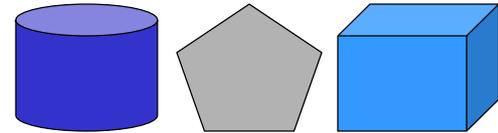


Datenintegration

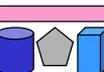
Datenintegration



Kapitel 1: Einführung

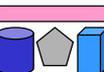
Dr. Anika Groß
Sommersemester 2016

Universität Leipzig
Institut für Informatik
<http://dbs.uni-leipzig.de>



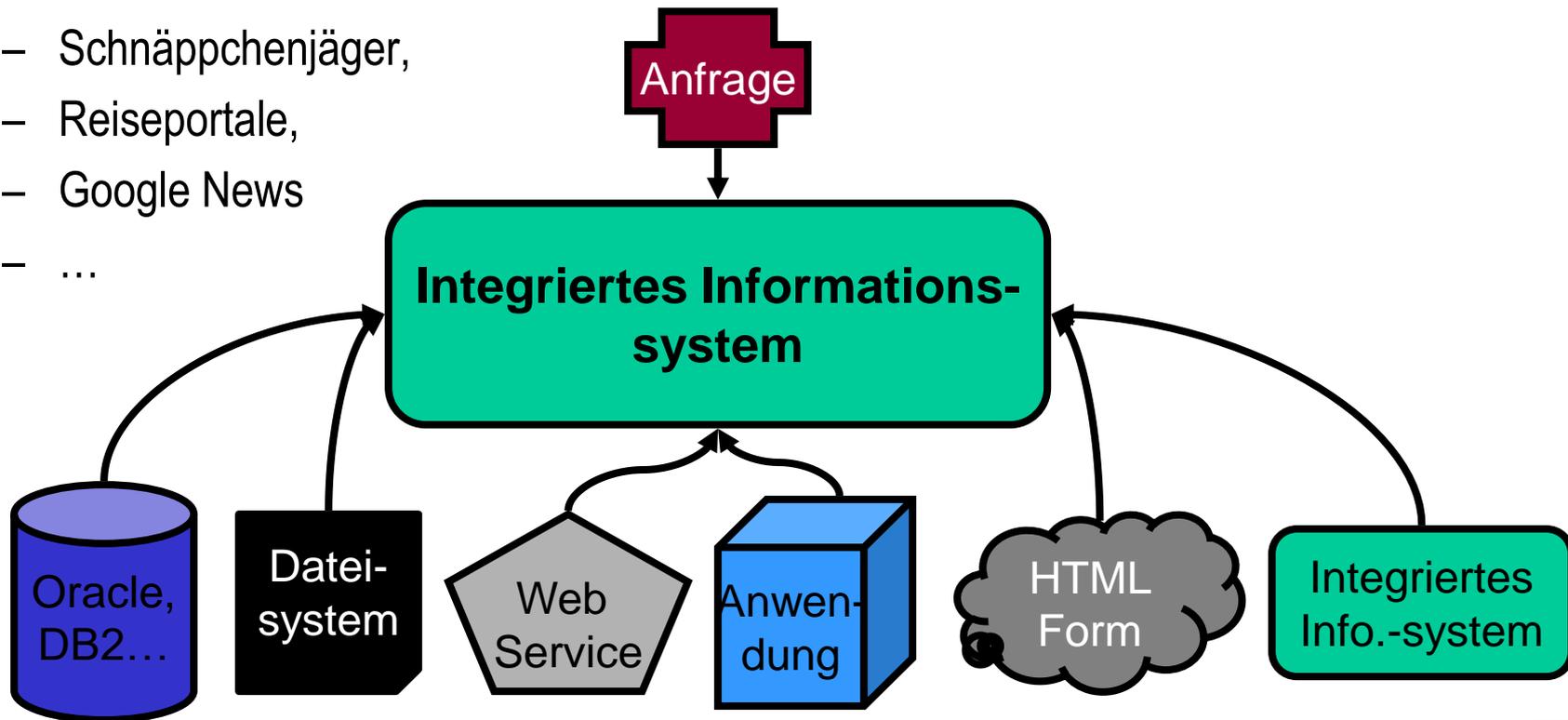
Inhalt

- Begriffsdefinition
- Anwendungsgebiete
- Informationssysteme und integrierte Informationssysteme
- Integration am Beispiel



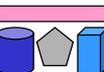
Integrierte Informationssysteme

- Zusammenführung von Daten und Inhalt verschiedener Quellen zu einer einheitlichen Informationsmenge
- Beispiele
 - Metasuchmaschinen
 - Data Warehouses
 - Schnäppchenjäger,
 - Reiseportale,
 - Google News
 - ...



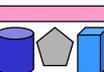
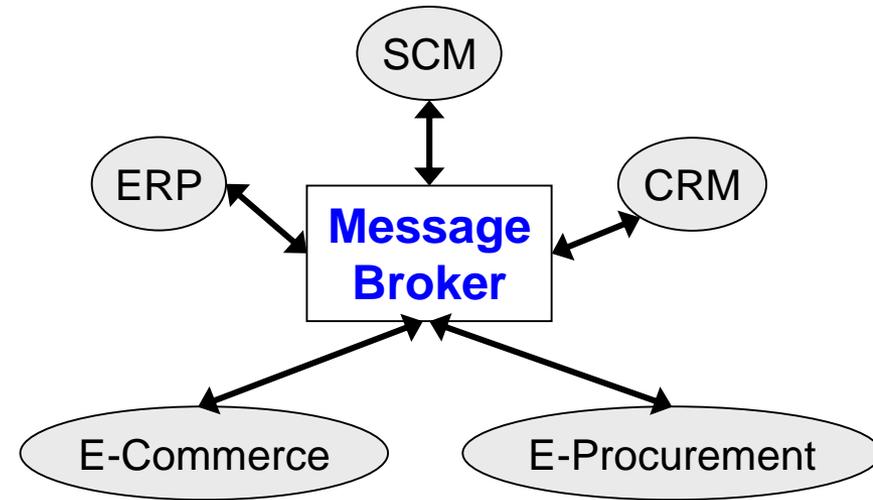
Daten-/Informationsintegration

