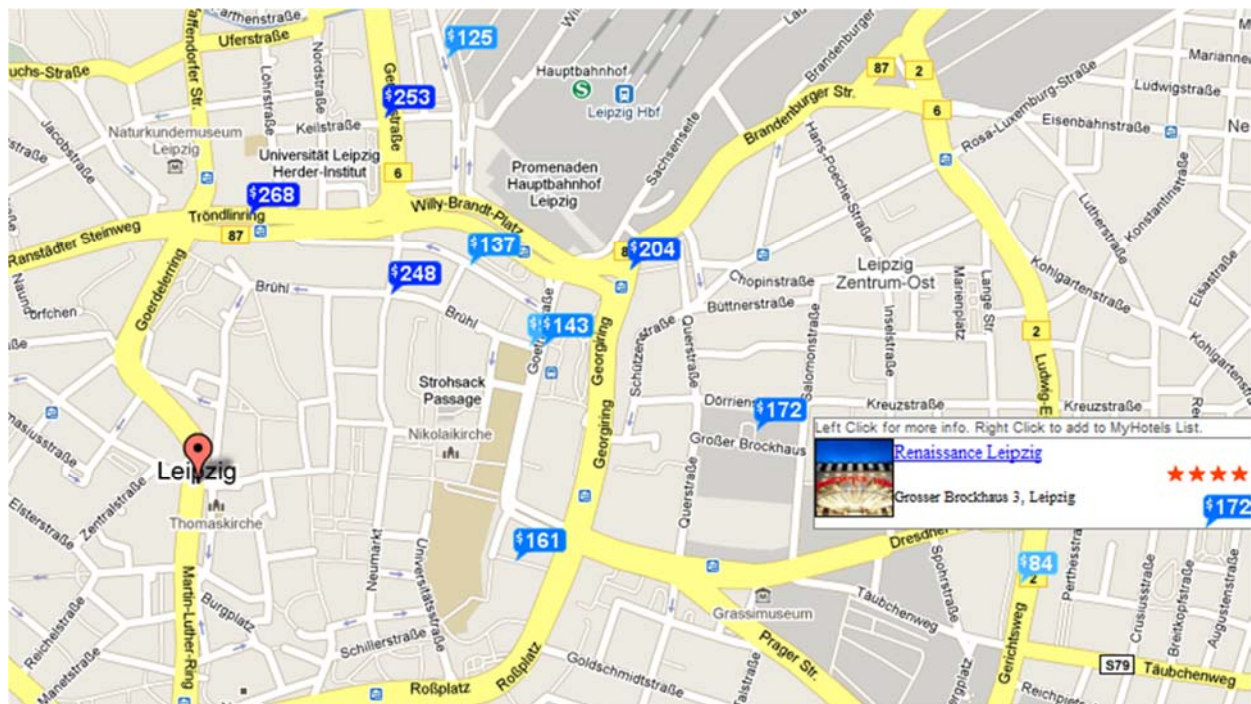
# Mashups

- ▶ "A web mashup is a web page or **application** that **combines data** from two or more external **online sources**." (ProgrammableWeb)
- ▶ "A mashup is a web application that combines data from **more than one source** into an **integrated experience**." (Wikipedia)

