Hasso Plattner Institut

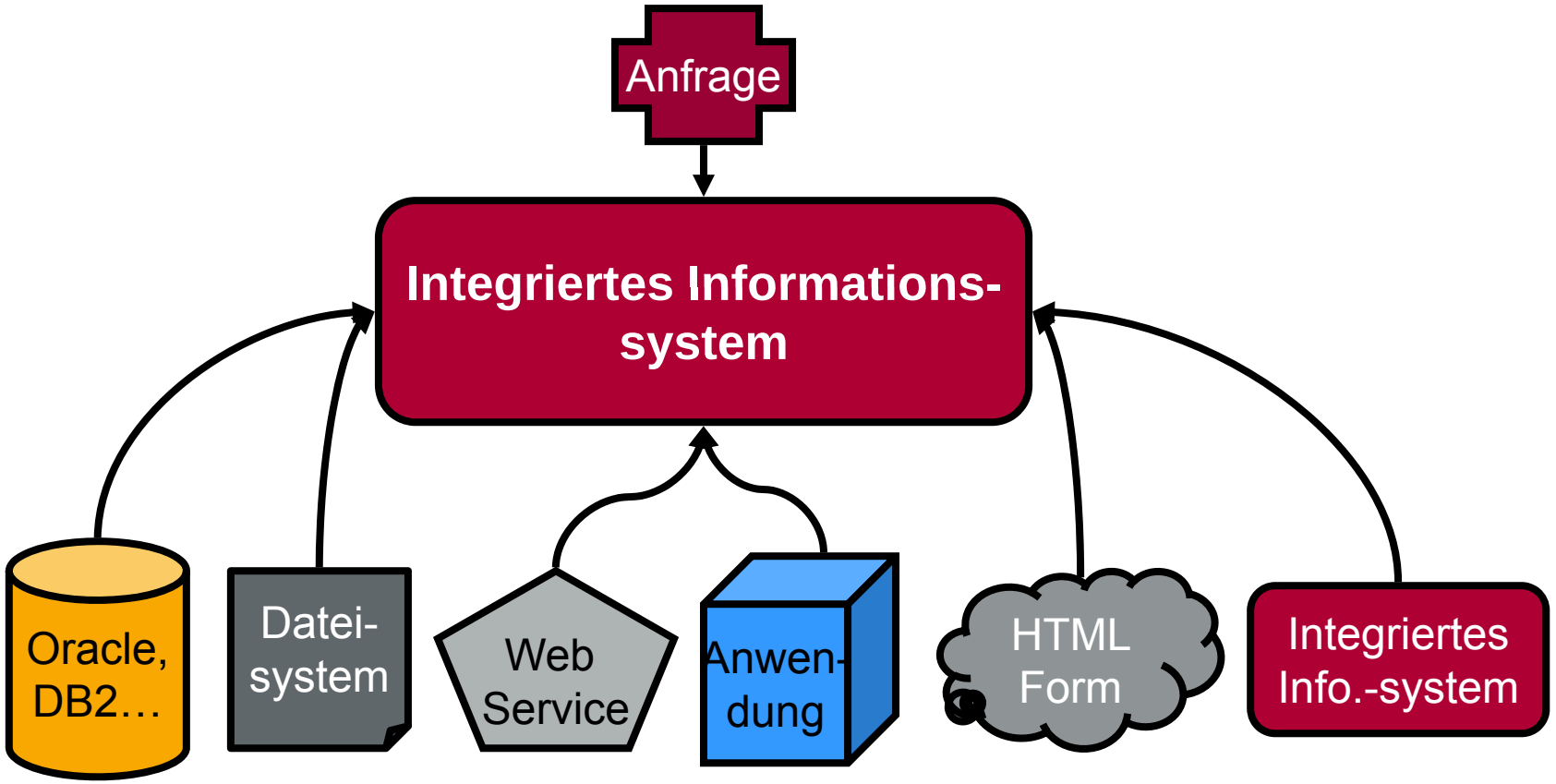
IT Systems Engineering | Universität Potsdam

Informationsintegration Einführung

14.4.2008

Felix Naumann

Integrierte Informationssysteme



Überblick

- Vorstellung der Arbeitsgruppe
- Organisatorisches
- Informationssysteme
- Informationsintegration am Beispiel
- Ausblick auf das Semester

Arbeitsgruppe Informationssysteme

Paul **Führing**

project **ViQTOR**

Patricia **Hobro**

Prof. Felix **Naumann**

Jens **Bleiholder**

project **fusem**

Karsten **Draba**

DQ Assessment

Data Fusion

Information Integration

Information Quality

project **HumMer**

Duplicate Detection

Melanie **Weis** & Sascha **Szott**

Data Cleaning

Armin **Roth**

project **System P**

Peer Data Management Systems

Matching

Service-Oriented Systems

Ontologies

Mohammed **AbuJarour**

Frank **Kaufer**

Data Integration for Life Science Data Sources

project **Aladin**

Jana **Bauckmann**

Alexander **Albrecht**

project **XClean**

Personal Information Management

Data Profiling for Schema Management

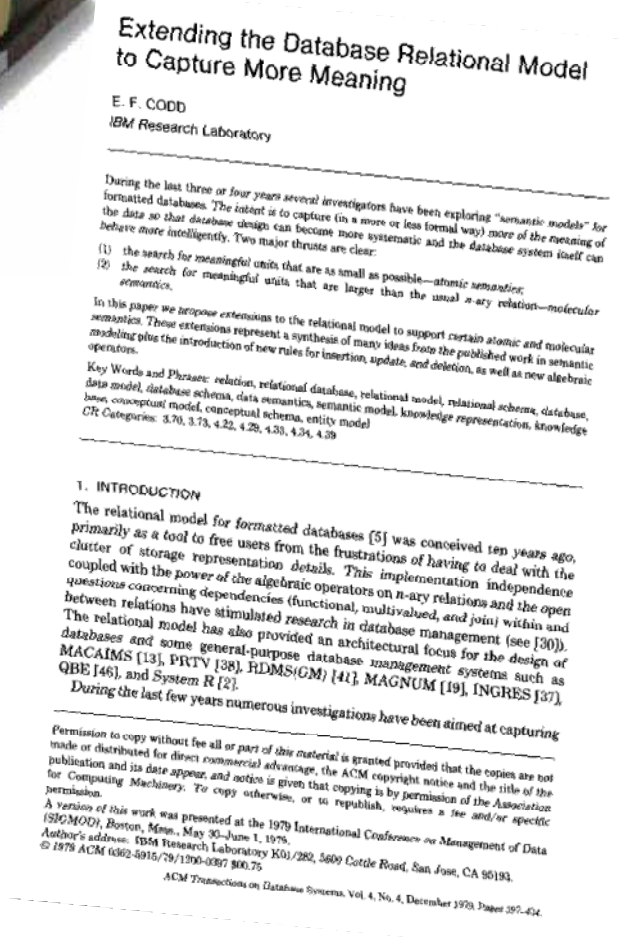
Lehrveranstaltungen in diesem Semester

Vorlesungen

- DBS II
- Informationsintegrati

Seminare

- Bachelor: Beauty is our Business
- Bachelor: www.ligageschichte.de
- Master: Duplikaterkennung
- Forschungsseminar



Extending the Database Relational Model to Capture More Meaning

E. F. CODD

IBM Research Laboratory

During the last three or four years several investigators have been exploring "semantic models" for formatted databases. The intent is to capture (in a more or less formal way) more of the meaning of the data so that database design can become more systematic and the database system itself can behave more intelligently. Two major thrusts are clear:

(1) the search for meaningful units that are as small as possible—atomic semantics;
(2) the search for meaningful units that are larger than the usual n-ary relation—molecular semantics.

In this paper we propose extensions to the relational model to support certain atomic and molecular semantics. These extensions represent a synthesis of many ideas from the published work in semantic modeling plus the introduction of new rules for insertion, update, and deletion, as well as new algebraic operators.

Key Words and Phrases: relation, relational database, relational model, relational schema, database, data model, database schema, data semantics, semantic model, knowledge representation, knowledge base, conceptual model, conceptual schema, entity model