- Informationsintegration ist die korrekte, vollständige und effiziente Zusammenführung von Daten und Inhalt verschiedener, heterogener Quellen zu einer einheitlichen und strukturierten Informationsmenge zur effektiven Interpretation durch Nutzer und Anwendungen.
- Begriffe “Datenintegration” und “Informationsintegration” werden synonym gebraucht
 - Informationsintegration = Integration der Metadaten und der Instanzdaten
- Ziel: Mehrwert, der durch Kombination von Daten entsteht
 - Anfragen, die “bessere” Ergebnisse durch Verwendung mehrerer (anstatt nur einer) Datenquellen liefern
 - Anfragen, die nur durch Verwendung mehrerer Datenquellen beantwortet werden können

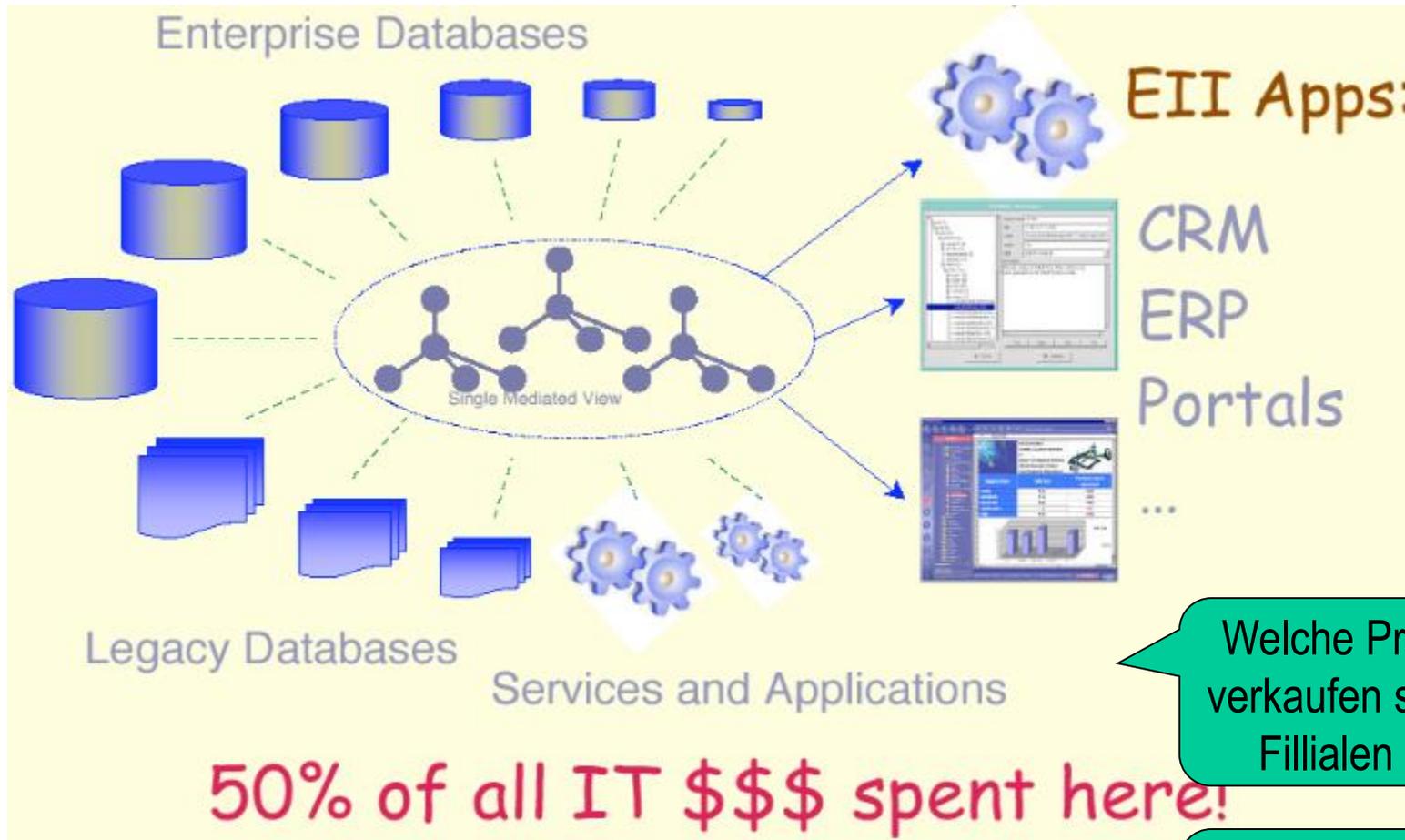


Vergleich: Enterprise Application Integration

- „Verwandt, aber anders“
 - Enterprise Application Integration
 - Middleware (CORBA, J2EE, .Net, ...)
 - Systemintegration
 - Business Process Integration
- Enterprise Application Integration
 - Nachrichtenbasiert, keine Anfragen
 - Informationsverteilung
 - Aktion beim Eintreten eines Ereignisses
- Information Integration
 - Anfragebasiert
 - Annahme eines (praktisch) statischen Datenbestands
 - Aktion
 - Erst bei Anfrage (virtuelle Integration)
 - In regelmäßigen Zyklen (materialisierte Integration)