## Leichtgewichtiger Ansatz zur Datenintegration

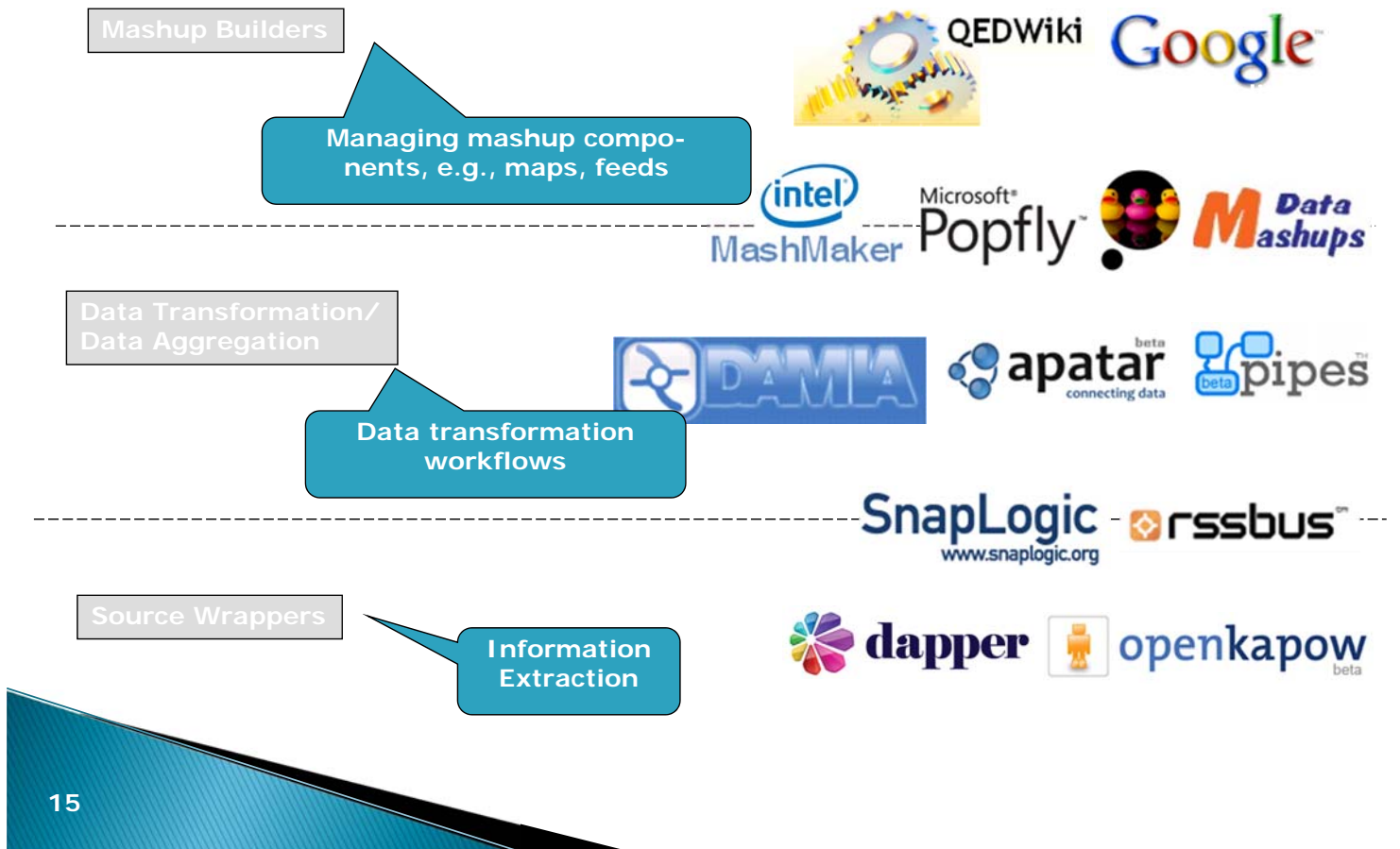
13

## Mashup – Beispiel: Hotelpreise



14

# Mashup Tools: Overview

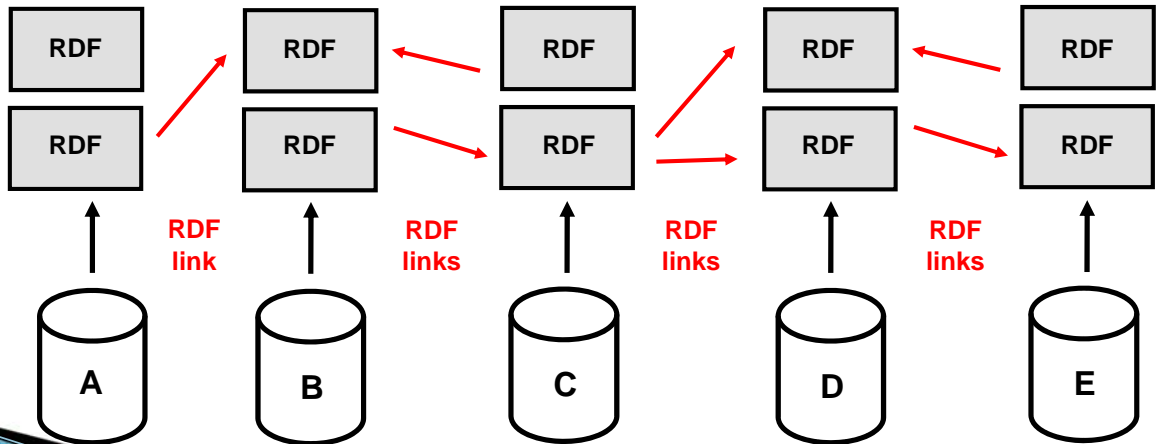


## Mashups: aktueller Stand

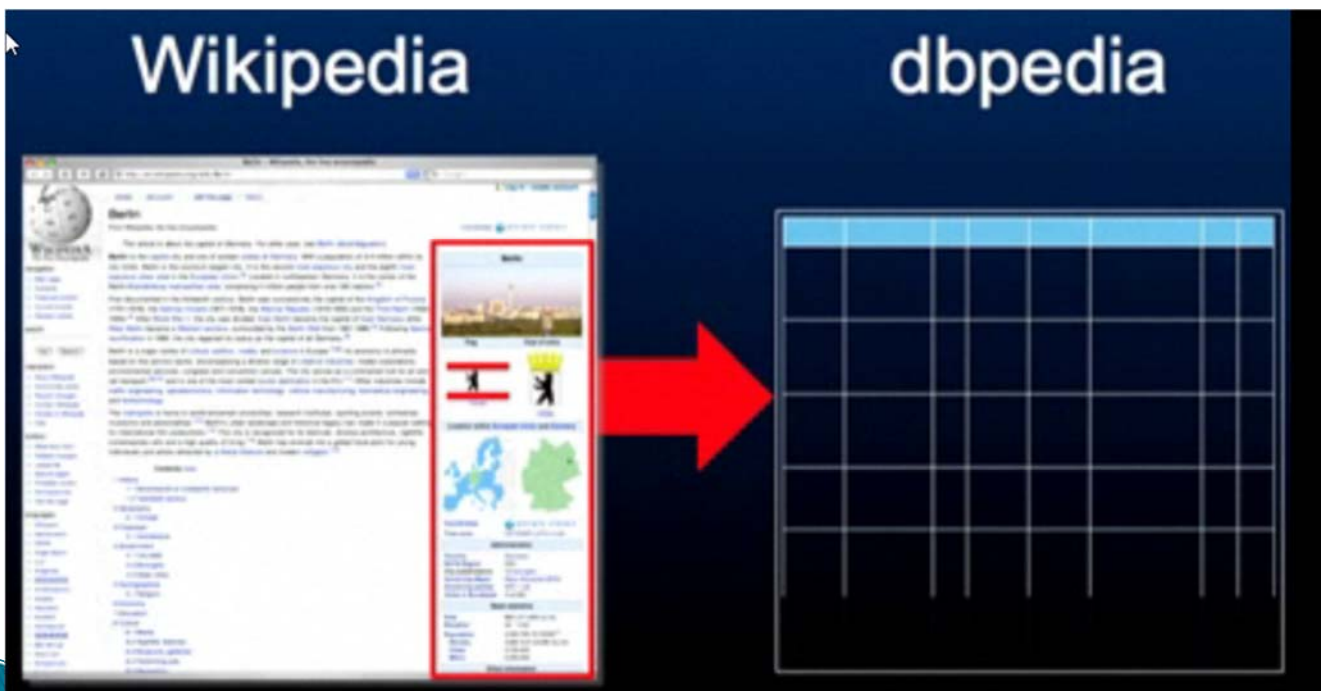
- ▶ Einfache Formen der Datenintegration
  - Map-Visualisierung über Koordinaten/Adressen
  - einfache Workflows, zB zum Mischen von Feeds/Suchmaschinenergebnissen
  - zum Teil GUI-basierte Definition (z.B. Yahoo Pipes)
- ▶ Einschränkungen
  - Einfache Ergebnis-Verarbeitung (Merge statt Match)
  - Unzureichende Datenqualität
  - Keine Anpassung von Anfragen für unterschiedliche Quellen
- ▶ Enterprise-Mashups
  - IBM Mashup Center, JackBe Presto, Serena
  - Fokus auf Unternehmensdaten, weniger Webdaten



