Innovationslabor

Semantische Integration von Webdaten

Prof. Dr. Erhard Rahm

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

http://dbs.uni-leipzig.de/format

UNIVERSITÄT LEIPZIG
Abteilung Datenbanken
am Institut für Informatik

Programmablauf

- Überblicksvortrag Prof. Rahm
- Feedback / Diskussion
- Vorstellung der Prototypen
  - Frau Maßmann
  - Dr. Thor
- Feedback / Diskussion
- Get-Together (Felix-Klein-Hörsaal)

BMBF: ForMaT

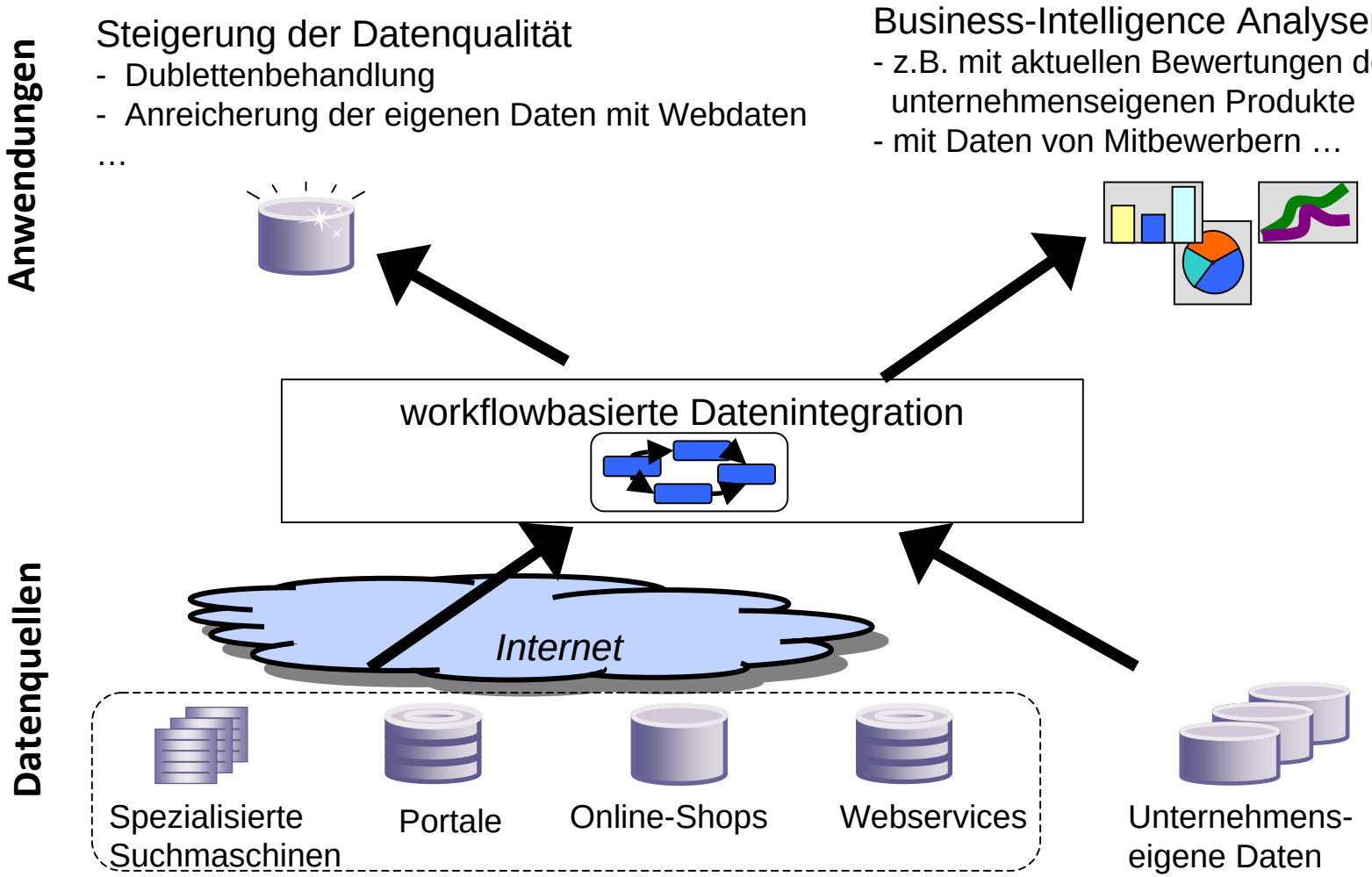
GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
ForMaT
UNTERNEHMEN REGION
Die BMBF-Innovationsinitiative Neue Länder

- Förderprogramm:
  - Forschung für den Markt im Team
  - Im Rahmen von Unternehmen Region (Innovationsinitiative für die Neuen Länder), http://www.unternehmen-region.de
- Ziel
  - Abstrakt: frühzeitiger Wissens- und Technologietransfer von der öffentlichen Forschung in die Wirtschaft
  - Konkret: Firmenausgründung nach Ende der Förderung
- Ansätze
  - Analyse der Forschungsuntersuchungen auf Verwertungseignung
  - Förderung der interdisziplinären Zusammenarbeit
- Förderung in mehreren Runden mit jeweils zwei Phasen
  - Phase 1: Potential-Screening, 6 Monate
  - Phase 2: Weiterentwicklung innerhalb von Innovationslabor, 2 Jahre
- Stand (Runde 2):
  - Derzeit 27 Initiativen in Phase 1 (ausgewählt aus ca. 100)

Semantische Integration von Webdaten

- Initiative der Universität Leipzig, Abt. Datenbanken (Prof. Rahm)
- Aktuell noch in Phase 1, dem Potenzial-Screening
- Ziel: weitgehende Automatisierung von Datenintegrationsaufgaben mit hoher Datenqualität, insbesondere für Webdaten
  - Schnellere Umsetzung als mit Data-Warehouse-Ansätzen (Offline-Datenaufbereitung)
  - On-Demand Datengewinnung (fokussierte Suche) , Datenabgleich und Analyse
  - Gegenüber einfachen Mashup-Lösungen Anwendbarkeit auf großen Datenmengen und Unterstützung für hohe Datenqualität
- Basiert auf eigenen Forschungsergebnissen / Prototypen
  - *Workflow-basierte Datenintegration mit fokussierter Websuche:* iFuice
  - *Objekt-Matching, Dublettenbehandlung*: MOMA und STEM
  - *Schema und Ontologie-Matching*: COMA und COMA++