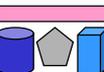
Anwendungsgebiet 1: Business



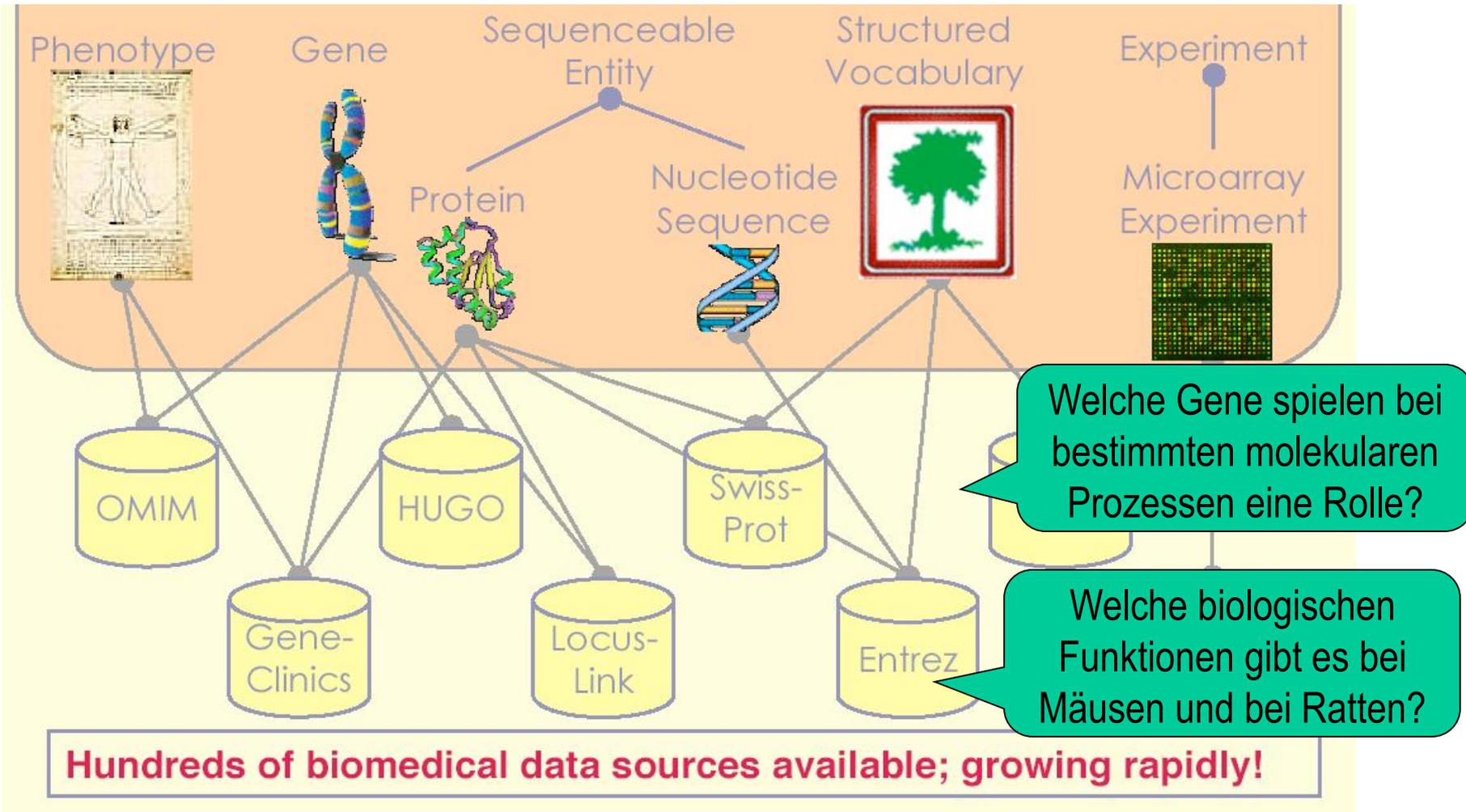
Welche Produktgruppen verkaufen sich in welchen Filialen am besten?

Wie erfolgreich sind unsere Marketing-Kampagnen?

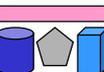
Alon Y. Halevy: Structures, Semantics and Statistics. VLDB 2004



Anwendungsgebiet 2: Wissenschaft



Alon Y. Halevy: Structures, Semantics and Statistics. VLDB 2004



Informationssystem: Swissprot-Datei

```
ID  RNGTPCHI  standard; RNA; ROD; 1016 BP.
XX
DT  01-AUG-1991 (Rel. 28, Created)
DT  04-MAR-2000 (Rel. 63, Last updated, Version 2)
XX
DE  Rat GTP cyclohydrolase I mRNA, complete cds.
XX
KW  GTP cyclohydrolase I.
XX
OS  Rattus norvegicus (Norway rat)
OC  Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC  Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Rattus.
XX
RN  [1]
RP  1-1016
RX  MEDLINE; 91093270.
RX  PUBMED; 1985963.
RA  Hatakeyama K., Inoue Y., Harada T., Kagamiyama H.;
RT  "Cloning and sequencing of cDNA encoding rat GTP cyclohydrolase I: The
RT  first enzyme of the tetrahydrobiopterin biosynthetic pathway";
RL  J. Biol. Chem. 266(2):765-769(1991).
XX
FT  CDS          128..853
FT                /codon_start=1
FT                /db_xref="GOA:P22288"
FT                /db_xref="SWISS-PROT:P22288"
FT                /EC_number="3.5.4.16"
FT                /gene="GTP cyclohydrolase I"
FT                /product="GTP cyclohydrolase I"
FT                /protein_id="AAA41299.1"
FT                /translation="MEKPRGVRCCTNGFPERELPRPGASRPAEKSRPPEAKGAQPADAWK
FT                AGRPRSEEDNELMLPNLAAAYSSILRSLGEDPQRQGLLKTWPRAATAMQFFTKGYQETI
FT                SDVLDNAIFDEHDHENVIKDIDMFSMCEHHLVPFVGRVHIGYLPNKQVGLGSKLARIV
FT                EIYSRRLQVQERLTKQIAVAITEALQPAGVGVVIEATHMCMVMRQVQKMNSTVTSTML
FT                GVFREDPKTREFFLTLIRS"
SQ  Sequence 1016 BP; 236 A; 279 C; 291 G; 210 T; 0 other;
    gacttcgaac ctcattcggg gcagaactcc tgtcccgggtg acagccacag gtcacggccg      60
    ccggctaagc cgagccgcag cgcttggttag caccttaggg tgtctcggga gcaatcggcg      120
    cgggtccatg gagaagccgc ggggtgtaag gtgcaccaat gggttccccg agcggggagct      180
    ...
    catcaggagc tgaacttccg tgtgcgagcc ccggtttgca gacccccgct gaggccagcg      900
    ttatctgtct cgattgtaca ttccagttcc agttggtata cttgtcaact ttattttctca      960
    ccatgaattg tattaataa ttatttatag agatgtcaaa taaagtgat caact          1016
```