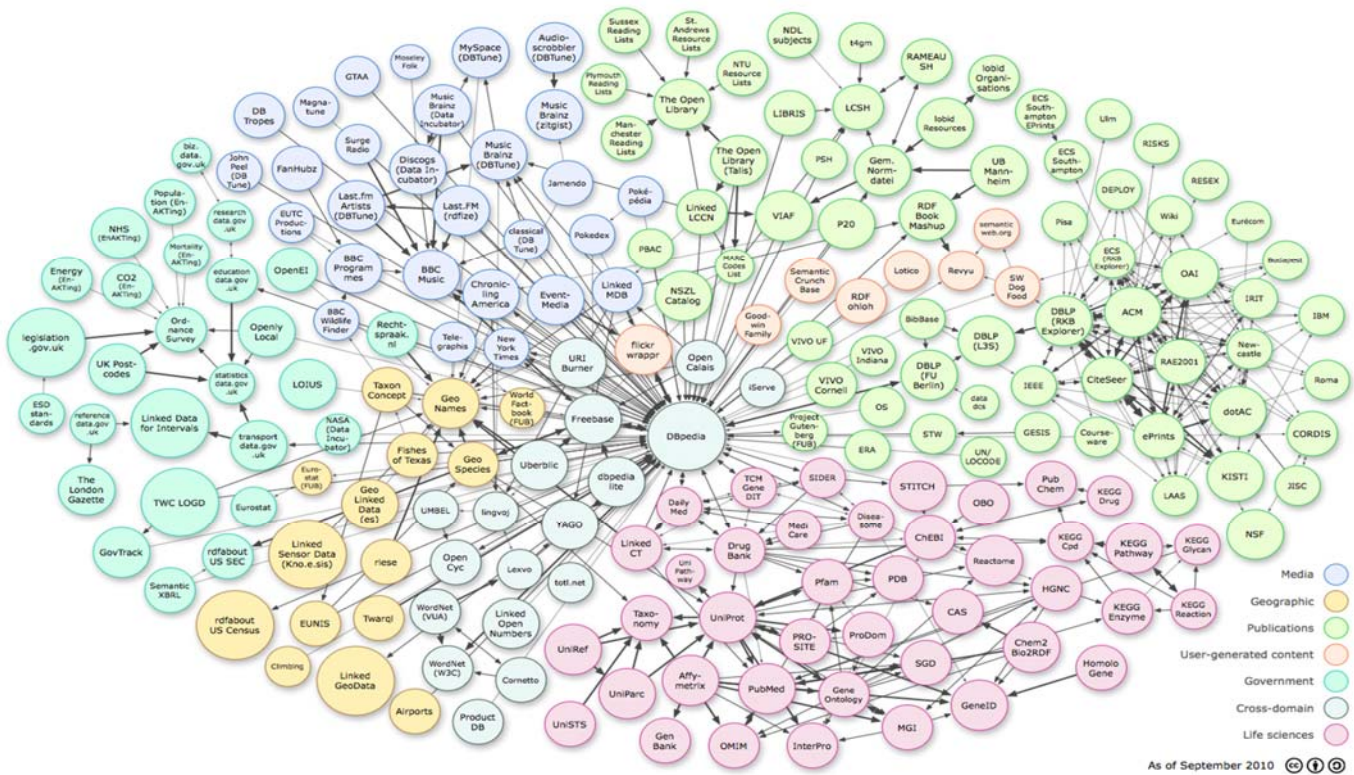
- ▶ Semantic Web-Initiative zur Vernetzung strukturierter Webdaten
  - Nutzung von URI, RDF-Tripeln und RDF-Links
- ▶ Daten/Link-Extraktion aus Webseiten (Übergang vom Dokumenten-Web zum Daten-Web)



## DBpedia



# Linked Datasets

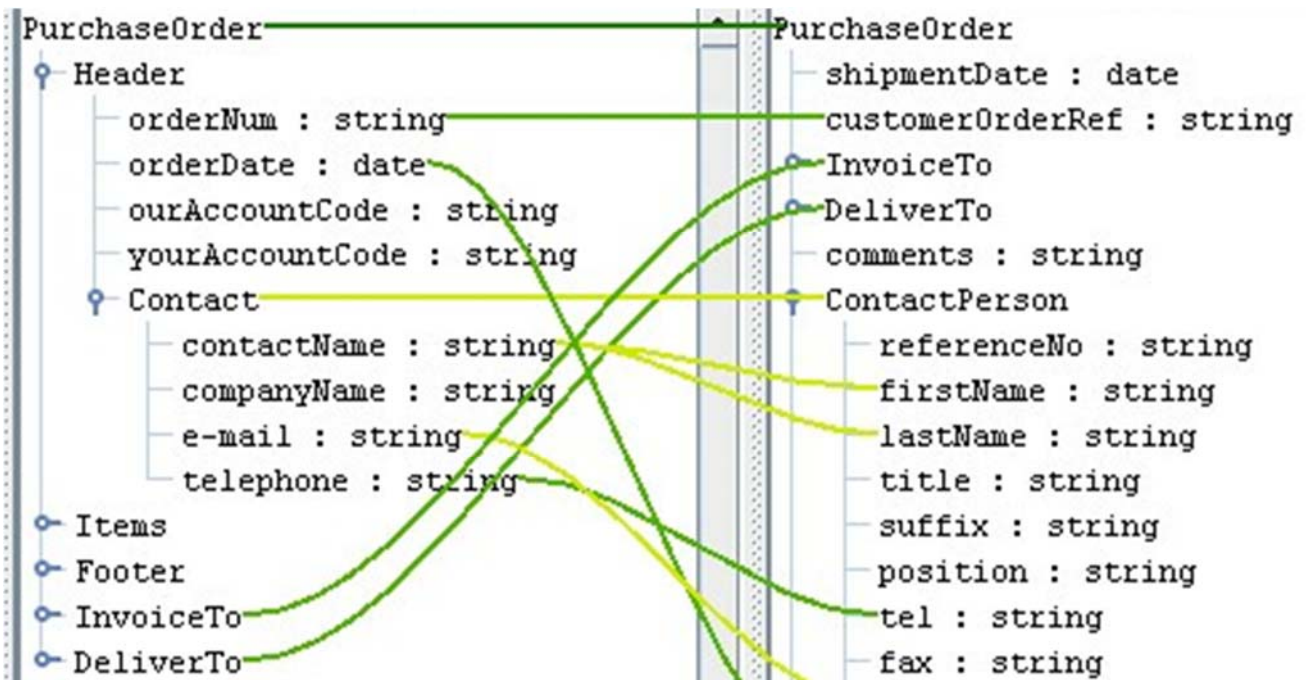


ca. 90% aller Links im Life Science-Bereich (2009)