Semantische Integration von Webdaten



Projektbeteiligte



- Prof. Dr. E. Rahm
- Dipl.-Inf. S. Maßmann
- Dr. A. Thor


- Prof. Dr. C. Lindemann
- Dipl.-Inf. C. Hoffmann


- Prof. Dr. H. Löbler
- Dipl.-Kfm. R. Müller

Vortragsinhalte

- Einleitung

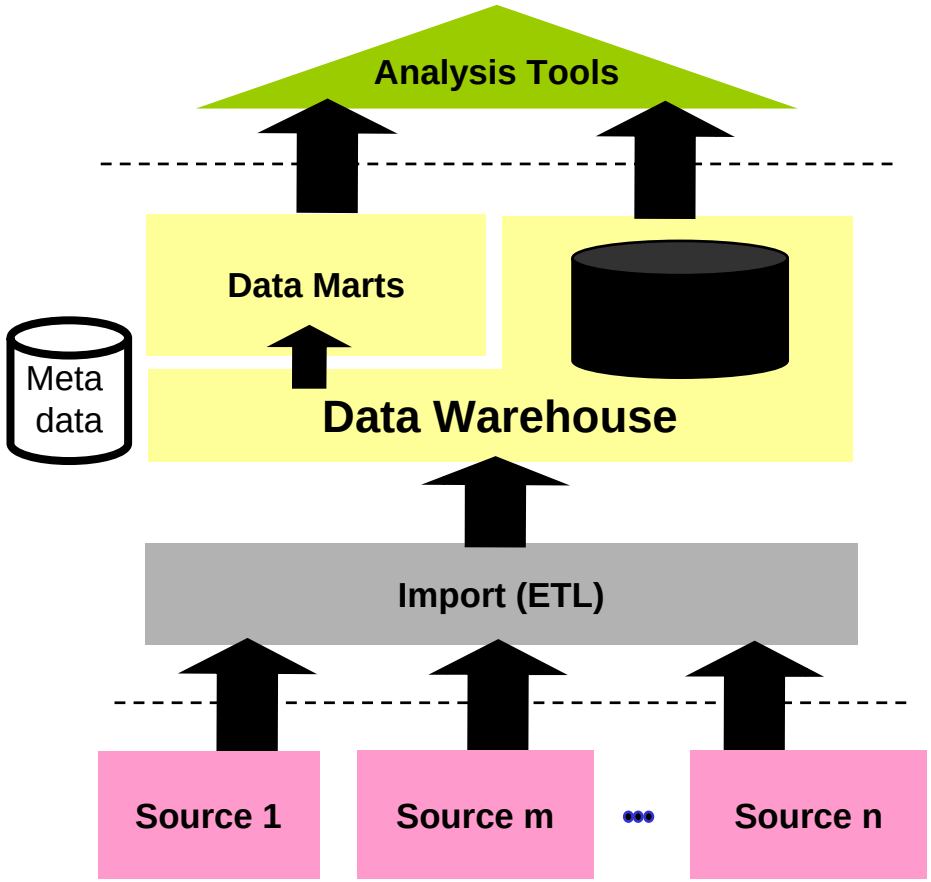
- Derzeitige Ansätze zur Datenintegration
  - Warehousing, EII
  - Mashups

- Eigene Forschungsarbeiten / Prototypen
  - Schema/Ontologie Matching: COMA++
  - Objekt-Matching: MOMA, STEM
  - Workflowbasierte Integration: iFuice

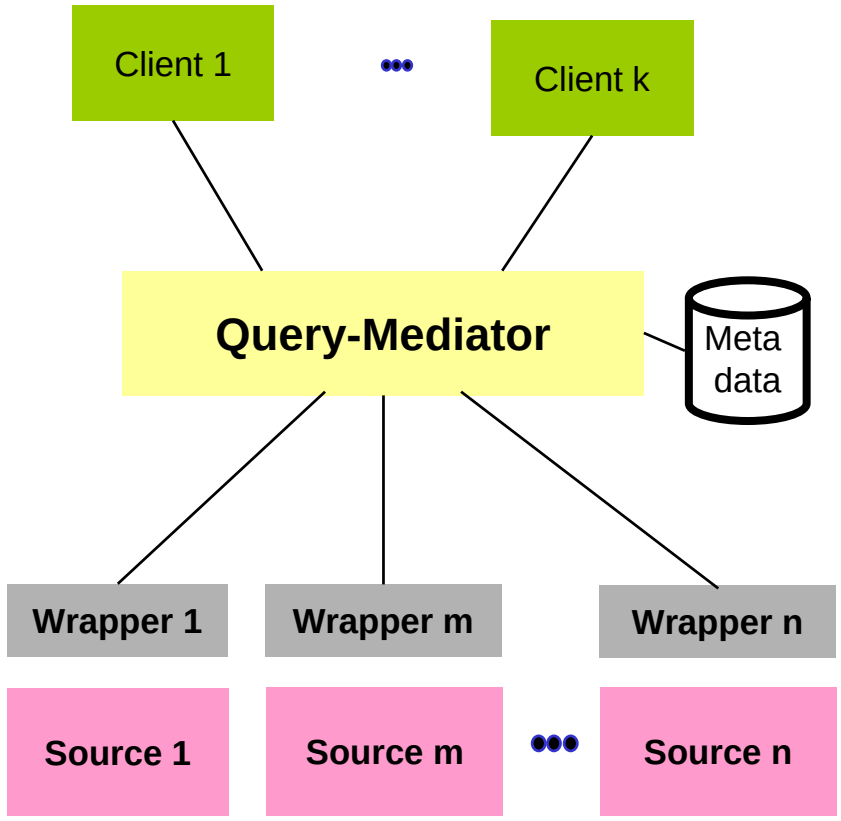
- Feedback / Diskussion

Physische vs. virtuelle Datenintegration

Data Warehousing, ETL
(physische Datenintegration)



Enterprise Information Integration, EII
(virtuelle Integration, Query–Mediatoren)