CR Categories: 3.70, 3.73, 4.22, 4.29, 4.33, 4.34, 4.39

1. INTRODUCTION

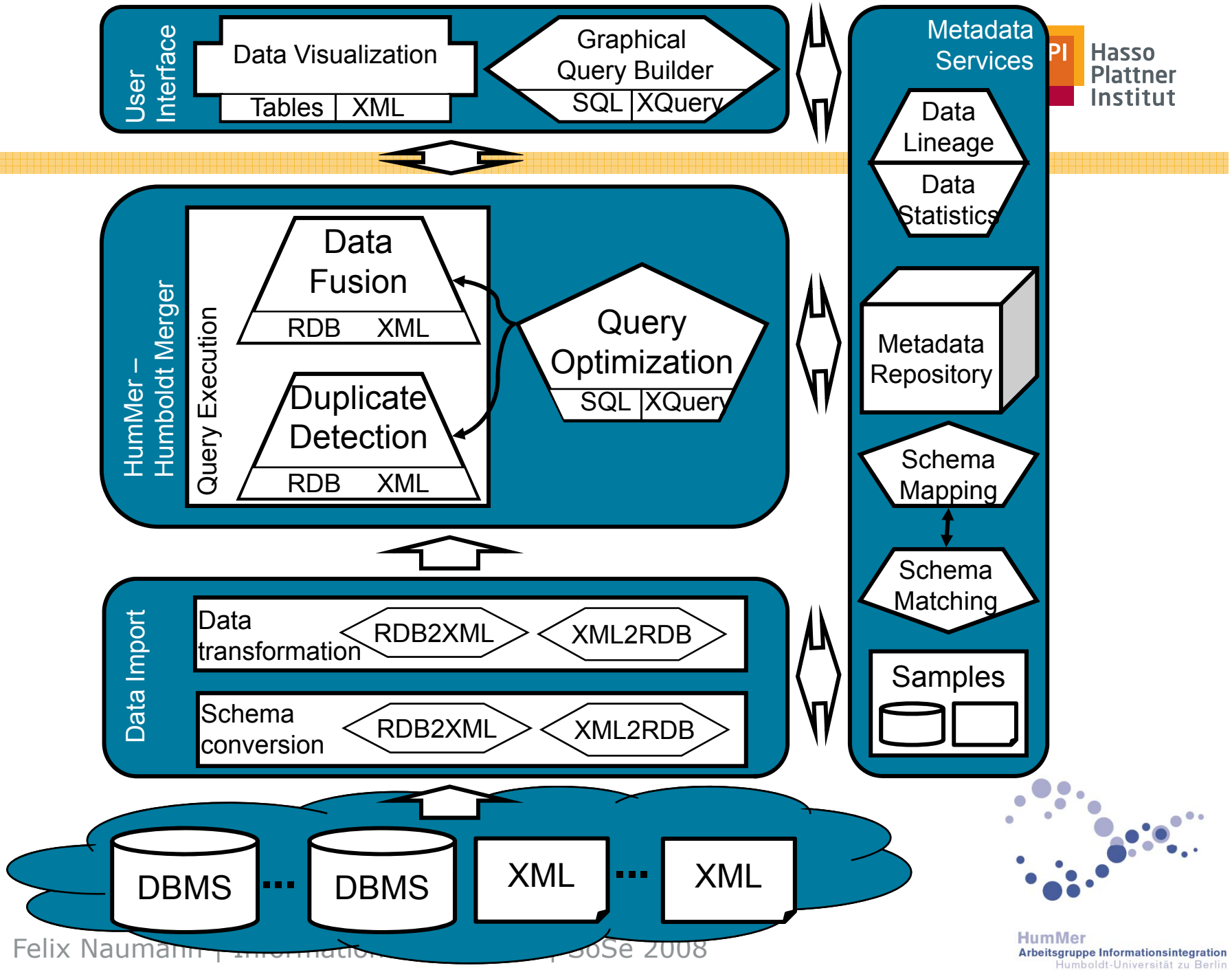
The relational model for formatted databases [5] was conceived ten years ago, primarily as a tool to free users from the frustrations of having to deal with the clutter of storage representation details. This implementation independence coupled with the power of the algebraic operators on n-ary relations and the open questions concerning dependencies (functional, multivalued, and join) within and between relations have stimulated research in database management (see [30]). The relational model has also provided an architectural focus for the design of databases and some general-purpose database management systems such as MACAIMS [13], PRTV [38], RDMS(GM) [41], MAGNUM [19], INGRES [37], QBE [46], and System R [2].

During the last few years numerous investigations have been aimed at capturing

A version of this work was presented at the 1979 International Conference on Management of Data (SIGMOD), Boston, Mass., May 30–June 1, 1979.

Author's address: IBM Research Laboratory K01/282, 5600 Cottle Road, San Jose, CA 95193.

ACM Transactions on Database Systems, Vol. 4, No. 4, December 1979, Pages 397–434.



Überblick

- Vorstellung der Arbeitsgruppe
- Organisatorisches
- Informationssysteme
- Informationsintegration am Beispiel
- Ausblick auf das Semester

Termine und Leistungserfassung

- Vorlesung
  - Montags 15:15 – 16:45
  - Donnerstags 13:30 – 15:00
- Praktikum / Übung
  - Ausgewählte Termine – WWW beachten
- Erste Vorlesung
  - 14.4.2008
- Letzte Vorlesung
  - 17.7.2008
- Feiertage
  - 1.5. Maifeiertag (Do)
  - 12.5. Pfingstmontag
- Sondertermine
  - 17.4. (Do): IBM Tag
  - 21.4.: Erste Übung
  - 24.4.: fällt leider aus ☹

- Prüfung
  - Mündlich, 30 Minuten
  - Erste Woche nach Vorlesungszeitraum
- 7 Übungstermine
  - Leitung: Alexander Albrecht
  - Theoretische und praktische Aufgaben
  - 2er Teams
- Voraussetzungen
  - Zur Teilnahme
    - Datenbankkenntnisse (DBS I)
  - Zur Prüfung
    - Besuch der Vorlesung
    - Aktive Teilnahme an den Praktikumsterminen
    - „Bestehen" der Übungsblätter

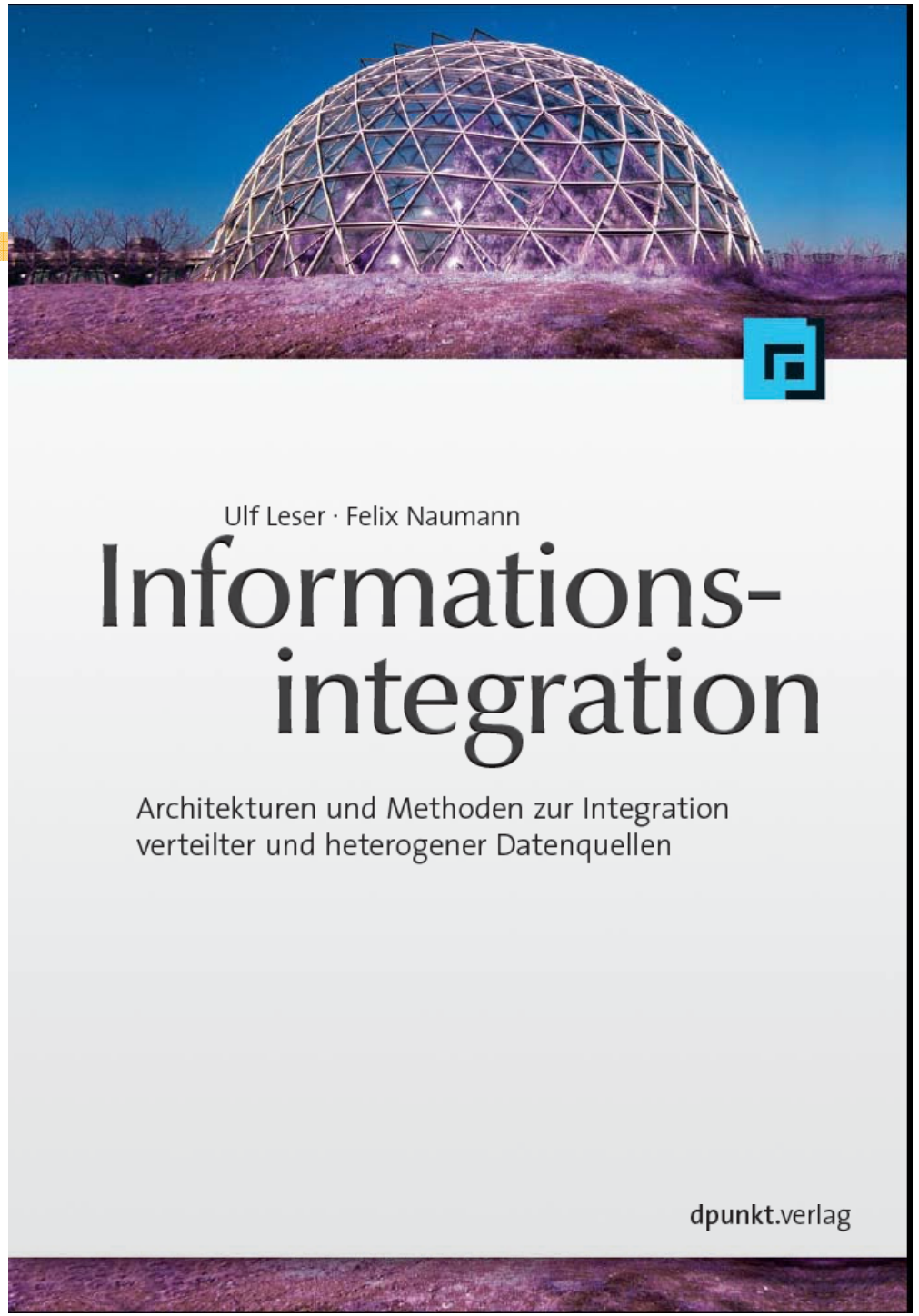
Feedback

- Evaluation am Ende des Semesters
- Fragen bitte jederzeit!
  - In der VL
  - Sprechstunde: Dienstags 15-16
  - Email: naumann@hpi.uni-potsdam.de
- Anregungen zur Verbesserung:
  - Z.B. zu
    - Gebrauch der Folien
    - Infos im WWW
  - Jeweils nach der VL oder in der Sprechstunde
  - Oder per Email: naumann@hpi.uni-potsdam.de

Lehrbuch

- Informationsintegration
- Ulf Leser und Felix Naumann
  - dpunkt Verlag
- 42 Euro
- X-mal in Bibliothek

- Fehler gefunden =>



Weitere Literatur

- Themen u.a. aus
  - Föderierte Datenbanksysteme. Konzepte der Datenintegration, Stefan Conrad, ISBN: 3540631763
  - Principles of Distributed Database Systems
    M. Tamer Özsu, Patrick Valduriez
    ISBN: 0136597076
- Jeweils Literaturhinweise in den Vorlesungen
- Alle genannten Artikel können von mir per Email angefragt werden. Oder:
  - Google Scholar: http://scholar.google.com/
  - DBLP: http://www.informatik.uni-trier.de/~ley/db/index.html
  - CiteSeer: http://citeseer.ist.psu.edu/
  - ACM Digital Library: www.acm.org/dl/
  - Homepages der Autoren