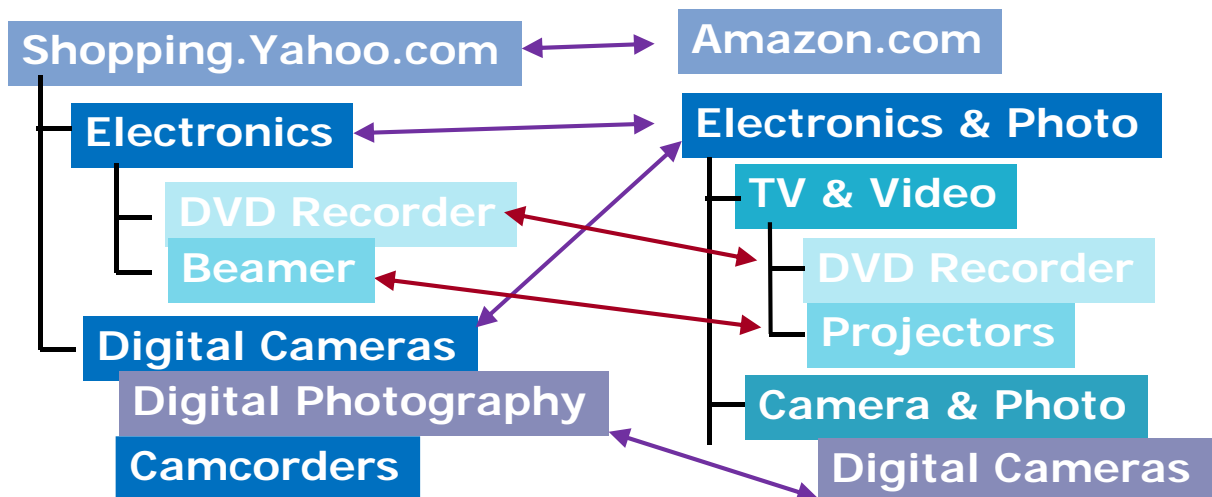
## Querschnittsprobleme

- ▶ Schema/Ontologie-Matching
- ▶ Schema/Ontologie-Merging
- ▶ Objekt-Matching (und -Fusion)

# Schema matching (COMA++)



# Matching von Produktkatalogen



# Integration von Web-Formularen

**1 Where and when do you want to travel?**

Leaving from:

Going to:

Departing: (MMDD/YY)   Anytime

Returning: (MMDD/YY)   Anytime

**2 Who is going on this trip?**

1 Adults (age 19 to 64)

0 Seniors (age 65 and over)

0 Children (age 18 and under)

**3 Do you have any preferences?**

Airline:  Class:  Economy / Coach



From:  To:

Round trip One way Multi city

Departure date: Feb 19 Morning

Return: Feb 19 Morning

Search by Schedule Price

Passengers: 1 Check

Depart City:

Destination City:

Depart Date: May 15 2004 SAT

Return Date: May 17 2004 MON

Passengers: 1 Adult 0 Child (Age 2 to 11)

Class: Economy

Leaving from:  Departure date: Feb 26 10am

Going to:  Return date: Mar 05 10am

Passengers: 1 Preferred cabin: Economy/Coach

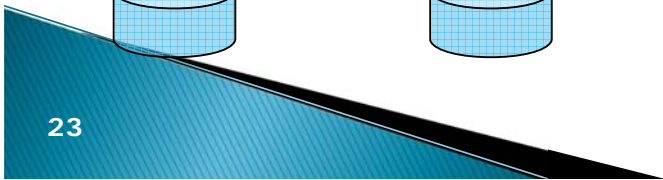
From: find a city name To: find a city name

enter airport or city enter airport or city

Departure Date: Feb 27 Return Date: Mar 2

No same-day returns.

Number of Tickets: 1 **START**



## Web Interface matching

Leaving from:  Departure date: Feb 26 Time: 10am

Going to:  Return date: Mar 05 Time: 10am

Passengers: 1 Preferred cabin: Economy/Coach

From: find a city name To: find a city name

enter airport or city enter airport or city

Departure Date: Feb 27 Return Date: Mar 2

No same-day returns.