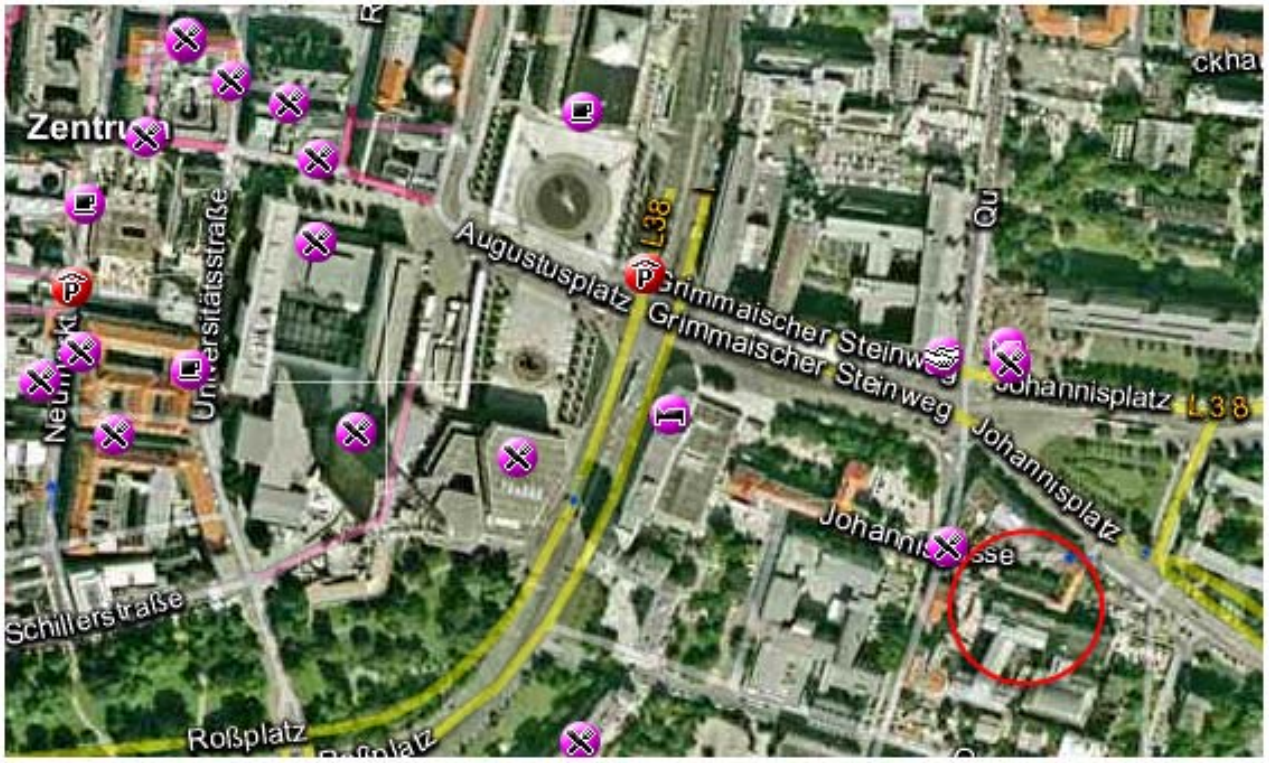
Probleme ETL / EII

- Hoher Realisierungsaufwand, lange Vorlaufzeiten
  - Einheitliche Sicht auf heterogene und verteilte Datenbestände
  - Aufwändige Erstellung und Wartung
    - für gemeinsames Schema sowie
    - für Mappings zu Quell-Schemas

- Unzureichende Unterstützung für Webdaten
  - Beschränkte Zugänglichkeit über Suchformulare bzw. Suchmaschinen
  - Größere Heterogenität der Daten
  - Geringe Datenqualität

Mashups – Datenintegration

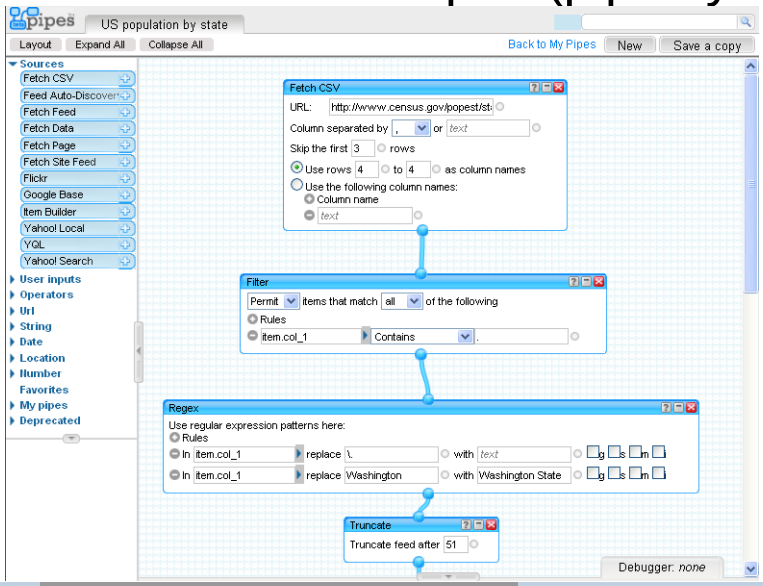
Mashup = interaktive Web-Applikation zu Kombination von Daten mehrerer Online-Datenquellen im Rahmen eines neuen Dienstes

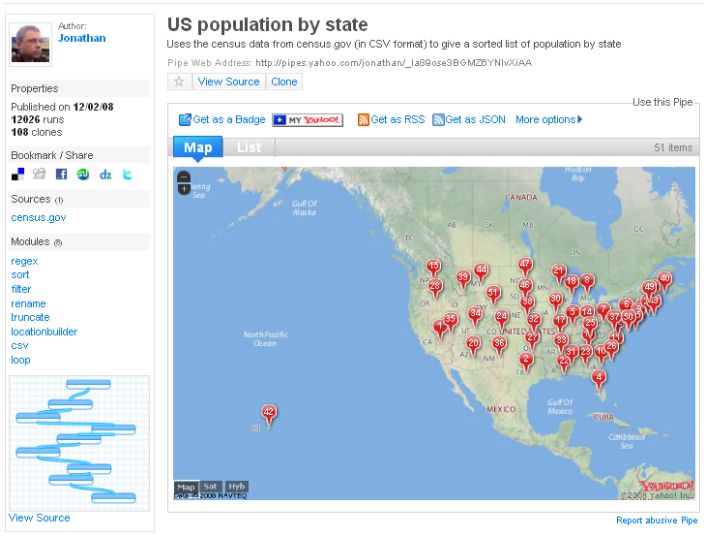


www.goyellow.de

Mashup-Tools

- Basieren oft auf RSS-Feed
- Angepasst auf einige konkrete Datenquellen mit meist geringen Datenmengen
- Schnell zu erstellen
- Einfache Operationen mit Daten möglich z.B. Filtern, Sortieren, Teilen
- Z.B. Yahoo-Pipes (pipes.yahoo.com)





Mashups: Merkmale und Probleme

- Schnelle Entwicklung einfacher Workflows
  - zum Teil Erstellung über GUI; z.B. Yahoo Pipes
- Nutzung von Web-APIs / Web Services
- Einfache Datenintegration
  - z.B. über geographische Koordinaten
  - Kombination von Suchergebnissen zu Keyword
- Einschränkungen
  - Einfache Anfragen (keine Transformation für unterschiedliche Quellen)
  - Einfache Ergebnisnachbearbeitung (Merge statt Match)
  - Unzureichende Vollständigkeit / Genauigkeit
  - Geringe Datenqualität (Dubletten)
  - Unternehmenstauglichkeit