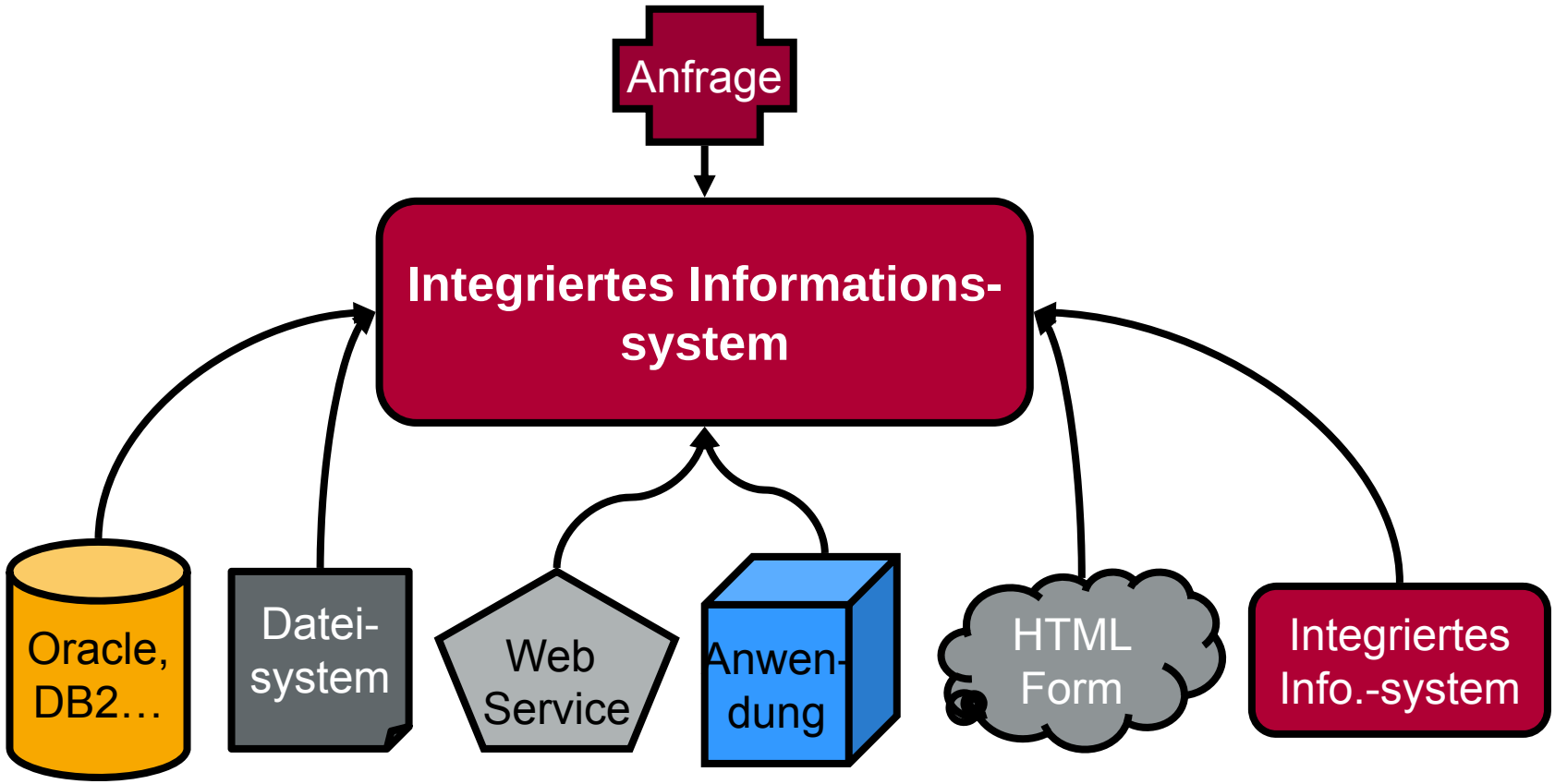
Vorstellung – Hörer

- Welches Semester?
- HPI oder IfI?
- Erasmus o.ä.?
- Datenbankkenntnisse?
  - Andere relevante Lehrveranstaltungen?
- Workshop/Seminar Duplikaterkennung?
- Ihre Motivation?
  - Schon mal integriert?
  - DWH?

Überblick

- Vorstellung der Arbeitsgruppe
- Organisatorisches
- Informationssysteme
- Informationsintegration am Beispiel
- Ausblick auf das Semester

Integrierte Informationssysteme



Was ist Informationsintegration?

Informationsintegration ist die Zusammenführung von Daten und Inhalt verschiedener Quellen zu einer einheitlichen Informationsmenge.

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.

Wo herrscht Informationsintegration?

- Im weiteren Sinne
  - Business-Integration
  - Application-Integration
  - Prozess-Integration (Workflow-Integration)
- Im engeren Sinne
  - Datenbanken und Informationssysteme
    - Verteilt
    - Autonom
    - Heterogen