Molecule type
Name

Date of creation and last update

Free text description

Keywords describing the molecule

Organism

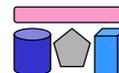
Article the sequence
was published in

Structural annotation (coding sequence)

Link to functional annotation
of resulting protein

Translated
protein
sequence

Sequence
of
bases



Informationssystem: Amazon Suchformular

The image shows a screenshot of the Amazon.de website's search interface for books. The browser window is titled "Amazon.de: Erweiterte Suche: Bücher" and the address bar shows the URL "www.amazon.de/Erweiterte-Suche-Bücher/b?ie=UTF8&node=124535011". The page features the Amazon.de logo, navigation links like "Mein Amazon", "Sonderangebote", and "Wunschzettel", and a search bar with "Bücher" entered. Below the search bar, there are tabs for "Bücher", "Erweiterte Suche", "Stöbern", "Bestseller", "Neuheiten", "Hörbücher", "Englische Bücher", "Taschenbücher", "Fachbücher", and "Sonderangebote". A banner for "Elektro-Großgeräte bei Amazon.de versandkostenfrei" is visible. The main content area is titled "Erweiterte Suche Bücher" and contains a search form with various filters and a "Jetzt suchen" button.

Erweiterte Suche

Bücher

- Zeitschriften
- Fremdsprachige Bücher
- Elektronik & Foto
- Computer & Zubehör
- Bürobedarf & Schreibwaren
- Musikinstrumente & Equipment
- Musik
- Klassische Musik
- MP3-Downloads
- DVD
- Software
- PC- & Video-Games

Erweiterte Suche Bücher

Suchen Sie nach fremdsprachigen Büchern? [Klicken Sie hier](#)

Suchbegriffe

Autor

Titel

ISBN (10- oder 13-stellig, ohne Bindestriche)

Verlag

Kategorie

Format

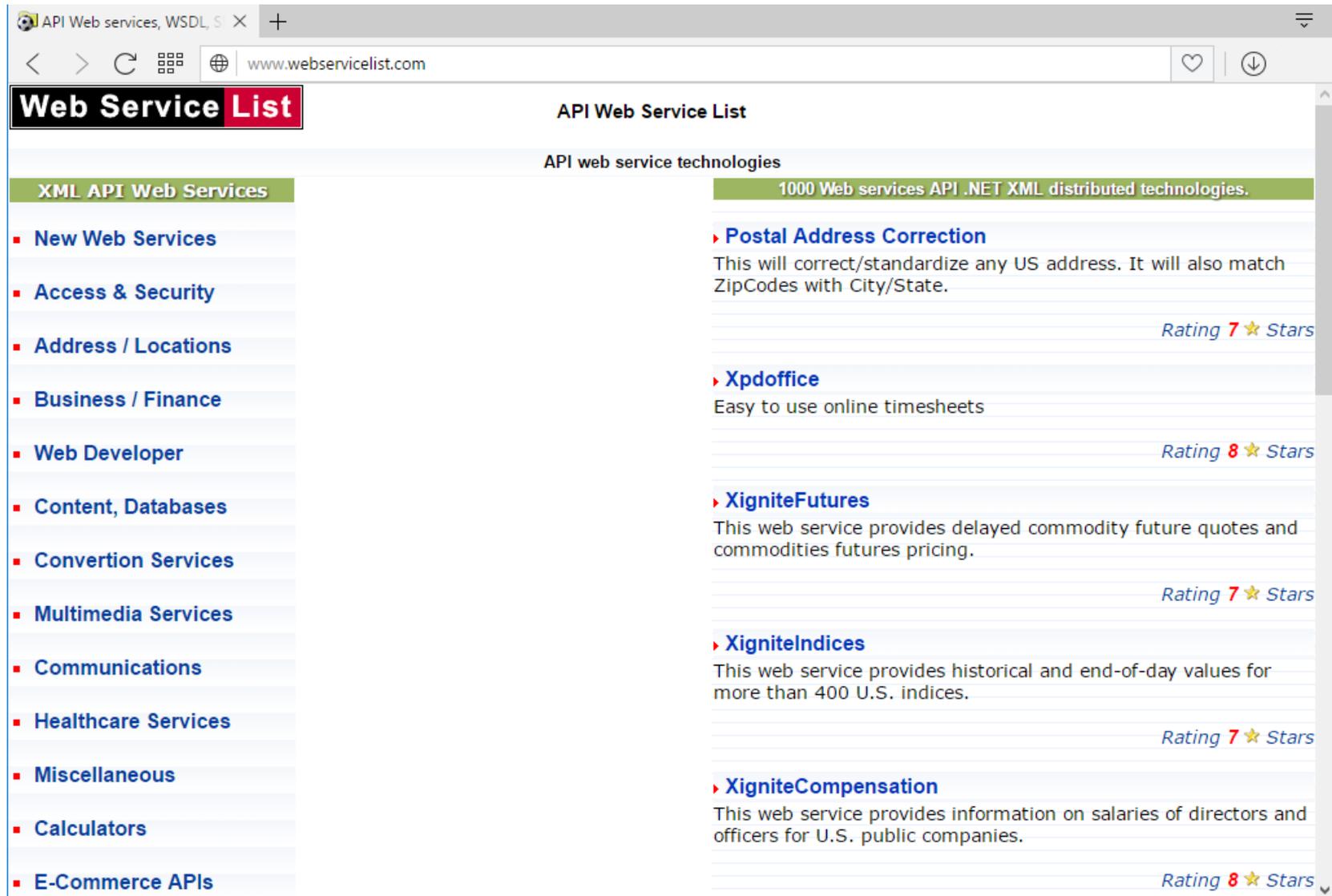
Erscheinungsdatum Monat Jahr

Anbieter

Sortieren nach

Jetzt suchen

Informationssystem: Web Services



API Web services, WSDL, S X +

< > ↻ 📄 🌐 www.webservicelist.com

Web Service List

API Web Service List

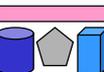
API web service technologies

XML API Web Services

- New Web Services
- Access & Security
- Address / Locations
- Business / Finance
- Web Developer
- Content, Databases
- Conversion Services
- Multimedia Services
- Communications
- Healthcare Services
- Miscellaneous
- Calculators
- E-Commerce APIs

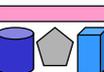
1000 Web services API .NET XML distributed technologies.