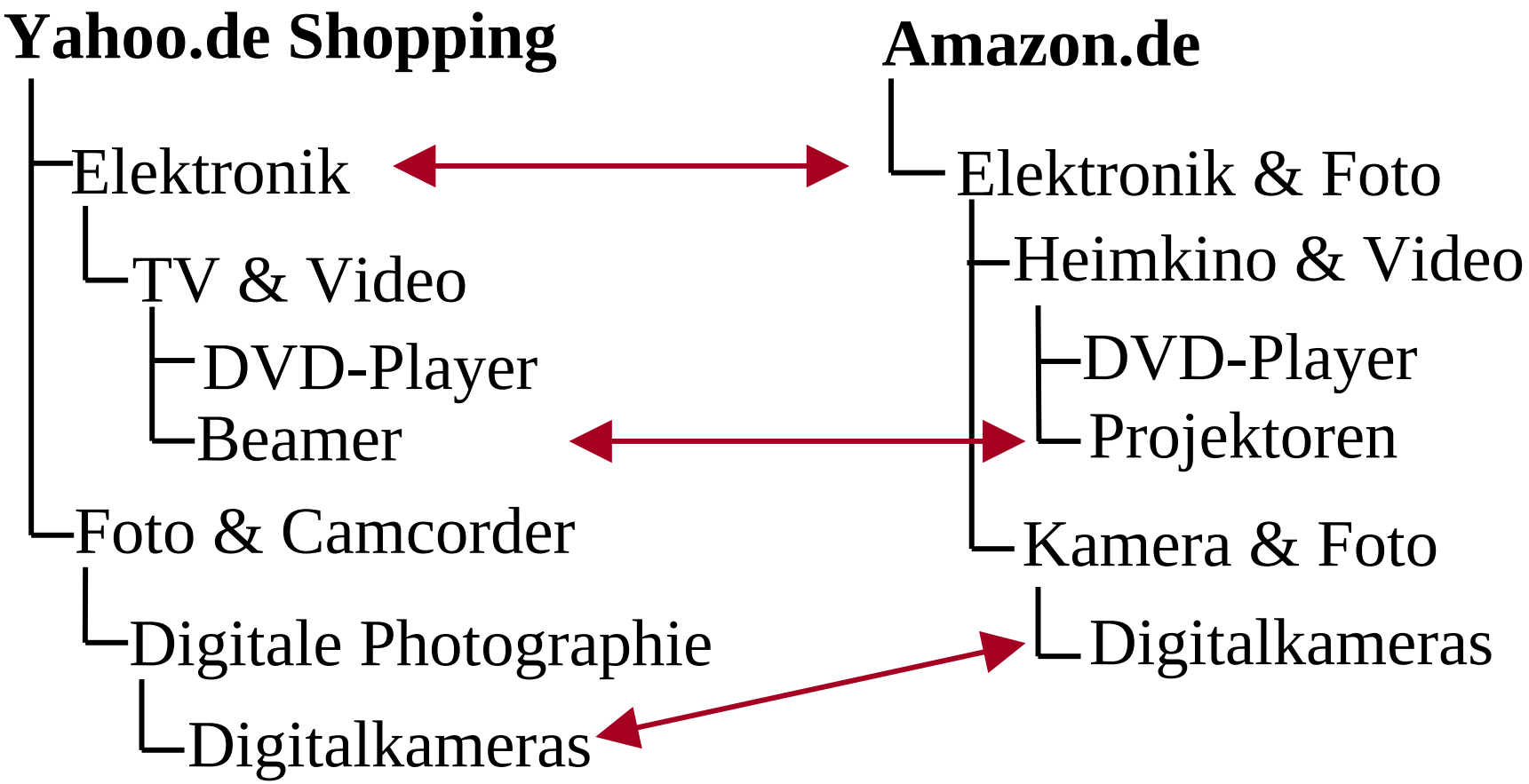
Vortragsinhalte

- Einleitung
- Derzeitige Ansätze zur Datenintegration
  - Warehousing, EII
  - Mashups
- Eigene Forschungsarbeiten / Prototypen
  - Schema/Ontologie Matching: COMA++
  - Objekt-Matching: MOMA, STEM
  - Workflowbasierte Integration: iFuice
- Feedback / Diskussion

Schema / Ontologie-Matching

- Finden semantischer Korrespondenzen zwischen 2 Schemas bzw. Ontologien
  - DB-Schemas, XML-Schemas
  - Inhaltskategorisierungen, Produktkataloge, Suchformulare, …
- Input: 2 Schemas (Ontologien) S_1 and S_2
  - Evtl. Dateninstanzen zu S_1 und S_2
  - *Hintergrundwissen*
- Output: *Mapping* zwischen S_1 und S_2
- Kritischer Schritt in zahlreichen Applikationen
  - Datenintegration: Mappings zwischen Datenquellen bzw. zwischen Datenquelle und globalem Schema
  - E-Business: XML-Datentransformation
  - Katalogintegration, Finden verwandter Inhalte …