Beispiele für Informationssysteme

- Dateisystem
  - WWW (HTML Dateien)
  - Desktop-Anwendungen (Textverarbeitung, etc.)
- Datei (Flat file)
  - Zeile /Token
  - Anfrage: Parser, File search, RegEx
  - Strukturiert / semi-strukturiert / unstrukturiert
  - Spezialfall: Markup-Datei
    - XML, HTML mit Struktur
    - Anfragespreache
  - Beispiele
    - Komma-delimited files
    - Annotated files
  - Einsatzgebiete
    - SwissProt

```
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="MEKPRGVRCTNGFPERELPRPGASRPAEKSRPPEAKGAQPADAWK
FT                   AGRPRSEEDNELNLPNLAAAYSSILRSLGEDPQRQGLLKTPWRAATAMQFFTKGYQETI
FT                   SDVLNDAIFDEDHDEMVIVKDIDMFSMCEHHLVPFVGRVHIGYLPNKQVLGLSKLARIV
FT                   EIYSRRLQVQERLTKQIAVAITEALQPAGVGVVIEATHMCMVMRGVQKMNSKTVTSTML
FT                   GVFREDPKTREEFLTLIRS"
SQ   Sequence 1016 BP; 236 A; 279 C; 291 G; 210 T; 0 other;
     gacttcgaac ctcattcggt gcagaactcc tgtcccggtg acagccacag gtcacggccg        60
     ccggctaagc cgagccgcag cgcttggtag caccttaggg tgtctcggga gcaatcgcgc       120
     cgggtccatg gagaagccgc ggggtgtaag gtgcaccaat gggttccccg agcgggagct       180
        ...
     catcaggagc tgaacttccg tgtgcgagcc ccggtttgca gacccccgct gaggccagcg       900
     ttatctgtct cgattgtaca ttccagttcc agttggtata cttgtcaact ttatttctca       960
     ccatgaattg tatttaataa ttatttatag agatgtcaaa taaaggtgat caactt          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)

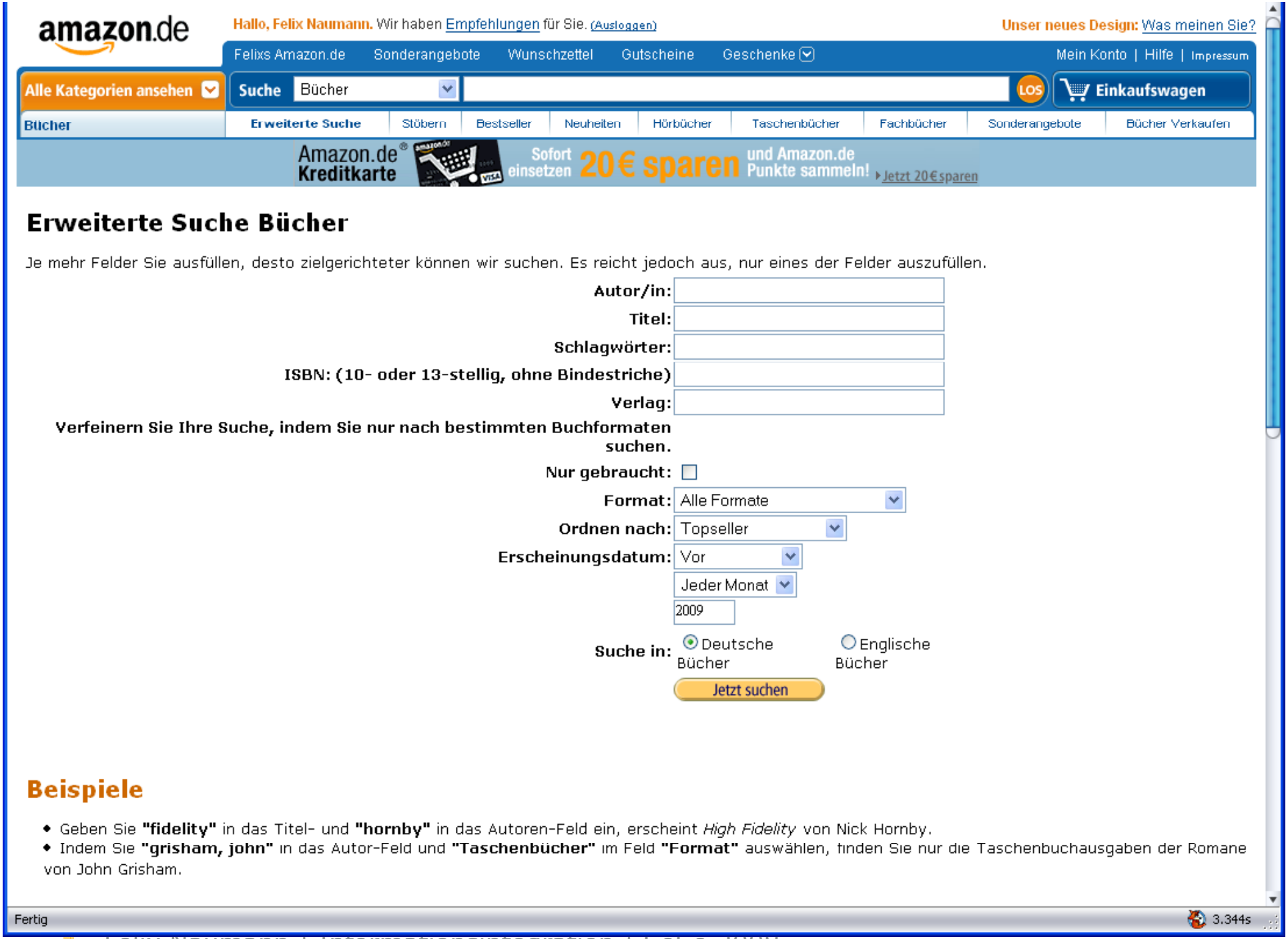
Link to functional annotation of resulting protein

Translated protein sequence

Sequence of bases

Beispiele für Informationssysteme

- Datenbank
  - Anfragesprachen, z.B. SQL
  - Relational, OO, Hierarchisch
  - XML DBMS
  - Data Warehouses
  - OLTP
- HTML Formular
  - Anfrage: Einfache Suche, komplexe Formulare
    - Radiobutton, dropdown-list, etc.
  - Struktur des Ergebnisses: wie Markup Datei: Flach, hierarchisch oder graph-basiert


amazon.de
Hallo, Felix Naumann. Wir haben Empfehlungen für Sie. (Ausloggen)
Unser neues Design: Was meinen Sie?
Felixs Amazon.de
Sonderangebote
Wunschzettel
Gutscheine
Geschenke
Mein Konto | Hilfe | Impressum
Alle Kategorien ansehen
Suche
Bücher
LOS
Einkaufswagen
Bücher
Erweiterte Suche
Stöbern
Bestseller
Neuheiten
Hörbücher
Taschenbücher
Fachbücher
Sonderangebote
Bücher Verkaufen
Amazon.de Kreditkarte
Sofort einsetzen
20 € sparen
und Amazon.de Punkte sammeln!
Jetzt 20 € sparen
Erweiterte Suche Bücher
Je mehr Felder Sie ausfüllen, desto zielgerichteter können wir suchen. Es reicht jedoch aus, nur eines der Felder auszufüllen.
Autor/in:
Titel:
Schlagwörter:
ISBN: (10- oder 13-stellig, ohne Bindestriche)
Verlag:
Verfeinern Sie Ihre Suche, indem Sie nur nach bestimmten Buchformaten suchen.
Nur gebraucht:
Format:
Alle Formate
Ordnen nach:
Topseller
Erscheinungsdatum:
Vor
Jeder Monat
2009
Suche in:
Deutsche Bücher
Englische Bücher
Jetzt suchen
Beispiele
Geben Sie "fidelity" in das Titel- und "hornby" in das Autoren-Feld ein, erscheint High Fidelity von Nick Hornby.
Indem Sie "grisham, john" in das Autor-Feld und "Taschenbücher" im Feld "Format" auswählen, finden Sie nur die Taschenbuchausgaben der Romane von John Grisham.
Fertig
3.344s

Beispiele für Informationssysteme

- Services
  - Z.B. Web Services
    - XML Dokument
    - Anfrage als XML Dokument
  - Einsatzgebiete
    - Intra-organisatorische Workflows
    - E-Marketplaces
    - Datenaustausch
    - Mashups


www.xmethods.net
Recent Listings [View the FULL LIST]
Publisher | Style | | Service Name | Description | Implementation
StrikeIron | DOC | Try It | Gale Group Web Domain Business Intelligence | Based on a website domain get in-depth financial and corporate information
StrikeIron | DOC | Try It | Gale Group Business Intelligence | Based on company name get in-depth financial and corporate information
StrikeIron | DOC | Try It | Gale Group Business Information | Standard financial and corporate information for 440,000 U.S. and international companies
StrikeIron | DOC | Try It | IPligence Geo IP Location | Comprehensive information on IP addresses
mharvanek | DOC | Try It | DuoShare Address Quality Integrator | DuoShare Address Quality Integrator standardizes, corrects, validates (to the delivery point), and enhances U.S. addresses using a number of USPS® certified processes.
CDYNE | DOC | Try It | CDYNE Phone Notify! | Calls any phone number and speaks text or sound file to the person. Also supports advanced Dialplans, Machine Detection and Incoming calls.
CDYNE | DOC | Try It | CDYNE SMS Notify! | Send SMS Text Messages to your mobile telephone devices without the use developing your own software!
StrikeIron | DOC | Try It | The Sports Network Men's College Basketball | Comprehensive sports data for NCAA Mens Division I Basketball
StrikeIron | DOC | Try It | Wall Street Horizon Interest Rate Calendar | Access current and historical interest rate information
EDIXMLOnline | DOC | Try It | EDIXMLOnline | Translate, generate & validate EDI files using X12/EDIFACT SEF files. Translate EDI to XML format. | MS .NET
StrikeIron | DOC | Try It | Wall Street Horizon Real-Time Company Earnings | Information to analyze and evaluate investments and future earning potential
phi1281 | RPC | Try It | Weight Watchers Point Calculator | Given the parameters of calories, fat grams, & fiber grams, computes the Weight Watchers point value for a given serving of food. | ColdFusion
SOATrader | RPC | Try It | Captcha Web Service | This Web service will create captcha
Fertig
6.828s

StrikeIron: Data as a Service

STRIKEIRON®
DATA AS A SERVICE

CONTACT REGISTER LOGIN

MARKETPLACE SOLUTIONS PRODUCTS DEVELOPERS PARTNERS NEWS COMPANY SUPPORT YOUR ACCOUNT

- Business Users
- Enterprise IT Professionals
- Application Developers
- Partners

STRIKEIRON: StrikeIron's Data Services give you access to live data to use now, integrate into applications or build into Web sites.

SEARCH WEB SERVICES

Search

All Categories

Browse

LATEST NEWS

StrikeIron Offers Gale Data on Salesforce.com's AppExchange

StrikeIron Hits Home Run with Major League Baseball Data from The Sports Network

StrikeIron Data Fuels Enterprise Mashup Platforms