- Postal Address Correction**
This will correct/standardize any US address. It will also match ZipCodes with City/State.
Rating 7 ★ Stars
- Xpdoffice**
Easy to use online timesheets
Rating 8 ★ Stars
- XigniteFutures**
This web service provides delayed commodity future quotes and commodities futures pricing.
Rating 7 ★ Stars
- XigniteIndices**
This web service provides historical and end-of-day values for more than 400 U.S. indices.
Rating 7 ★ Stars
- XigniteCompensation**
This web service provides information on salaries of directors and officers for U.S. public companies.
Rating 8 ★ Stars



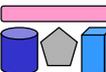
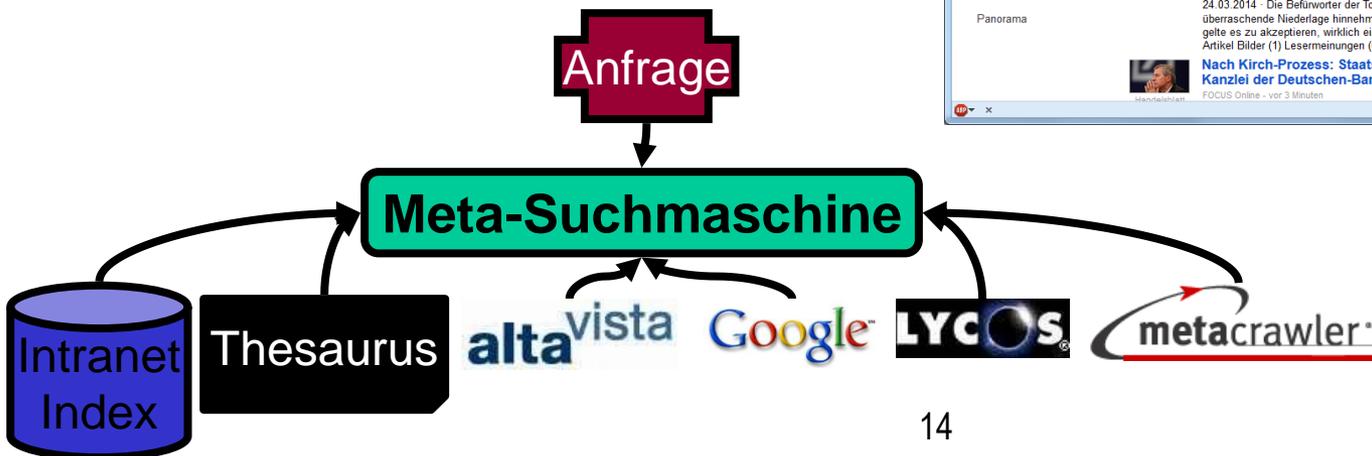
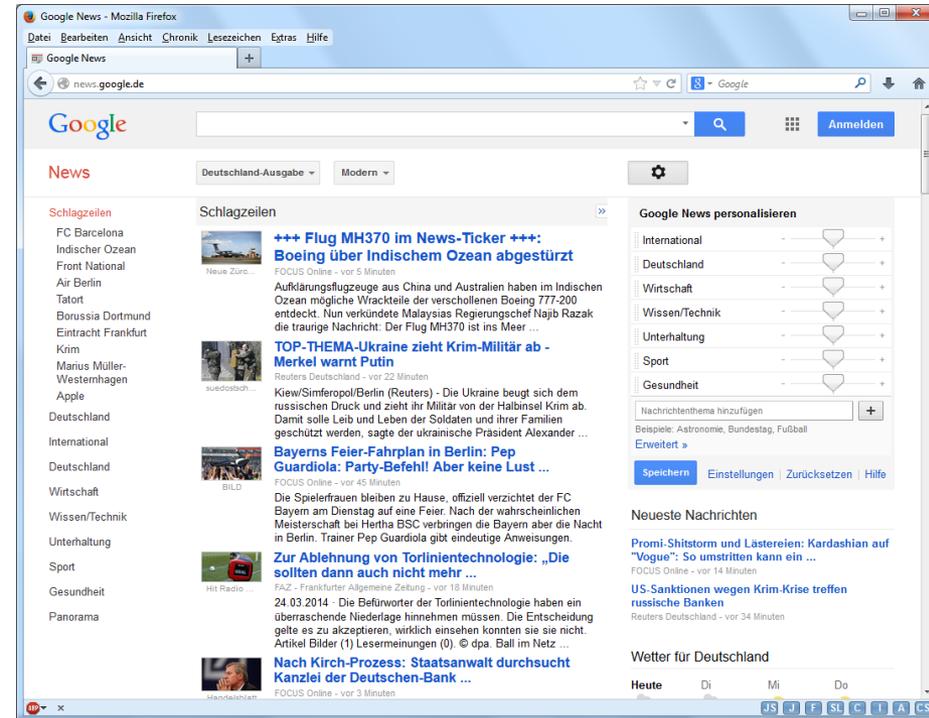
Informationssysteme: Übersicht (Auswahl)

| System | Informationseinheit | Anfrage | Struktur | Beispiele |
|---------------|-------------------------|--------------------------|-----------------------------------|--------------------------------------|
| Dateisystem | Flat file | File search, RegEx | Flach, hierarch. | NTFS, FTP |
| Datei | Zeile, Token | Parser | flach | CSV, Annotated Files |
| Markup-Datei | Tagged Text | Parser, Anfragesprache | Flach, hierarch., graphbasiert | XML, HTML |
| Datenbank | Tupel, Attribut, Objekt | Anfragesprache (SQL) | Relational, OO, hierarch. | RDBMS, OODBMS, XMLDBMS |
| HTML Formular | HTML Seite | Suchworte, Formular | Hierarch. | Such- und Anfrageformulare |
| Web Service | XML | XML | Hierarch. (Ergebnis) | Einfache Dienste, komplexe Workflows |
| Anwendung | Java-Objekt, Text | Anwendungsinterface, GUI | Objekt (Interface), Display (GUI) | Java, C++ |