Number of Tickets: 1 **START**



Interface 1 (S<sub>1</sub>)

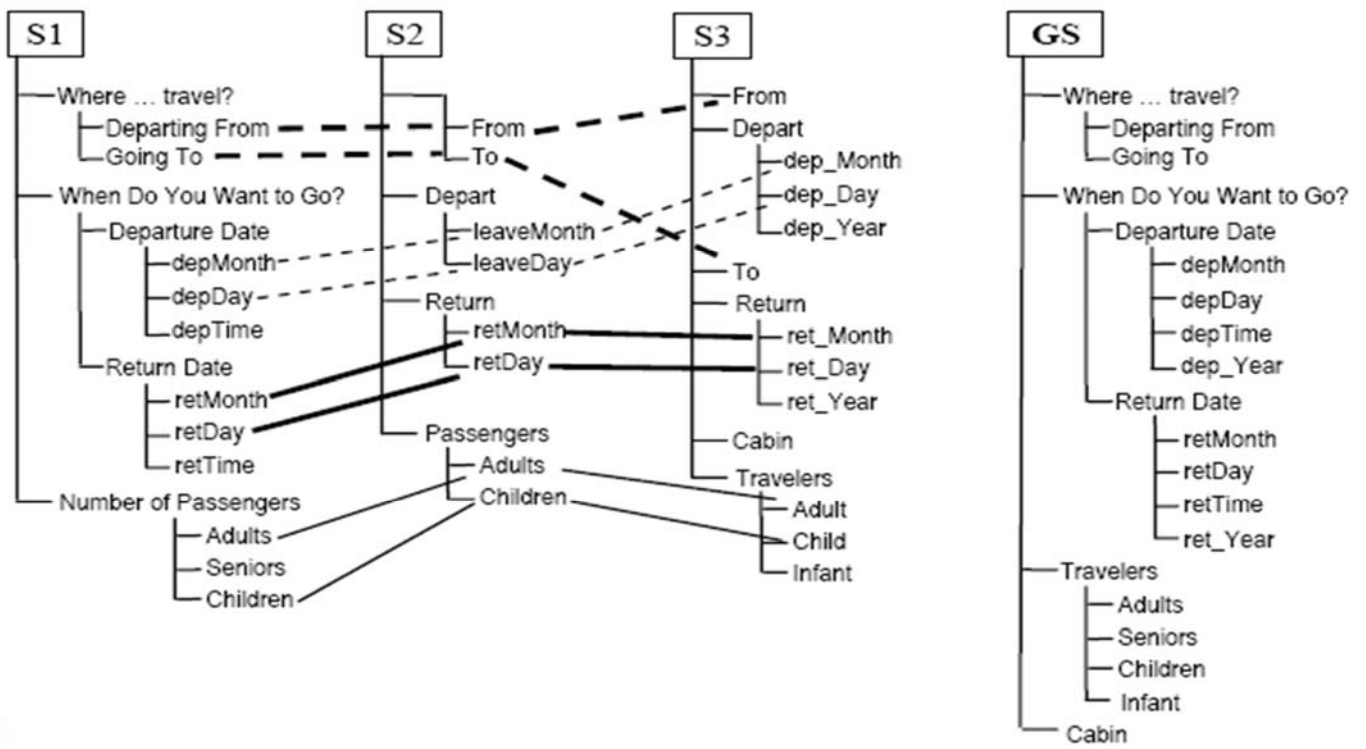
- Leaving from
- Going to
- Departure date
- Return date
- Passengers:
- Time
- Preferred cabin

Interface 2 (S<sub>2</sub>)