intelligent enterprise EDITORS' CHOICE AWARD 2008 COMPANIES TO WATCH

WHAT DO YOU WANT TO DO?

CONNECT

- Find Data Services
- Browse Solutions
- Cleanse & Enhance Data

CREATE

- Build Mashups in Excel
- Create Business Solutions
- Integrate into Salesforce.com

COLLABORATE

- New! Developer Community
- Download Code and Clients
- Access Free Data

DATA SPOTLIGHT

NEW WEB SERVICES

- D&B Business Prospect
- Gale Business Intelligence
- Midnight Trader Financial News
- MapQuest Driving Directions
- TSN Men's College Basketball
- Wall Street Horizon Company Earnings

HOT WEB SERVICES

- US Address Verification
- Global SMS Pro
- Sales and Use Tax
- Email Verification
- Foreign Exchange Rates

PARTNERS

StrikeIron is proud to partner with the following industry leaders:

Blockieren

jamcracker Enabling the On Demand World

 Terms of Use | Privacy | Trademarks | Contact Us | Site Map

Fertig 20.328s

726 APIs

Search APIs

2946 Mashups

Search Mashups

Vertical Markets

Social Platforms
Mapping
Video
Shopping
Government
Mobile+Telephony

first time here? take a tour

Popular Directory Searches

Celebrity Mashups
Video Mashups
Popular New Mashups
All Popular Mashups
Maps Mashups

Keeping you up to date with mashups, APIs and the Web as platform. Learn more »

Find	Track	Do	Share
API Directory	Industry News	Learn	Sign-up
Mashup Directory	Market Trends	Search Code	Members
Most Popular	Major Players	Win Contests	Add Links
Mashup Matrix	Tag Cloud	Discuss	Talk to us

Latest News Updates

more news »

Google Chart API's New Schematic Maps »

Graphic, schematic maps can very useful in a wide range of charting scenarios and thanks to a recent upgrade, now you can use the Google Chart API to create them. »

Posted April 11, 2008. Continue reading »

April 12, 2008

New Mashups »

- 2008 Beijing Olympics Torch Relay Path
- Svraka
- Enomalism Elastic Computing
- WebFOCUS for Google Maps

Mashup of the Day »

2008 Beijing Olympics Torch Relay Path

New APIs »

- ThisNext
- PilotOutlook
- Summize Twitter Search
- Enomalism
- Mortgage Marvel

Top Tags

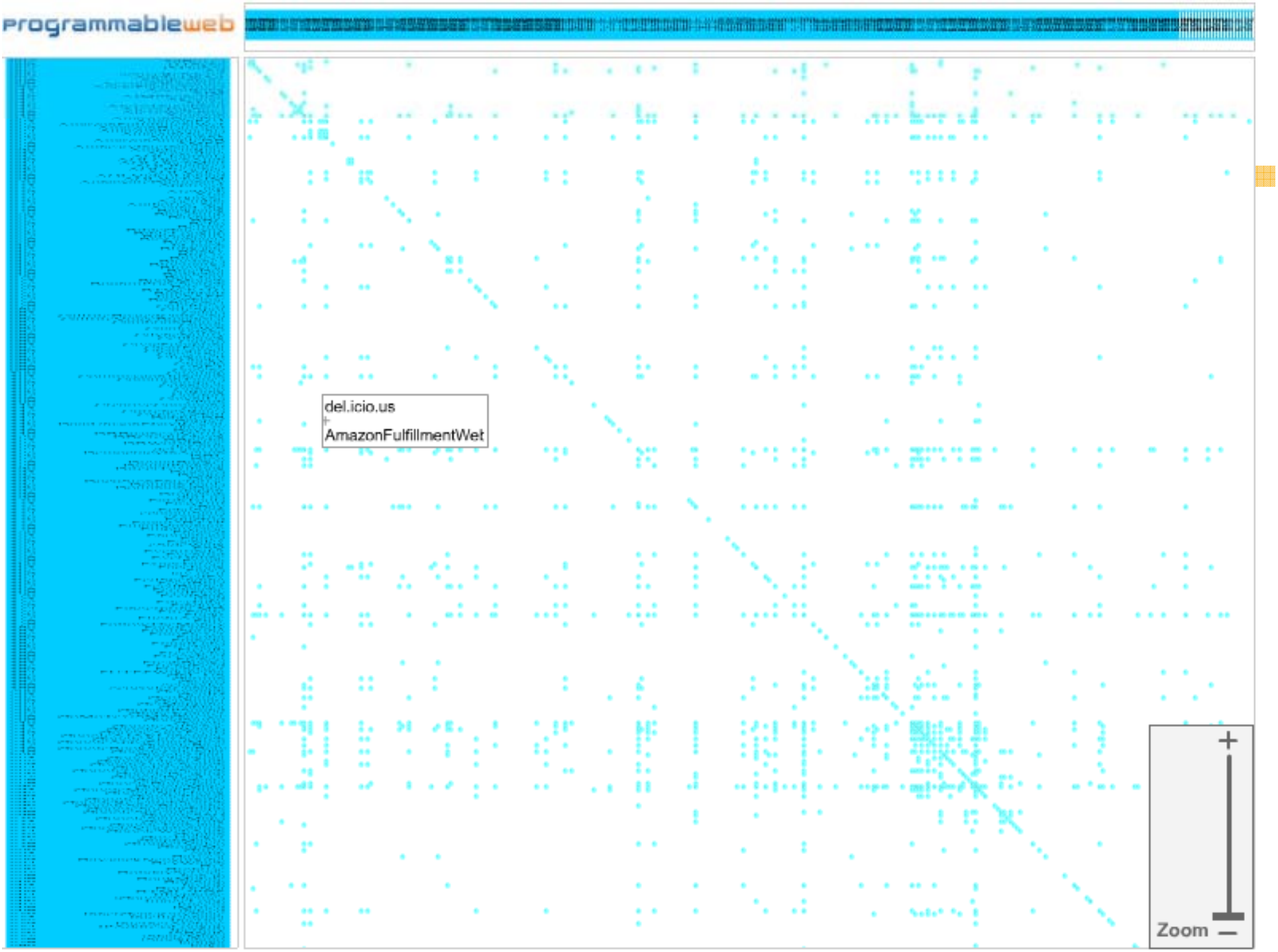
mapping (1602)
photo (438)
shopping (389)
search (348)
video (293)
travel (276)
news (188)
sports (176)

programmableweb

		001	002	003	004	005	006	007	008	009	010	011	012	013	014	015	016	017	018	019	020	021	022	023	024	025	026	027	028	029	0
001	23																														
002	30Boxes	4	1																			1					1				
003	411Sync	1	120															1		1											
004	43Things			2																1		1									
005	A9				9															3											
006	activeRenderer																														
007	AdobeOnAir					5																									
008	AdobeShare						1																								
009	AevumObscurum																														
010	AgentFactorTravel																														
011	AgentRank																														
012	AIM												6							2											
013	AIMPhoneline													1																	
014	Akismet																														
015	Alexa															5				1											
016	AlexaThumbnail																2		1								1				
017	AlexaTopSites		1															4	1												
018	AlexaWebInfo																1	1	5												
019	Amazon		1	1	3								2			1				238		5					4				
020	AmazonDevPay																														
021	AmazonEC2	1		1																5		29			1	1	10				
022	AmazonFulfillmentWeb																														
023	AmazonHistorical																														
024	AmazonPayments																					1			2	1	1				
025	AmazonQueue																					1			1	2	2				
026	AmazonS3	1															1			4		10			1	2	30				

AmazonPayments
A9

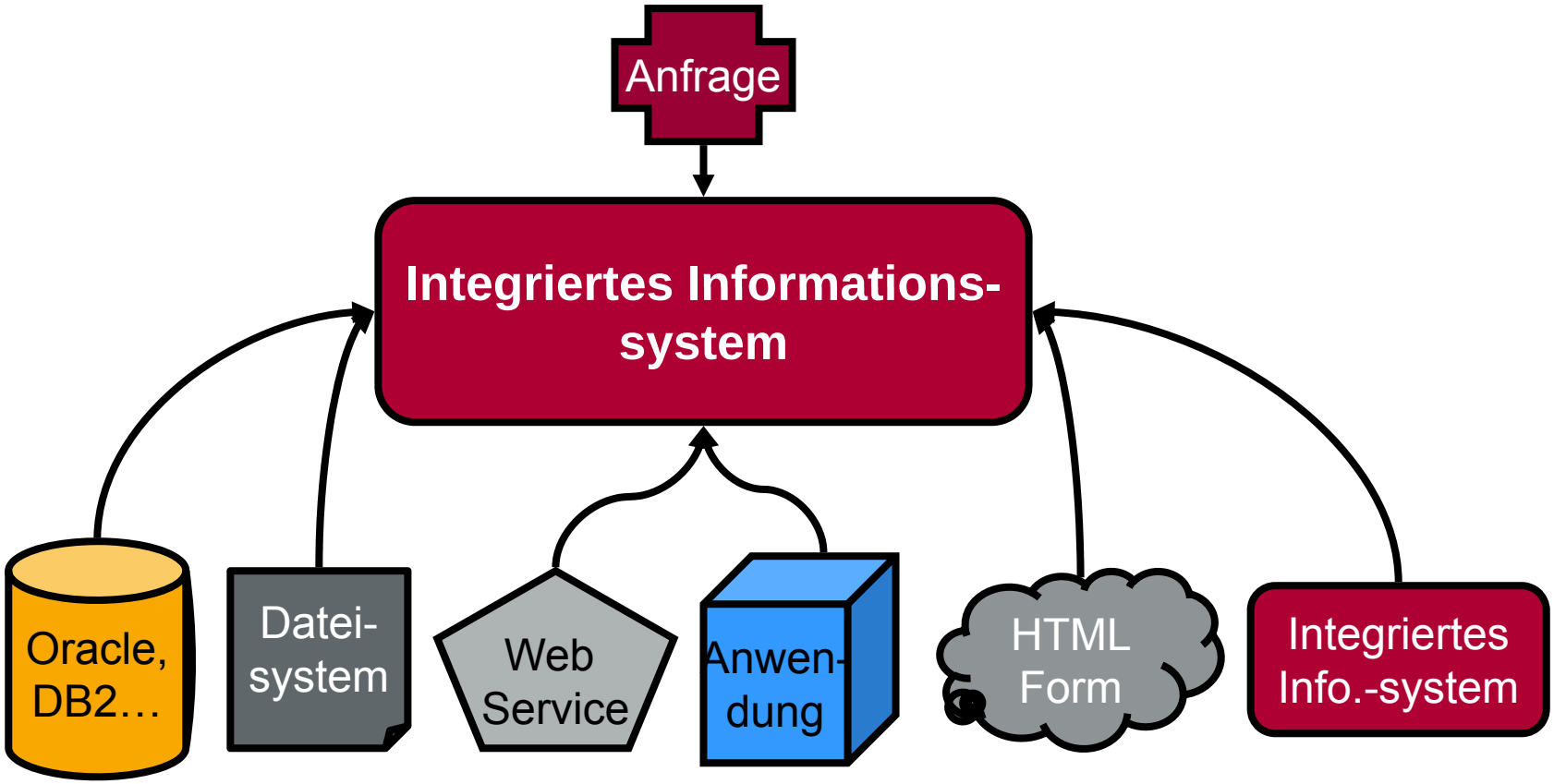
Zoom


programmableweb
del.icio.us
AmazonFulfillmentWeb
Zoom

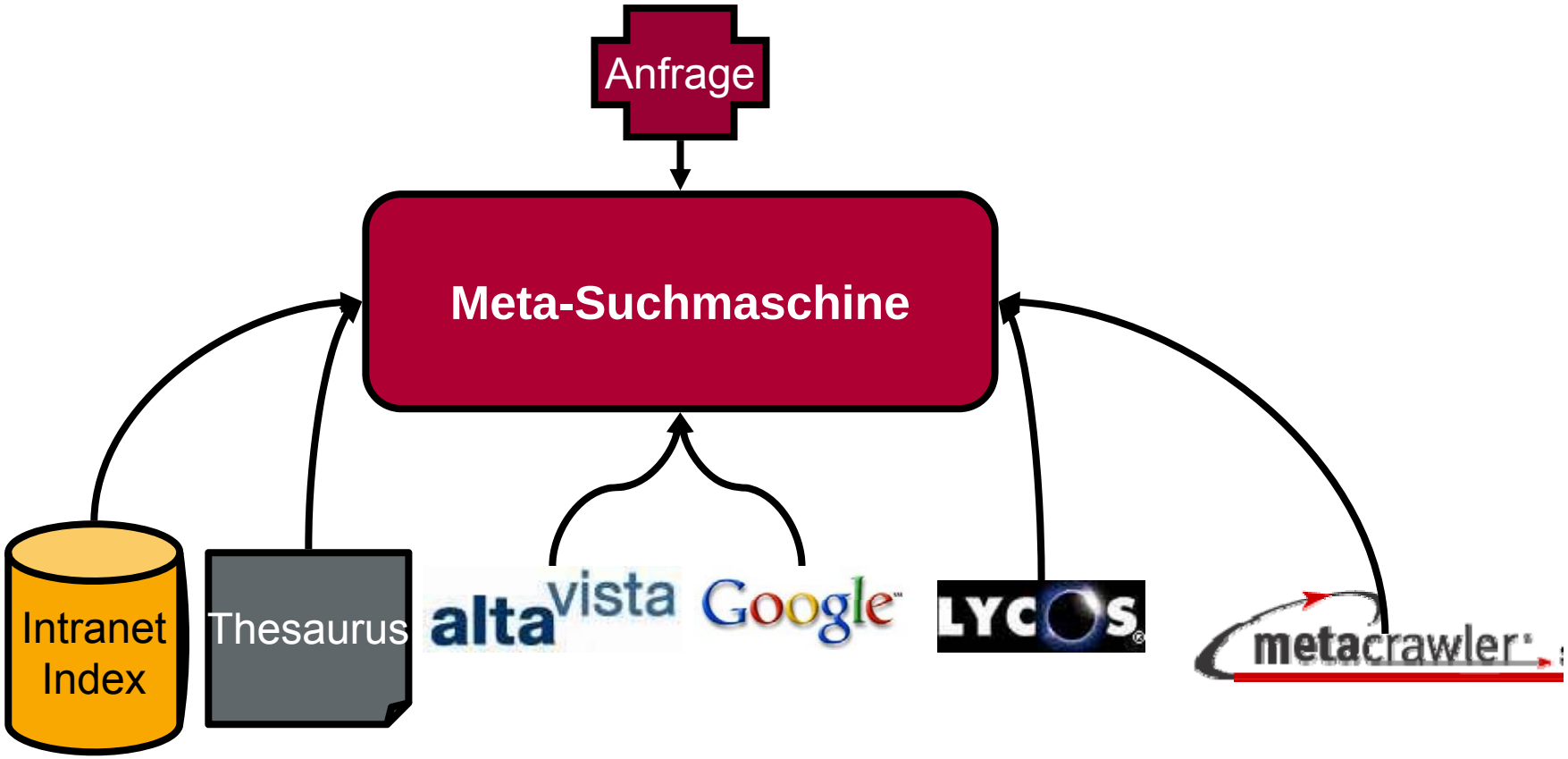
Beispiele für Informationssysteme

- Anwendung
  - Anfrage via Anwendungsschnittstelle oder GUI
  - Struktur: Objekt (Interface), Display (GUI)
  - Einsatzgebiete
    - Komplexe Analysen (Data Mining, Statistik)
- Integriertes Informationssystem
  - Verhält sich in Anfrage, Struktur und Informationseinheit je nach Design:
    - DBMS
    - HTML Formular
    - Web Service
  - Einsatzgebiete: Meta Search, Life Sciences, Int. Unternehmen, Intranets

Integrierte Informationssysteme



Integrierte Suchmaschinen

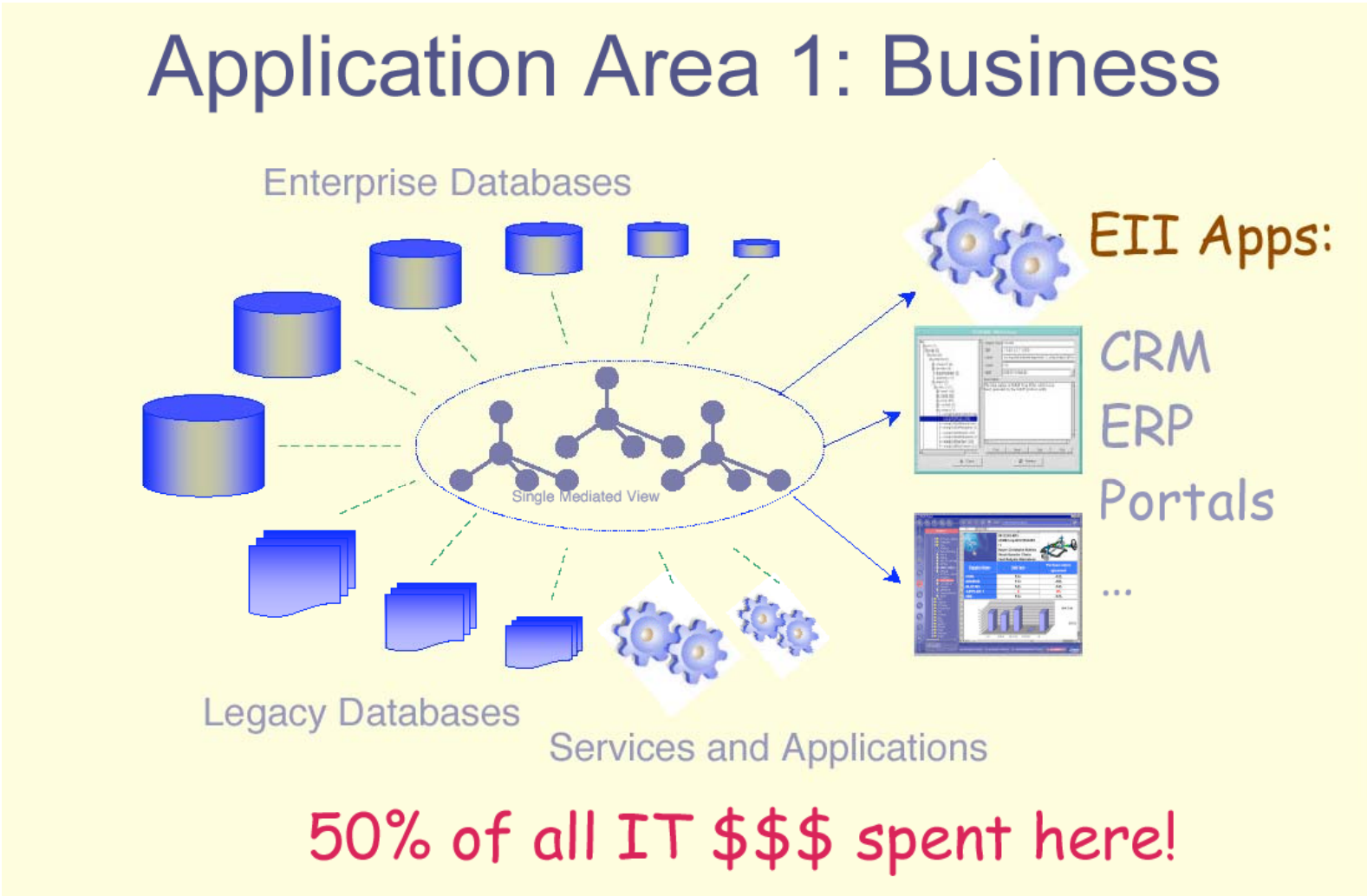


Weitere Beispiele?

Integration = Abstraktion

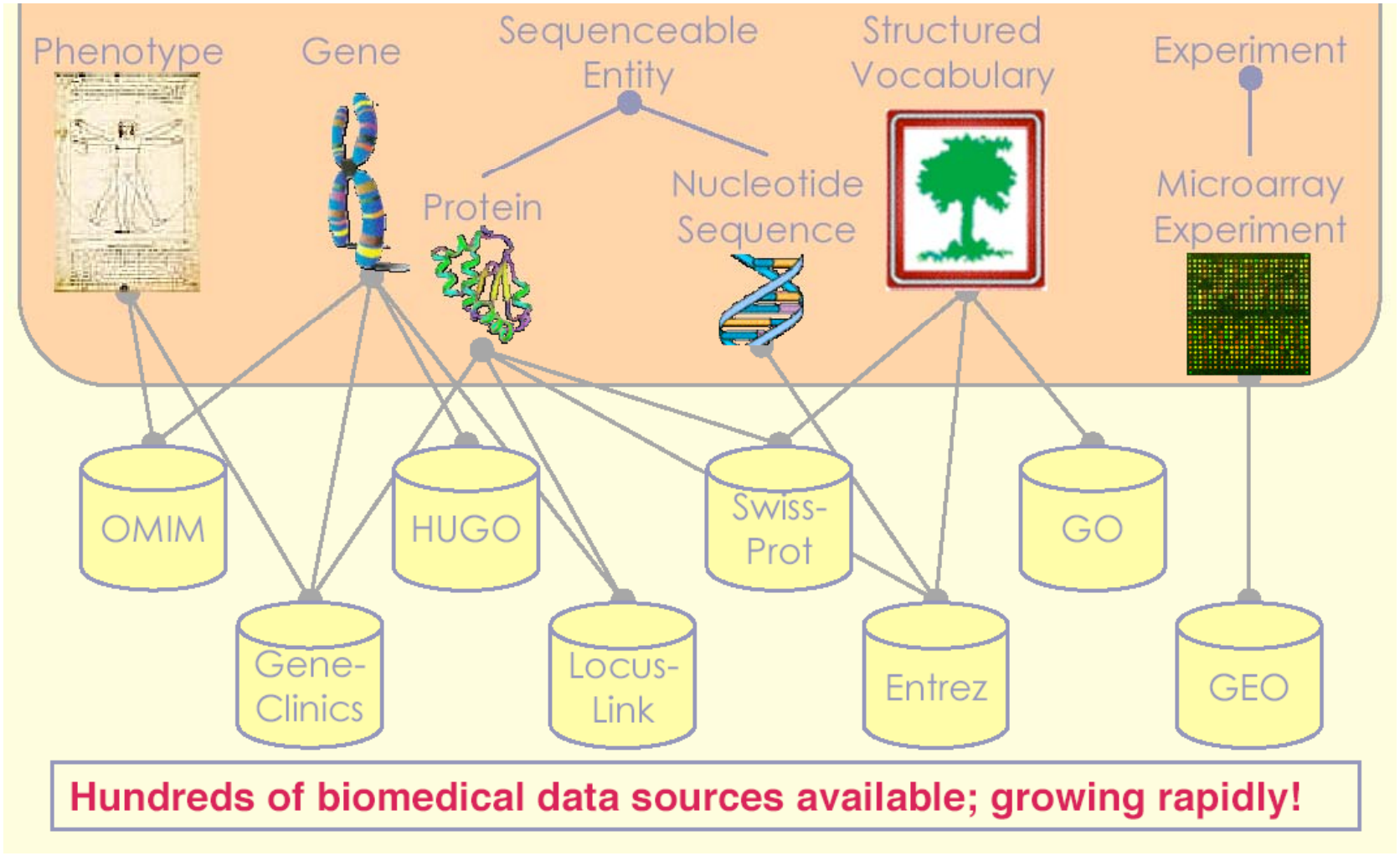
1. Logisches DB-Design abstrahiert von physischem DB-Design
   - Datenunabhängigkeit
   - Anfragen: Prozedural vs. deklarativ
2. Informationsintegration abstrahiert von logischen DB Design
   - Quellenunabhängigkeit (Speicherort)
   - Datenmodell- und Syntaxunabhängigkeit
   - Unabhängigkeit von semantischen Unterschieden (hoffentlich!)

Anwendungsgebiet 1: Business [Halevy04]



Anwendungsgebiet 2: Wissenschaft

[Halevy04]



Anwendungsgebiet 3: Das Web [Halevy04]

Informationsintegration: Ein altes Problem

- Seit 50 Jahren auf der Forschungsagenda
- Frühe Systeme in den 70ern
- Integration per Hand natürlich noch früher
- Neue Probleme
  - Viele, viele Quellen
  - Heterogenität
  - Neue Arten von Daten (XML, GIS, OO,...)
  - Neue Arten von Anfragen (Search, UDFs,...)
  - Neue Arten von Ergebnissen (Ranking, Visualisierung, ...)
  - Neue Arten von Nutzern (Laien, Manager, Admins, ...)
- Alon Halevy: „It‘s plain hard!“ [Halevy04]

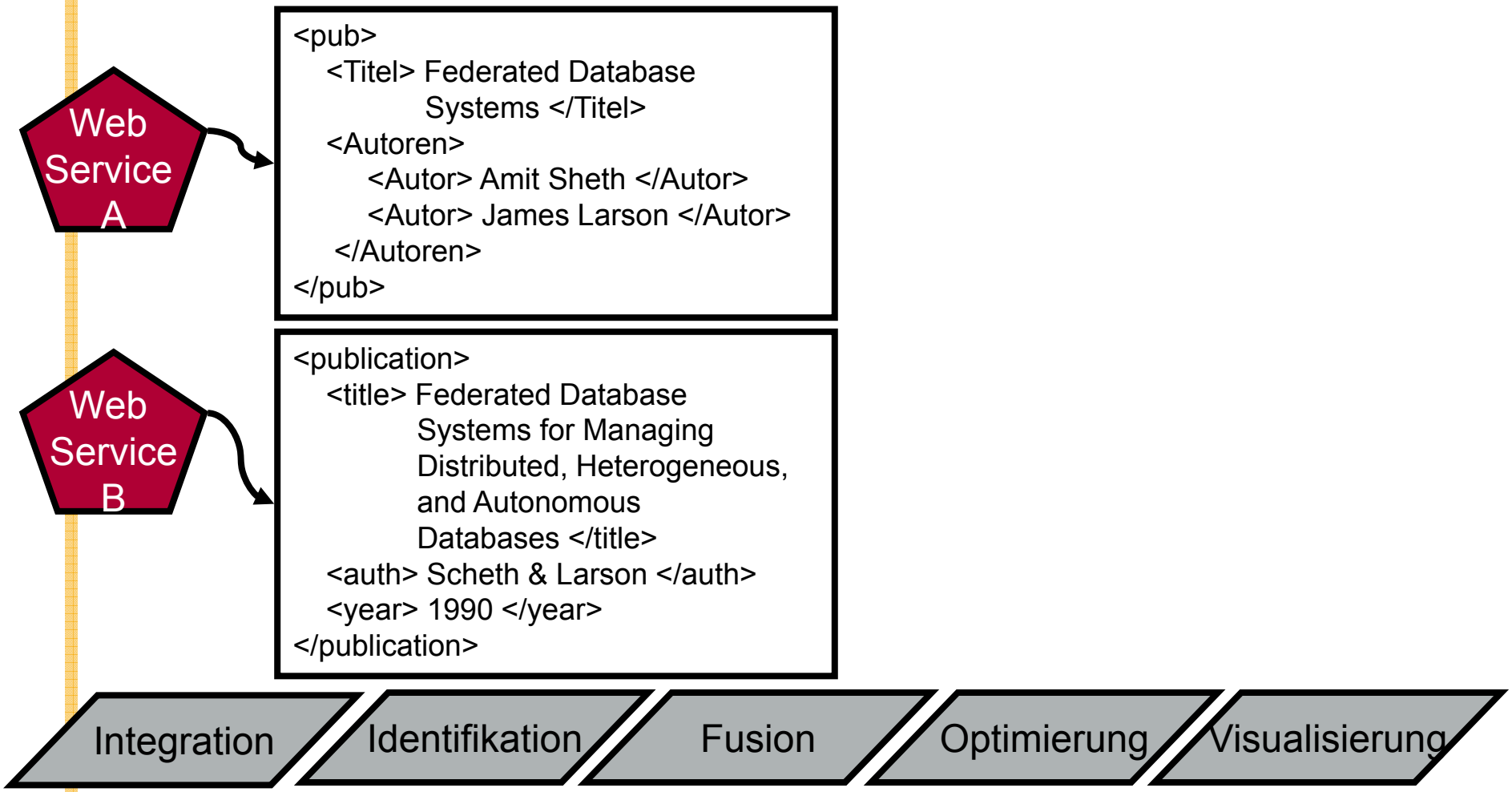
Warum ist es so schwer? [Halevy04]

- System-bedingte Gründe
  - Verschiedene Plattformen
  - Anfragebearbeitung über mehrere Systeme
- Soziale Gründe
  - Finden relevanter Daten in Unternehmen
  - Beschaffen relevanter Daten in Unternehmen
  - Menschen zur Zusammenarbeit überreden
- Logik-bedingte Gründe
  - Schema- und Datenheterogenität
  - Dies ist unabhängig von der jeweiligen Integrationsarchitektur.

Überblick

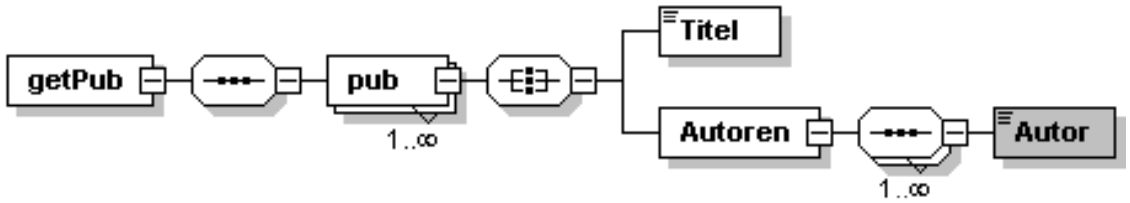
- Vorstellung der Arbeitsgruppe
- Organisatorisches
- Informationssysteme
- Informationsintegration am Beispiel
- Ausblick auf das Semester

Informationsintegration



Beispiel – Web Service A

- Standort: Trier
- Operation:
  - getPubByAuthor(firstName, lastName)
  - getPubByTitle(title)
- Output-Struktur:



Beispiel – Web Service A Output