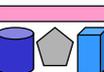
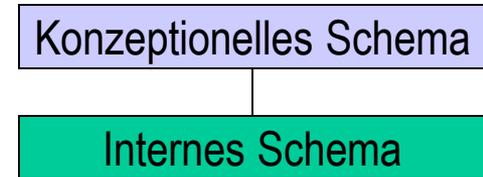
Integriertes Informationssystem

- Verhält sich in Anfrage, Struktur und Informationseinheit je nach Design:
 - DBMS, HTML Formular, Web Service, ...
- Beispiele
 - Data Warehouses
 - Föderierte Datenbanken
 - Portale, News-Aggregatoren
 - Meta-Suchmaschine
 - ...



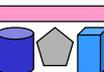
Integration = Abstraktion

- Logisches DB-Design abstrahiert von physischem DB-Design
 - Datenunabhängigkeit
 - Anfragen: Prozedural vs. deklarativ
- Informationsintegration „abstrahiert“ vom logischen DB Design vieler Datenbanken
 - Quellenunabhängigkeit
 - Ortsunabhängigkeit
 - Datenmodellunabhängigkeit
 - Formatunabhängigkeit
 - Unabhängigkeit von semantischen Unterschieden
 - Erscheint wie ein einheitliches Informationssystem



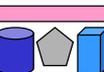
Warum ist Integration so schwer?

- System-bedingte Gründe
 - Verschiedene Plattformen
 - Anfragebearbeitung über mehrere Systeme
 - Quellen ändern sich dauernd
- Soziale Gründe
 - Finden relevanter Daten in Unternehmen
 - Menschen zur Zusammenarbeit überreden
 - Einhalten von Verabredungen und Standards
- Logik-bedingte Gründe
 - Heterogenität auf allen Ebenen
 - Semantik von Begriffen ist immer kontextabhängig
 - Semantik ist einfach schwer zu beschreiben



Integration = Ein uraltes Problem

- Seit 50 Jahren auf der Forschungsagenda
- Frühe Systeme in den 70ern
 - Hartkodierte Transformationsregeln
 - Fehleranfällig, teuer, unflexibel
- Neue Probleme
 - Viele, viele Quellen
 - Neue Arten von Daten (EXCEL, XML, GIS, OO,...)
 - Neue Arten von Anfragen (Ranking, Spatial, Mining ...)
 - Neue Arten von Nutzern (Laien, Manager, ...)
 - Neue Anforderungen (24x7x365, schnell, Ad-Hoc, Online)
 - Neue Anwendungen
 - Self-Service, eCommerce, eProcurement
 - Integration über Unternehmensgrenzen hinweg; Supply chain management
 - Strategische Unternehmensunterstützung
 - Wissensmanagement

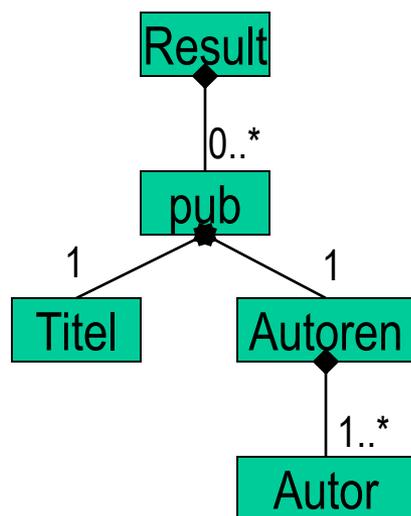


Integration am Beispiel

- Ausgangspunkt: Zwei Web-Services zur Suche nach wissenschaftlichen Publikationen mit unterschiedlichen Formaten und Operationen
- Ziel: Integrierter Web-Service, der beide Services “vereinigt”

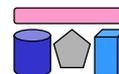
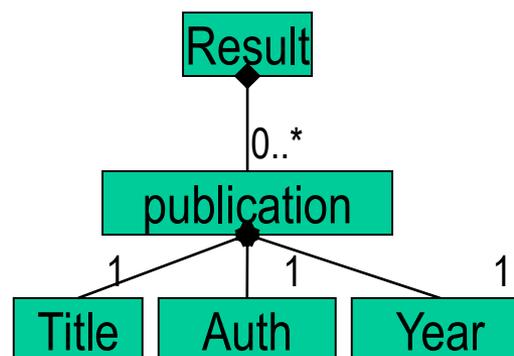
Webservice A

- Operationen
 - getPubByAuthor (firstName, lastName)
 - getPubByTitle (title)
- Output-Struktur