Match-Beispiel: Produktkataloge

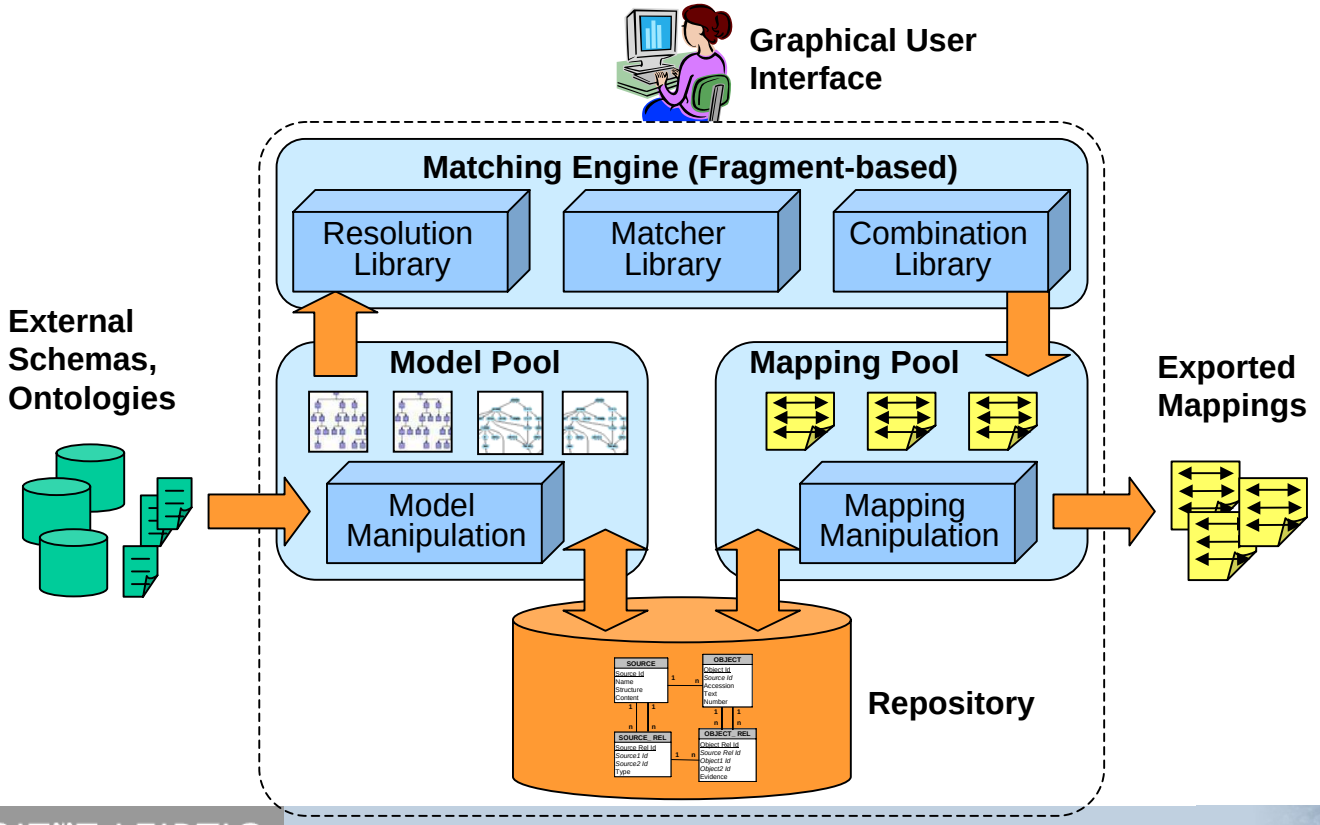


Derzeitige Situation

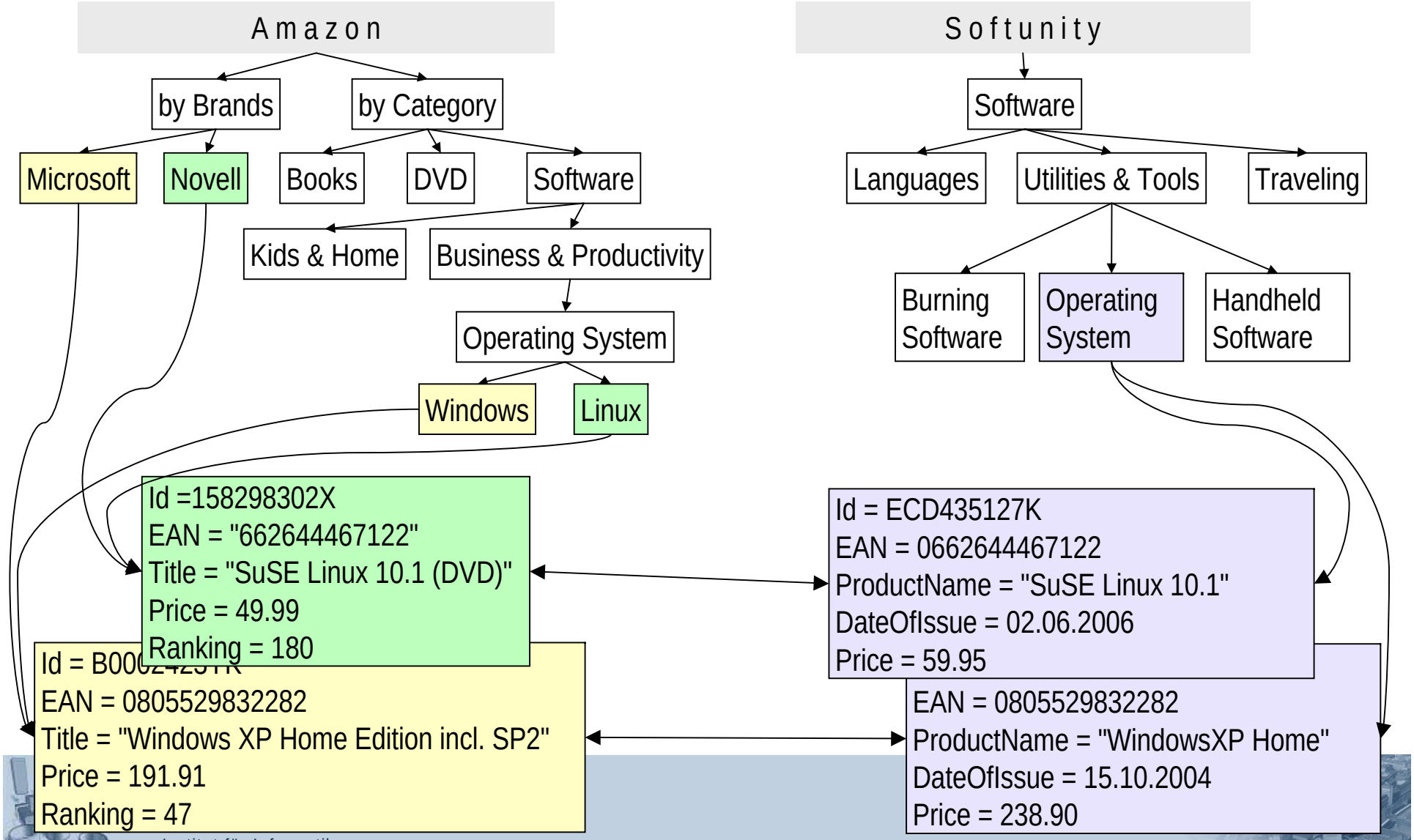
- Finden der Mappings ist Flaschenhals
  - Weitgehend manuell
  - Aufwändig und fehleranfällig
- Zunehmende Dringlichkeit
  - Wachsende Nutzung verteilter Daten
  - Starke Zunahme an XML-Daten und -Schemas
  - Zunahme an Ontologien / semantischen Kategorisierungen
- Skalierbarkeit erfordert semi-automatische Tools
  - Vollautomatische Lösungen aufgrund semantischer Heterogenität nicht perfekt
  - Namensproblematik (Synonyme, Homonyme)
  - begrenzte Mächtigkeit von Metadaten / Schemasprachen

Generische Match-Plattform

- Unterstützt relationale und XML-Schemas, OWL-Ontologien
- Flexible Kombination mehrerer Match-Algorithmen
- Match-Strategien für große Schemas
- Wiederverwendung vorhandener Match-Ergebnisse



Instanz-basiertes Ontologie-Matching



Objekt-Matching-Problem

- Identifikation semantisch äquivalenter Objekte
  - innerhalb einer Datenquelle oder zwischen verschiedenen Quellen
  - um Objekte zu integrieren/mischen, zu vergleichen, Dubletten zu eliminieren, etc.