```
<xs:element name="pub" maxOccurs="unbounded">
    <xs:complexType>
        <xs:all>
            <xs:element name="Titel" type="xs:string" nillable="true"/>
            <xs:element name="Autoren">
                <xs:complexType>
                    <xs:sequence maxOccurs="unbounded">
                        <xs:element name="Autor" type="xs:string" nillable="false"/>
                    </xs:sequence>
                </xs:complexType>
            </xs:element>
        </xs:all>
    </xs:complexType>
</xs:element>
```

Beispiel – Web Service B

- Standort: HPI
- Operation: myPubs(Autor, Jahr)
- Struktur:

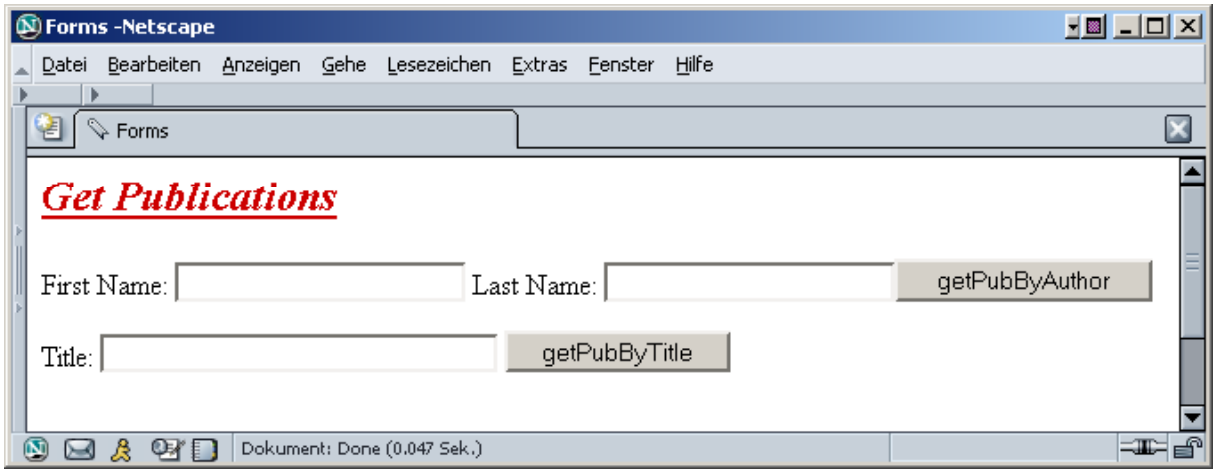
Beispiel – Web Service B Output