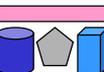
Webservice B

- Operation
 - myPubs (Autor, Jahr)
- Output-Struktur

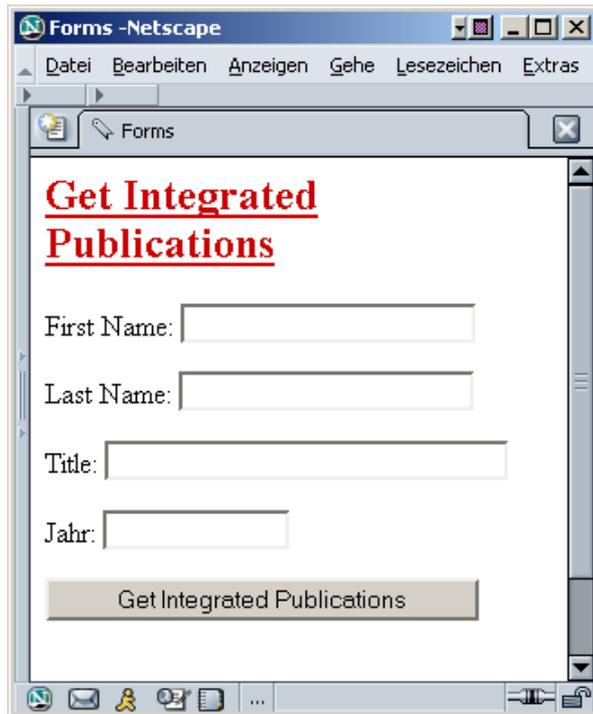


Vorgehensweise

1. Nutzerschnittstelle
2. Schema Integration / Schema Mapping
3. Anfrageumwandlung
4. Anfrageoptimierung
5. Requests an Services abschicken & Antworten einholen
6. Objektidentifikation
7. Integrationssschritte
 - Konfliktlösung etc.
 - Entscheidung kleinster gemeinsamer Nenner?
 - Durchführung (deklarativ, prozedural)
8. Anzeige beim Nutzer



1. Nutzerschnittstelle



The image shows a screenshot of a Netscape browser window. The title bar reads "Forms - Netscape". The menu bar includes "Datei", "Bearbeiten", "Anzeigen", "Gehe", "Lesezeichen", and "Extras". The address bar shows "Forms". The main content area displays the following form:

Get Integrated Publications

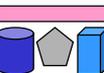
First Name:

Last Name:

Title:

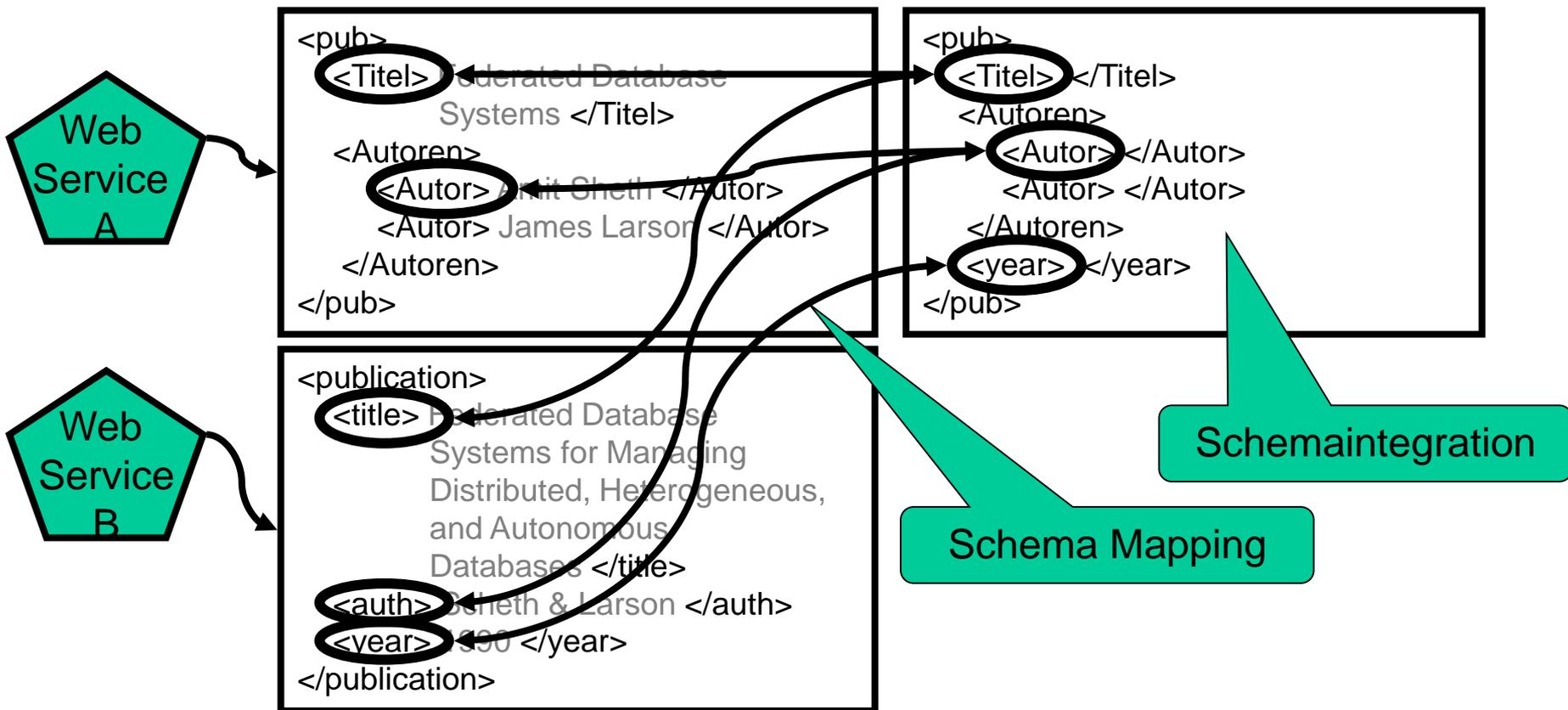
Jahr:

The browser's status bar at the bottom shows various system icons and a scroll bar on the right side of the window.