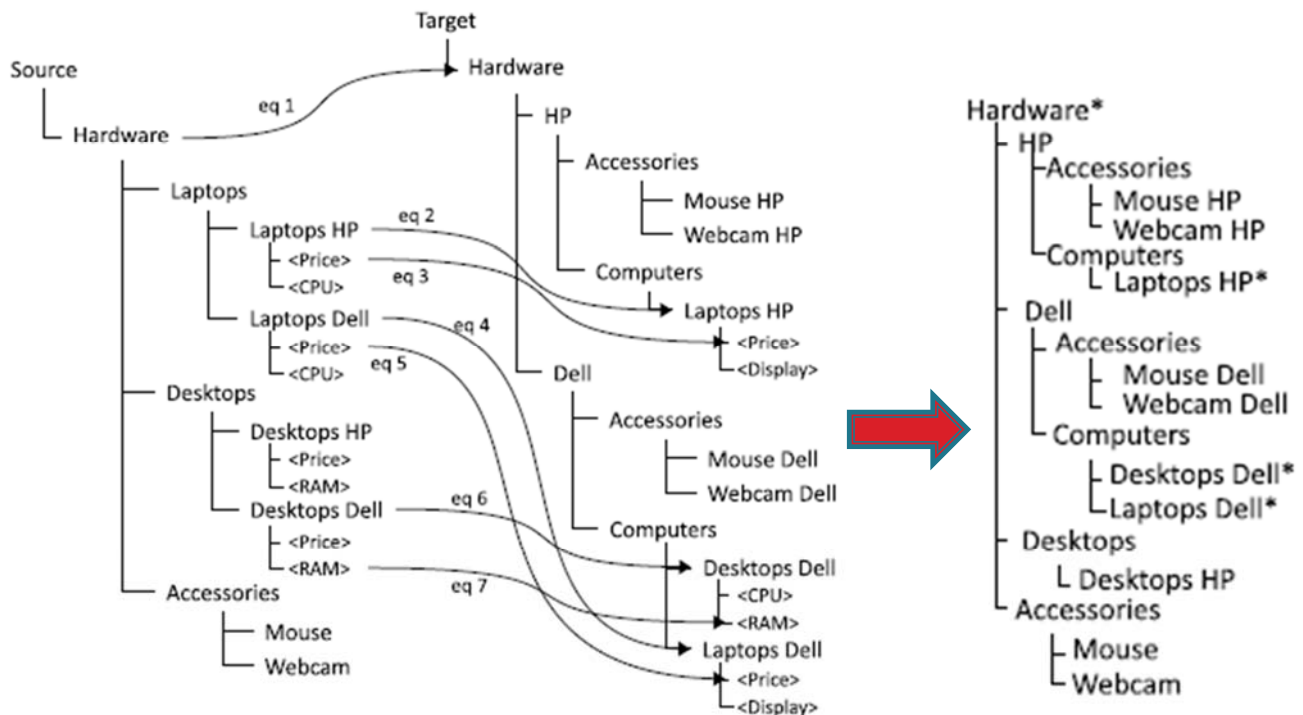
- From
- To
- Departure date
- Return date
- Number of tickets

- N-Wege Matching (**Holistic Schema Matching**)
- Clustering ähnlicher Attribute zum Mischen

# Mischen von Web-Formularen



# Mischen von Produktkatalogen



# Duplikate in Webdaten

**Canon VIXIA HF S10 Camcorder - 1080p - 8.59 MP - 10 x optical zoom** \$975 new  
Flash card, 32 GB, 1y warranty, F/1.8-3.0  
The VIXIA HF S10 delivers brilliant video and photos through a Canon exclusive 8.59 megapixel CMOS image sensor and the latest version of Canon's advanced image processor, ...  
★★★★★ 12 reviews - [Add to Shopping List](#)  
from 52 sellers [Compare prices](#)

**Canon ( VIXIA HF S10 iVIS Dual Flash Memory Camcorder** \$899.00 new  
Canon HF S10 iVIS Dual Flash Memory Camcorders SPECIAL SALE PRICE: \$899  
Display both English/Japanese + we supplu all English manuals in English as PDF. ....  
[Add to Shopping List](#)  
Made in Japan Online

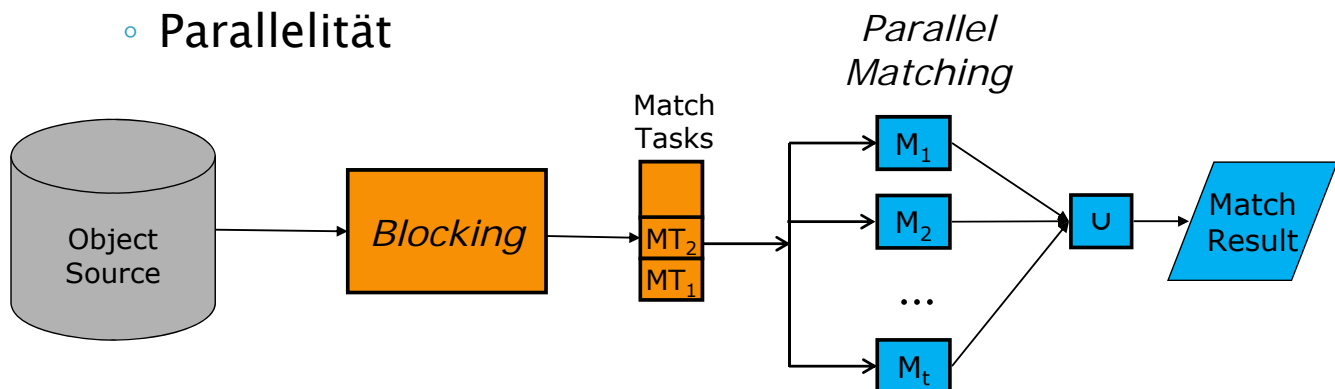
**Canon VIXIA HF S10** \$999.00 new  
Dual Flash Memory High Definition Camcorder The Next Step Forward in HD Video  
Canon has a well-known and highly-regarded reputation for optical excellence, ...  
[Add to Shopping List](#)  
Performance Audio  
[2 seller ratings](#)