```
<xs:element name="publication" maxOccurs="unbounded">
    <xs:complexType>
        <xs:all>
            <xs:element name="Title" type="xs:string" nillable="true"/>
            <xs:element name="Auth" type="xs:string" nillable="false"/>
            <xs:element name="Year" type="xs:string" nillable="false"/>
        </xs:all>
    </xs:complexType>
</xs:element>
```

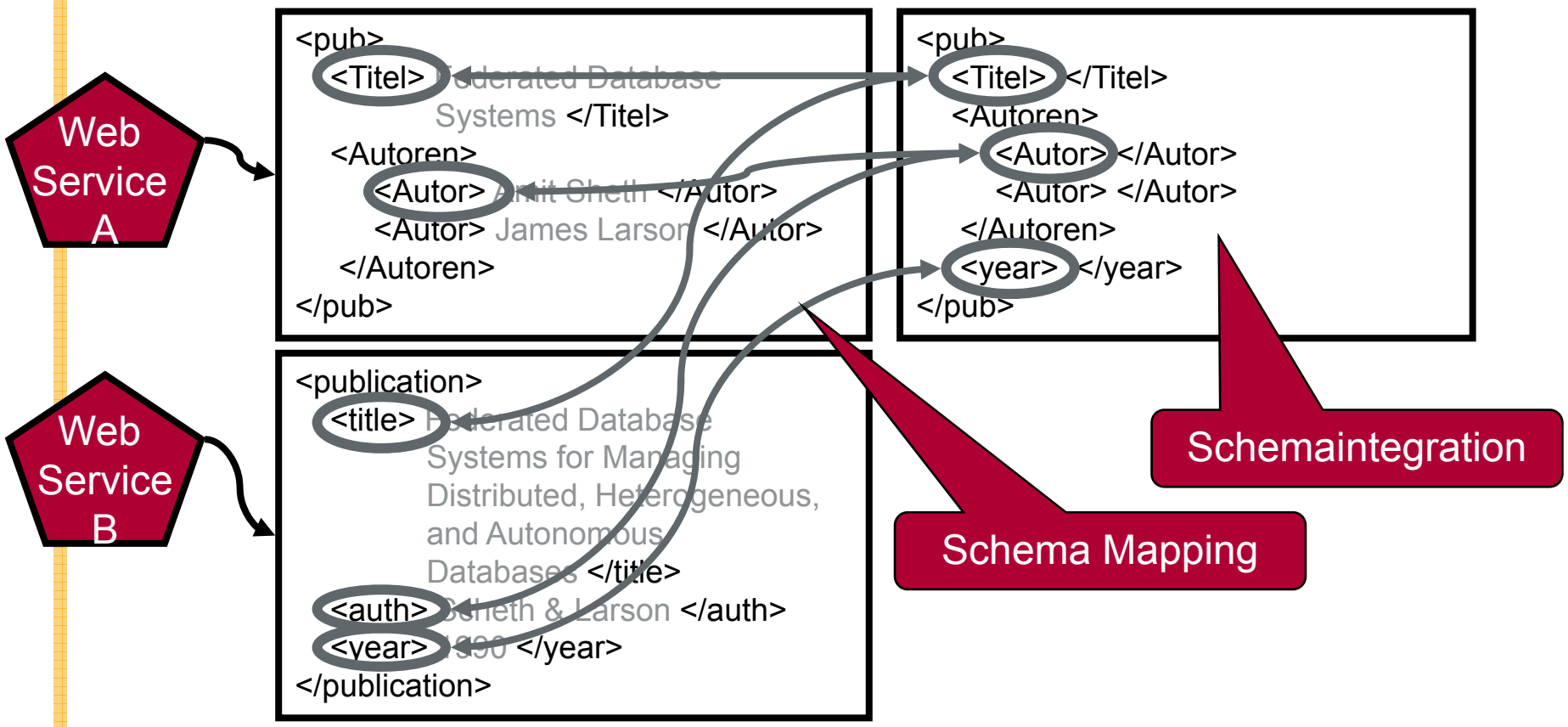
Integration von Web Services A & B

1. Nutzerschnittstelle
2. Schema Integration / Schema Mapping
3. Anfrage-Umwandlung
4. Zeit abschätzen (Optimierung)
5. Requests an beide Services abschicken
6. Antworten einholen
7. Objektidentifikation
8. Integrationsschritte
    1. Konfliktlösung etc.
    2. Entscheidung kleinster gemeinsamer Nenner?
    3. Durchführung (deklarativ, prozedural)
9. Anzeige beim Nutzer

Nutzerschnittstellen



Informationsintegration

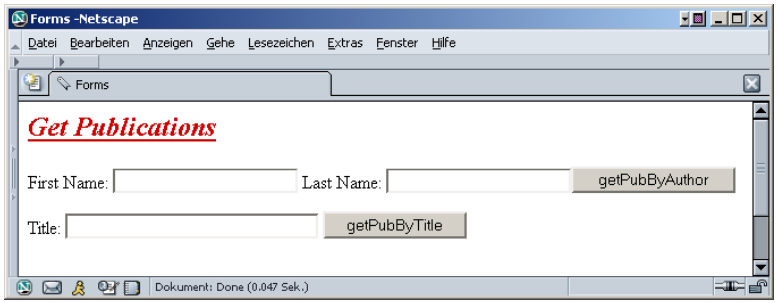


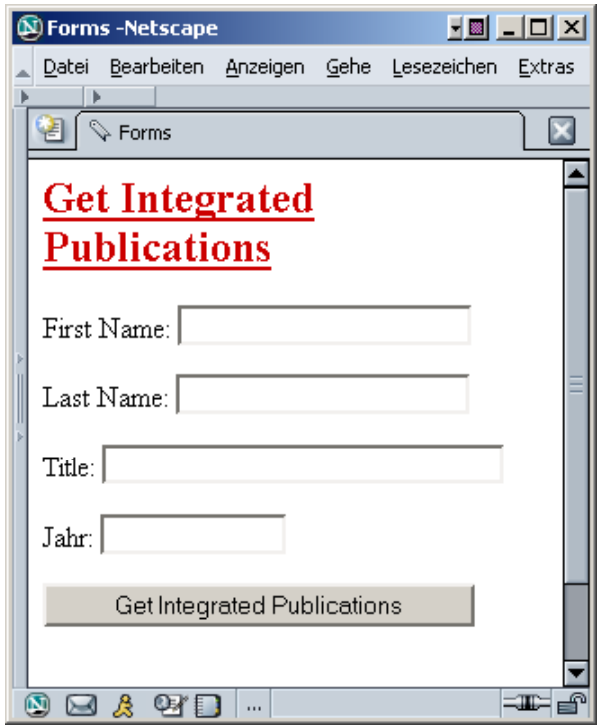
Modellierung durch eine Menge von Anfragen (Views)

Anfrage Umwandlung

Integration der Anfrage durch Mediator:

- Integrierte Schnittstelle
- Z.B.
  Concat(First Name, Last Name)
  = Autor





Anfrageoptimierung

- Was ist besser: Eine schnelle Antwort oder vollständige Antwort?
  - Web Service A in Trier (remote)
  - Web Service B in Griebnitzsee (local)
  - Web Service A hat mehr Attribute und mehr Objekte.
  - Web Service B hat weniger Attribute.
- Außerdem:
  - Eine Suche nach „year" kann nur durch Web Service B beantwortet werden.
  - Transformationen können teuer sein.

Zwei Resultate

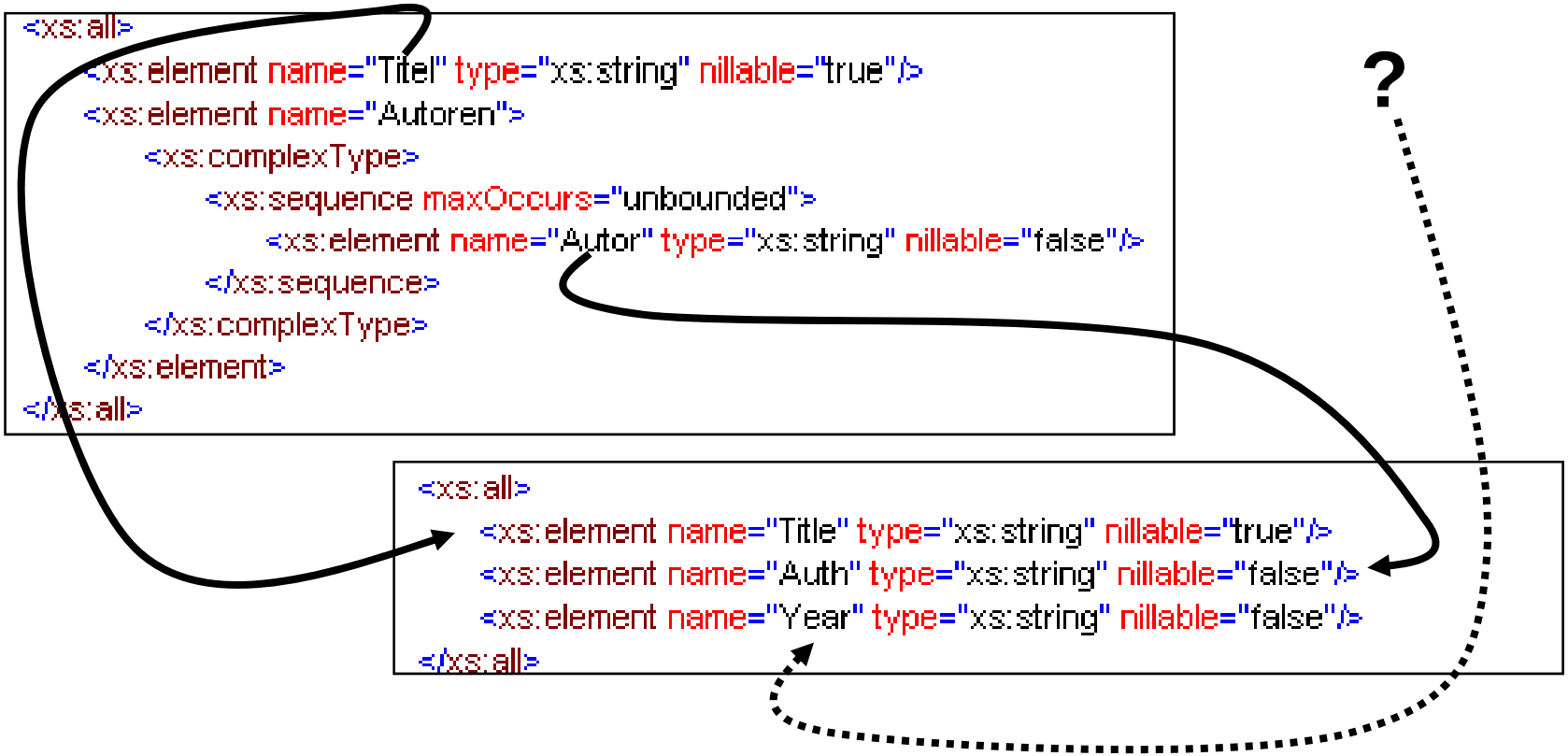
Web Service A