Quelle1: Kontakt

KID	***Name***	***Strasse***	***Stadt***	***Frau***
11	Kristen Schmid	Hanse Pl 2	Berlin	1
24	Christian Schmied	Hanse Str 2	Berlin	0

Quelle2: Kunde

Knr	***Nachname***	***Vorname***	***Geschl***	***Adresse***	***Telefon***
11	Schmid	Chris	M	Hansestr. 2, 18182 Bentwich	
493	Schmid	Kris L.	W	Hansa-Platz 2, 10557 Berlin	030-9627621

Objekt-Matching-Problem für Webdaten

Viele heterogene Dubletten
- unterschiedliche Namen
- nur teilstrukturiert
- unterschiedliche Formate
- heterogene Inhalte ...

CANON VIXIA HV30
CANON VIXIA HV30 _zustand: Neu - price inkl.
Zur Einkaufsliste hinzufügen

€656,90
www.ATLANTIV.com

Canon VIXIA HV30 High Definition Camcorder (PAL)
Canon VIXIA HV30 High Definition Camcorder **Canon HV30** inherits the HV20's **Canon** HD Camera System – the unique combination of a genuine **Canon** HD Video Lens, **...**.
★★★★½ 285 Beurteilungen - Zur Einkaufsliste hinzufügen

€814,48
DSLR Cameras Division - Phones Corporation Ltd.

Canon VIXIA HV30 DV Camcorder
Canon VIXIA HV30 DV Camcorder Digital, DV, NTSC, Bis zu 3.15 Megapixels Einzelbild, 10x Optischer Zoom, 0,5 kgMit dem HD-Camcorder **HV30** präsentiert **Canon ...**.
Zur Einkaufsliste hinzufügen

€749,00
komponentenfabrik.de

CANON VIXIA HV30 Ladegeräte, **VIXIA HV30** Ladegeräte
CANON VIXIA HV30 Ladegeräte, neu Ladegeräte für **CANON VIXIA HV30**, 1 Jahr Garantie, Akku lokalisiert in Deutschland.
Zur Einkaufsliste hinzufügen

€19,01
kaufen-akkus.de

Canon VIXIA HV30 Camcorder
The **Canon VIXIA HV30** Camcorder is a versatile High Definition camcorder which **...** The **Canon VIXIA HV30** lets moviemakers adapt the camcorder to their needs **...**.
Zur Einkaufsliste hinzufügen

€1.139,22
DigitalRev.com
3 Händlerbewertungen

Google Produktsuche

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

von €532 bis €873 bei 76 Händ

Flash card, SD Memory Card, SDHC-Speicherk

Der attraktive HF100 speichert ganz bequem 1.9
dabei einfach atemberaubend brillante HD Videc

Der HF100 speichert HD-Videos direkt auf SDH
teilen. Die Karten lassen sich nämlich prima ver

Der HF100 ist ein Camcorder im superkompakte
Speicherung ohne mechanisch bewegliche Kom

Der HF100 zeichnet in 1.920 x... mehr »

Zur Einkaufsliste hinzufügen

Preise vergleichen | **Technische Spezifikationen** | Verw

Allgemein - Hauptmerkmale - Speicher / Speicher - Objektivsystem - Zusät
Batterie - Universal Product Identifiers

Allgemein

Produkttyp	Camcorder
Breite	7.3 cm
Tiefe	12.9 cm
Höhe	6.4 cm
Gewicht	380 g

Hauptmerkmale

Sensorauflösung Camcorder	3.31 Mpix
Effektive Videoauflösung Camcorder	2.07 Mpix
Effektive Fotoauflösung 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 1/3,20"
- 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 Ø: ★★★★½ (15 Meinungen)
Testnote Ø: 1,8 (16 Testberichte)

Hersteller: Canon

Alle Angaben ohne Gewähr

PREIS SUCHMASCHINE

Abbildung ähnlich

Canon HF100 günstig bei ebY ersteigern

Preisvergleich ab € 553,95 | **Meinungen** | Preisverlauf

Meinung zu Canon HF100

Klasse Cam 09.09.2008

Hallo, nach Vergleich mit einer Panasonic 3CCD (HDC-SD9) muss ich sagen, das im Low-Light-Bereich die Canon mehrere Nasen Vorsprung vor der 3CCD-Pansonic 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. Insbesonders 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 Weissableicheinstellungen, oder den manuellen Weissabgleich bevorzugen (wenn möglich). Ich werde die Canon behalten und die Panasonc zurückgeben. Für das Geld gibts z.Zt. nichts besseres.

Meine Wertung: ★★★★★ **Autor: skeptiker**

Canon HF100 Test (Camcorder)

Merken und Vergleichen

Testberichte.de

Bild 1/2

GUT **1,6**

586,95 Euro bei amazon.de

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

Produktdaten: 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 Kantenabhebung, bei der auffällige Schärfungssäume auftreten. ..."

586,95 bei Amazon.de Camcorder suchen bei Otto.de

Im Testbericht lesen

Download (3,00 €)

video Ausgabe 8/2008 Platz 1/3 im Test

„sehr gut" (75 von 100 Punkten)

Preis/Leistung: "hervorragend" „Testsieger"

„Plus: schärfstes Bild im Test; manuelle Tonaussteuerung; Mikro- und Kopfhörerbuchse; optischer Bildstabilisator. Minus: Zubehörschuh nur für Canon-Equipment."