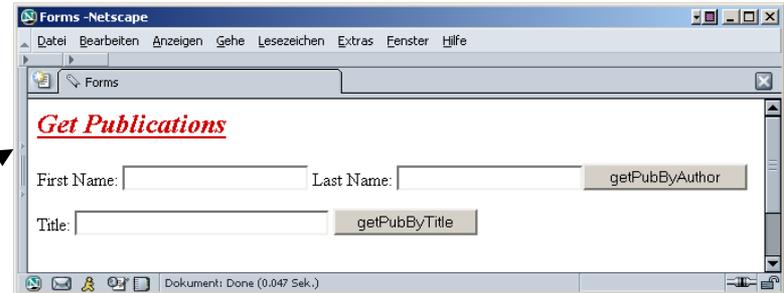
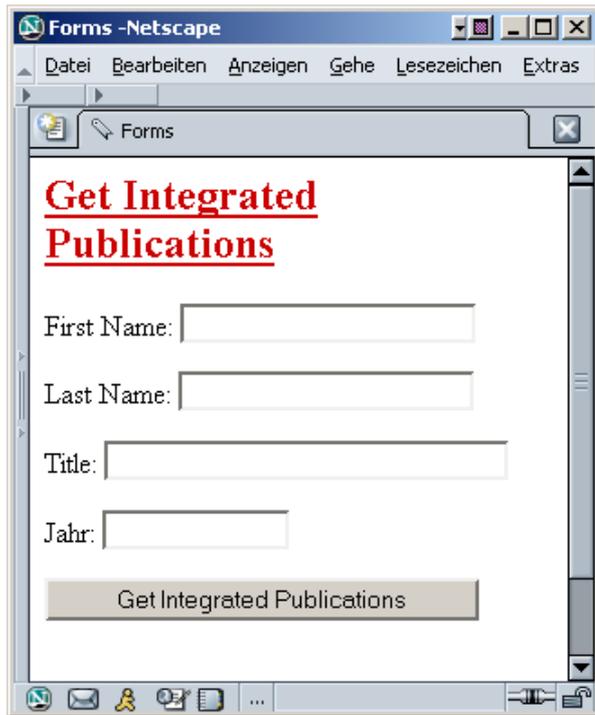
2. Schema Integration / Schema Mapping

- Erstellung eines integrierten (globalen) Schemas
 - “integrierte” Gesamtsicht auf die Daten
- Zuordnung der Elemente der Quellschemas zum integrierten Schema

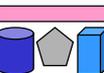


3. Anfrageumwandlung

- Integration durch Mediator
 - Nimmt Anfrage entgegen und berechnet Ergebnis unter Zugriff auf Quellen

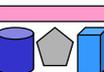


Autor =
concat(firstName, lastName)



4. Anfrageoptimierung

- Eine schnelle Antwort oder eine vollständige Antwort?
- Geschwindigkeit
 - Web Service A in USA
 - Web Service B in Deutschland
 - Welches System ist schneller? Selektivität?
- Vollständigkeit
 - Web Service A hat weniger Attribute, aber mehr Objekte
 - Web Service B hat mehr Attribute, weniger Objekte, aber ist schneller
 - Eine Suche nach „year“ kann nur durch Web Service B beantwortet werden, eine Suche nach Titel nur von A
 - Web Service A hat alle Autoren, B nur einen

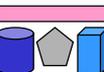


5. Antworten einholen

- Zwei Web-Service-Aufrufe ... zwei Ergebnisse

```
<Result>
  <pub>
    <Titel>MOMA - A Mapping-based Object Matching System</Titel>
    <Autoren>
      <Autor>Andreas Thor</Autor>
      <Autor>Erhard Rahm</Autor>
    </Autoren>
  </pub>
  <pub>
    <Titel>Data Cleaning: Problems and
      Current Approaches</Titel>
    <Autoren>
      <Autor>Erhard Rahm</Autor>
      <Autor>Hong-Hai Do</Autor>
    </Autoren>
  </pub>
</Result>
```

```
<Result>
  <publication>
    <Title>A Mapping-based Object Matching System</Title>
    <Auth>Thor, A.; Rahm, E.</Auth>
    <Year>2007</Year>
  </publication>
  <publication>
    <Title>Citation Analysis of Database Publications</Title>
    <Auth>Rahm, E.; Thor, A.</Auth>
    <Year>2005</Year>
  </publication>
</Result>
```



6. Objektidentifikation

- Referenzieren zwei Datensätze die gleiche Publikation?
 - Keine eindeutige Id → (generische) String-Vergleiche → hinreichend ähnlich?

```
<pub>
  <Titel>MOMA - A Mapping-based Object Matching System</Titel>
  <Autoren>
    <Autor>Andreas Thor</Autor>
    <Autor>Erhard Rahm</Autor>
  </Autoren>
</pub>
```

```
<publication>
  <Title>A Mapping-based Object Matching System</Title>
  <Auth>Thor, A.; Rahm, E.</Auth>
  <Year>2007</Year>
</publication>
```

Edit-Distance = 7
Ähnlichkeit = 84%

Ähnlichkeitsmaß?



7. Integrationsschritte

- Während der Integration
 - Konfliktlösung (welche Werte)
 - Informationsfusion
 - Restrukturierung
 - ...

8. Anzeige beim Nutzer

- Visualisierung der
 - Datenherkunft
 - Qualität
 - veränderte Daten
 - Operationen
 - ...

```
<Result>
  <pub>
    <Titel>MOMA - A Mapping-based Object Matching
      System</Titel>
    <Autoren>
      <Autor>Andreas Thor</Autor>
      <Autor>Erhard Rahm</Autor>
    </Autoren>
    <Year>2007</Year>
  </pub>
  <pub>
    <Titel>Data Cleaning: Problems and Current
      Approaches</Titel>
    <Autoren>
      <Autor>Erhard Rahm</Autor>
      <Autor>Hong-Hai Do</Autor>
    </Autoren>
  </pub>
  <pub>
    <Titel>Citation Analysis of Database Publications</Titel>
    <Autoren>
      <Autor>Rahm, E.</Autor>
      <Autor>Thor, A.</Autor>
    </Autoren>
    <Year>2005</Year>
  </pub>
</Result>
```

Konfliktlösung

Informationsfusion

Neustrukturierung

Zusammenfassung

- Begriffsdefinition
- Anwendungsgebiete zeigen Bedeutung von Integration
 - Gründe, warum Integration nötig und schwierig ist → Kap. 2
- Unterschiedliche Informationssysteme führen zu unterschiedlichen Anforderungen und Arten integrierter Informationssysteme
 - Anforderungen / Kriterien / Eigenschaften → Kap. 3
 - Architekturen von Integrationssystemen → Kap. 4
- Integration am Beispiel zeigt Notwendigkeit von ...
 - Anfrageverarbeitung → Kap. 5
 - Schemamanagement → Kap. 6
 - Datenfusion → Kap. 7