**Canon VIXIA HF S100 Flash Memory Camcorder** \$899.95 new  
\*\*\*\*Canon Video HF S100 Instant Rebate Receive \$200 with your purchase of a new Canon VIXIA HF S100 Flash Memory Camcorder. (Price above includes \$200 ....  
[Add to Shopping List](#)  
Arlingtoncamera.com  
[5 seller ratings](#)

**Canon Vixia Hf S10 Care & Cleaning** \$2.99 new  
Care & Cleaning Digital Camera/Camcorder Deluxe Cleaning Kit with LCD Screen Guard Canon VIXIA HF S10 Camcorders Care & Cleaning.  
[Add to Shopping List](#)  
shop.com  
★★★★☆ 38 seller ratings

# Paralleles Objekt-Matching

- ▶ Großer Suchraum, v.a. zur Evaluierung des Kartesischen Produkts der Eingabeobjekte
- ▶ Effizienz durch
  - Blocking und/oder
  - Parallelität



# SEMINAR

29

## Seminarziele

- ▶ Beschäftigung mit einem praxis- und wissenschaftlich relevanten Thema
  - kann Grundlage für Abschlussarbeit oder SHK-Tätigkeit sein
- ▶ Erarbeitung + Durchführung eines **Vortrags** unter Verwendung wissenschaftlicher (englischer) Literatur
- ▶ Diskussion
- ▶ **Schriftliche Ausarbeitung** zum Thema
- ▶ Hilfe und Feedback durch zugeteilten Betreuer

30

# Seminar: Anrechnungsmöglichkeiten

- ▶ **Masterstudium**
  - Teil der Module *Moderne Datenbanktechnologien (bzw. Anwendungsspezifische Datenbankkonzepte)*
  - *Seminarmodul (oder Masterseminar)*
- ▶ **Bachelorstudium**
  - *Seminarmodul (oder Bachelorseminar)*
  - *(Vertiefungsmodul in Kombination mit Real. v. IS)*
- ▶ **Alte Studiengänge (Diplom, etc.)**
  - Problemseminar

31

## Scheinvergabe / Modulprüfung

- ▶ **selbständiger Vortrag mit Diskussion (ca. 45 Minuten)**
  - Abnahme der Folien durch Betreuer
- ▶ **schriftliche Ausarbeitung (ca. 15 Seiten)**
  - Abnahme der Ausarbeitung durch Betreuer
  - Ausarbeitung soll zum Vortragstermin vorliegen (Vorträge ab 3. Januar 2011)
- ▶ **aktive Teilnahme an allen Vortragsterminen**
- ▶ **Modul-Workload: 30h Präsenzzeit,  
120 h Selbststudium**

32



# Seminar (3)

- ▶ max. 20 Teilnehmer
- ▶ Vortragstermine
  - Montags, 9 -10:45 Uhr
  - Montags 11 - 12:45 Uhr, JG 1-22, ab 3. 1. 2011
- ▶ Themenzuordnung
  - Koordinierungstreffen mit Betreuer bis **spätestens 5.11.2010**
  - ansonsten verfällt Seminaranmeldung
  - freiwilliger Rücktritt auch bis max. 5.11.2010

Nr.	Thema	Termin	Betreuer	Studenten
	<b>Linked Data</b>			
1	Einführung, DBpedia und Dynamik		M. Hartung	Eidon, AS, etc
2	Anfrageverarbeitung		T. Kirsten	Gröze, Verk.
3	Entity Resolution und Cleaning		H. Köpcke	(E)Hybrid, Gmiden
	<b>Data Spaces</b>			
4	Einführung und Systeme		D. Aumüller	Beer, We 1 für
5	Probabilistic Mappings		K. Wurdinger	(F)hili, etc
	<b>Mashups</b>			
6	Neue Mashup-Ansätze		S. Endrullis	Britschauken, (Kadins)
	<b>Aktuelle Match- und Merge-Techniken</b>			
7	Holistic Schema Matching		A. Groß	Böhmig, Lij
8	Schema- und Ontology-Merging		S. Raunich	(S)B, Ma
9	Community-based Matching		S. Maßmann	Thomas, H. etc
10	Entity Resolution in der Cloud		L. Kolb	Zellman, Beug