586,95 bei Amazon.de Camcorder suchen bei Otto.de

Download (1,20 €)

Informationsfusion

UNIVERSITÄT LEIPZIG

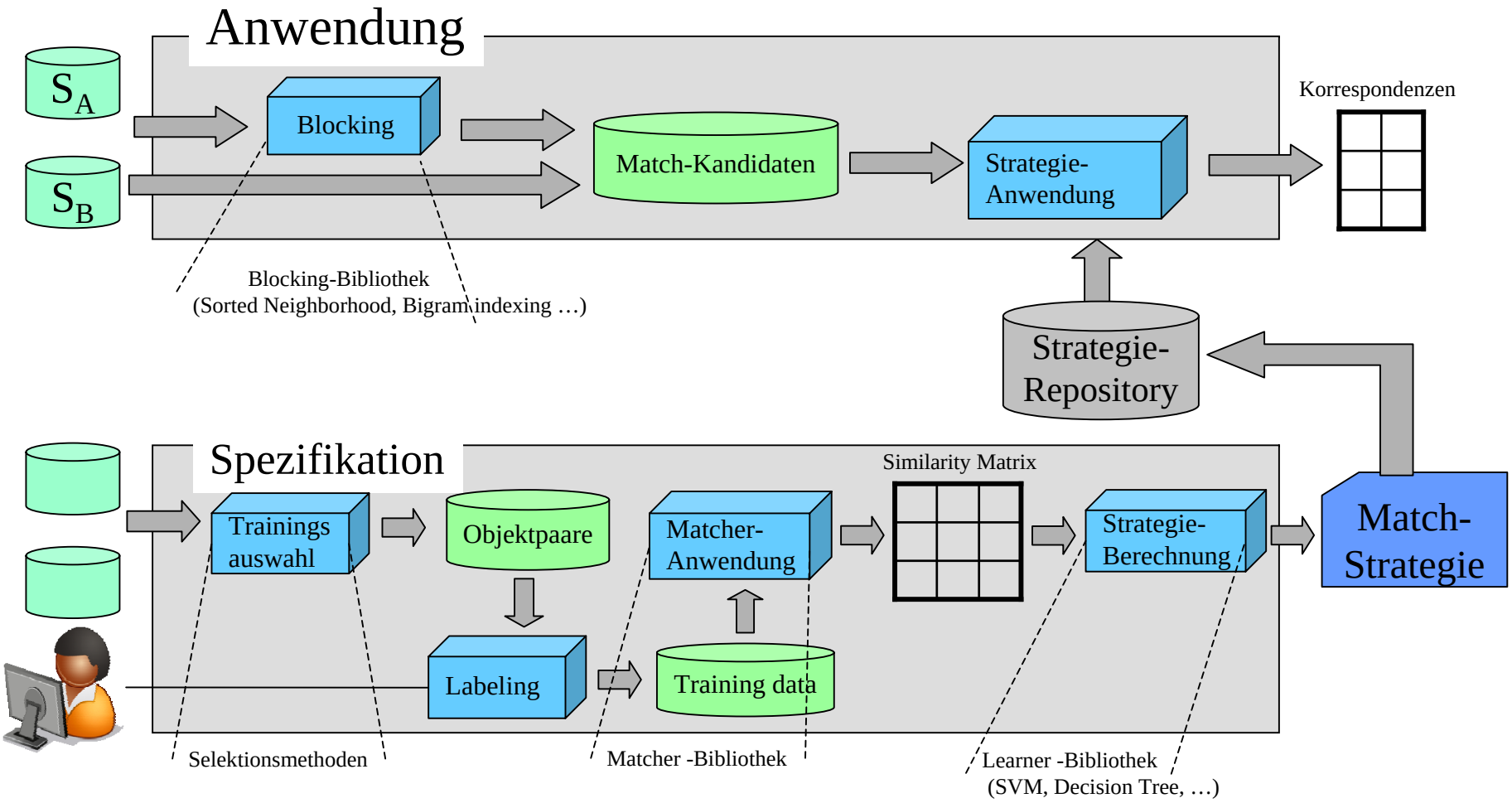
Abteilung Datenbanken
am Institut für Informatik

Prototypen zum Objekt-Matching

- Aufgabe: Erkennung äquivalenter Objekte (Dubletten) in Eingabedaten
  - Ergebnis: Mapping mit Korrespondenzen
- MOMA = **M**apping based **O**bject **Ma**tching
- STEM = **S**elf-**T**uning **E**ntity **M**atching
- Kombinierte Verwendung mehrerer Match-Algorithmen
  - String-Vergleich zwischen Attributwerten
  - Auswertung von Nachbarschaftsbeziehungen
- STEM: Nutzung von *Machine Learning-Verfahren* zur automatischen Parameteroptimierung (z.B. Schwellwerte für Attributähnlichkeiten)

LDS_A	$LDS_{A'}$	Sim
a_1	a'_1	1
a_2	a'_1	0.9
a_3	a'_3	0.8

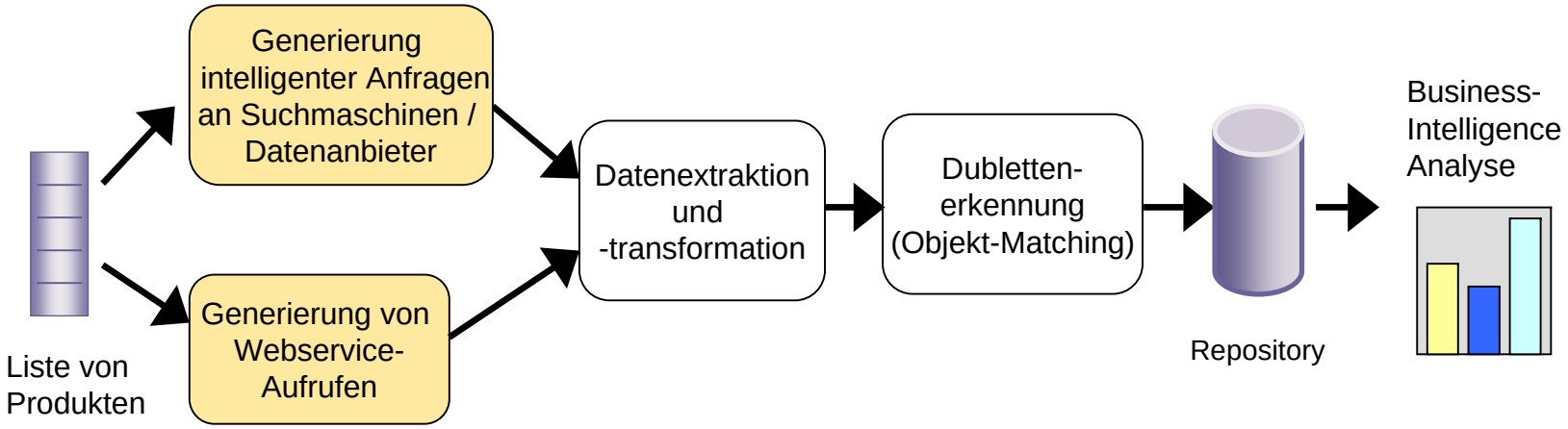
STEM Architektur



Workflow-artige Datenintegration mit iFuice

- iFuice-Framework: Erweiterung des Mashup-Ansatzes zur schnellen Realisierung von Datenintegrationsaufgaben
- Weitgehende Nutzung existierender Services, bereits integrierter Datensammlungen bzw. Portale
- Programmierung von Workflows durch Skript-Programme mit mächtigen mengenorientierten Operatoren (Anfragen, Matching, Mapping-Traversierung, Mengenverarbeitung etc.)
- Iterative Verbesserung von Suchanfragen zur Ergebnisverbesserung (Vollständigkeit, Genauigkeit)

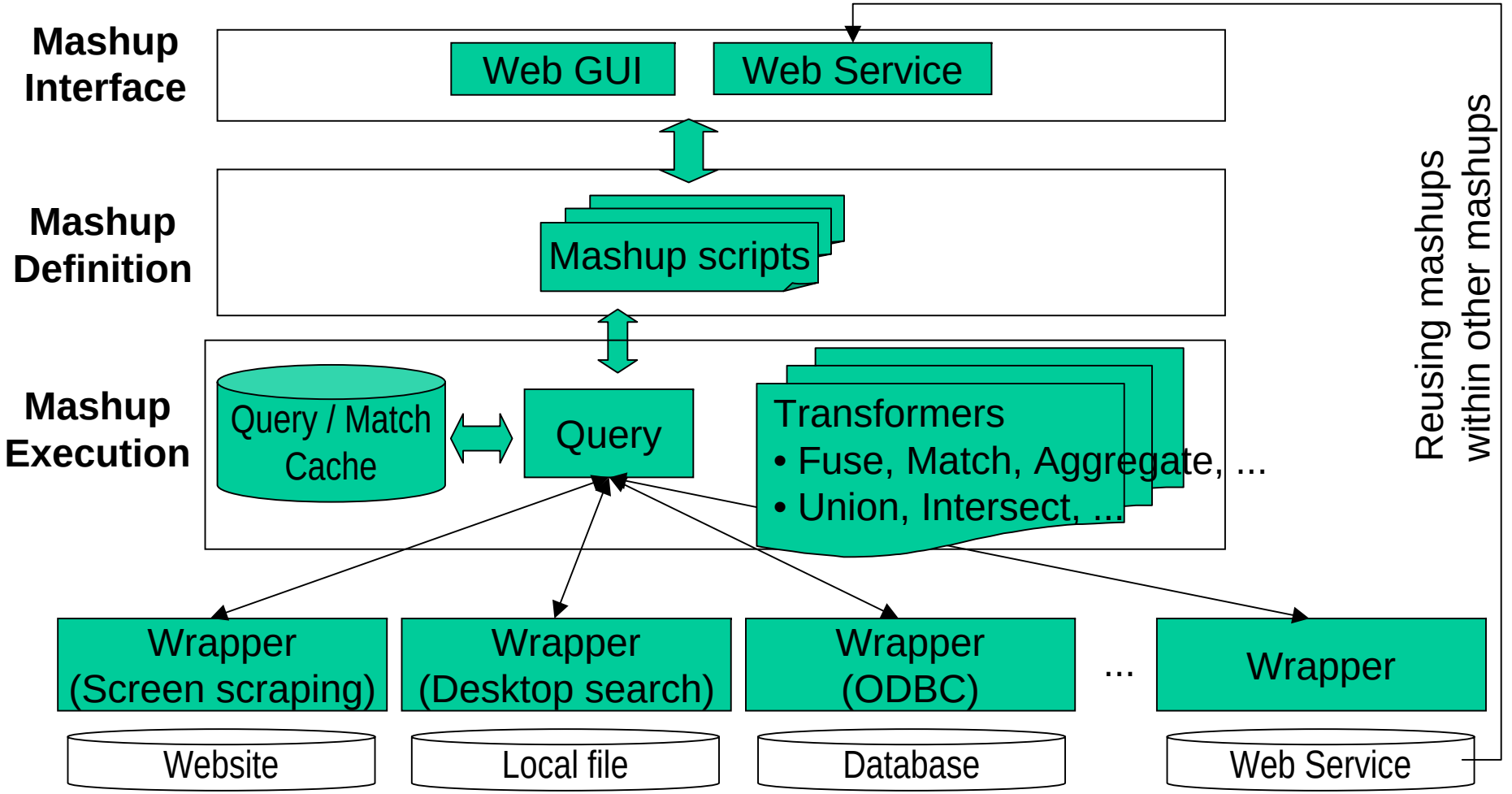
Beispiel-Workflows



Anwendungsbeispiele:

- Suche nach Zusatzinformationen (z.B. Nutzerbewertungen) in verschiedenen Quellen zu bestimmten Produkten
- Dynamische Preis/Verfügbarkeitsanalyse z.B. im Rahmen einer Metasuchmaschine (Hotels, Autos, etc.) auf Basis einer zuvor erstellten Präferenzliste
- Kopplung mehrerer Metasuchmaschinen: Flug, Hotel, Leihwagen
- Interaktive Recherche nach lokalen Anbietern bzw. Lieferanten bestimmter Dienstleistungen / Produkte

iFuice-Architektur



Vortragsinhalte

- Einleitung
- Derzeitige Ansätze zur Datenintegration
  - Warehousing, EII
  - Mashups
- Eigene Forschungsarbeiten / Prototypen
  - Schema/Ontologie Matching: COMA++
  - Objekt-Matching: MOMA, STEM
  - Workflowbasierte Integration: iFuice
- Feedback / Diskussion

Innovationslabor

- an Univ. Leipzig angesiedelt
- Erprobung / Weiterentwicklung der Techniken für Praxis/Unternehmenseinsatz
  - Workflow-basierte Datenintegration
  - Objekt-Matching / Dublettenbehandlung
  - Schema/Ontologie-Matching
- Kopplung mit verfügbaren Werkzeugen, z.B. von Datenbank/BI-Anbietern bzw. Open-Source-Lösungen
- Vorbereitung eventueller Ausgründungen

Feedback / Diskussion

- Wie beurteilen Sie die vorgestellten Verfahren (Stärken, Schwachstellen, Erweiterungsbedarf) ?
- Sehen Sie Anwendungsmöglichkeiten für die vorgestellten Verfahren?
  - zur Lösung konkreter Integrationsaufgaben
  - zur Kombination mit eigenen Entwicklungen/Produktangeboten
- Wie schätzen Sie die Marktrelevanz der Lösungsansätze ein?
- Welche Kooperationsformen mit dem Innovationslabor können Sie sich vorstellen, z.B.
  - gemeinsam betreute studentische Arbeiten
  - gemeinsame Projekte, z.B. zur Pilotanwendung der Werkzeuge
  - Beratung zu Datenintegrationsthemen