```
<?xml version="1.0" encoding="UTF-8" ?>
<!-- edited with XMLSPY v2004 rel. 2 U (http://www.xmlspy.com) by Felix Naumann
Universität zu Berlin)  -->
<!-- Sample XML file generated by XMLSPY v2004 rel. 2 U (http://www.xmlspy.com)
- <getPub xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:noNamespaceSchemaLocation="C:\Dokumente und Einstellungen\naumann\Eigene
  Dateien\Lehre\Winter03_04\VL InfoInt I WS03_04\getPub.xsd">
  - <pub>
      <Titel>Real-world Data is Dirty: The Merge/Purge Problem</Titel>
    - <Autoren>
        <Autor>Mauricio Hernandez</Autor>
        <Autor>Salvatore Stolfo</Autor>
      </Autoren>
    </pub>
  - <pub>
      <Titel>MAC: Merging Autonomous Content</Titel>
    - <Autoren>
        <Autor>Felix Naumann</Autor>
        <Autor>Jens Bleiholder</Autor>
      </Autoren>
    </pub>
</getPub>
```

Web Service B

```
<?xml version="1.0" encoding="UTF-8" ?>
<!-- Sample XML file generated by XMLSPY v2004 rel. 2 U (http://www.xmlspy.com)
- <myPub xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:noNamespaceSchemaLocation="C:\Dokumente und Einstellungen\naumann\Eigene
  Dateien\Lehre\Winter03_04\VL InfoInt I WS03_04\myPubs.xsd">
  - <publication>
      <Title>Merging Autonomous Content</Title>
      <Auth>Naumann</Auth>
      <Year>2003</Year>
    </publication>
  - <publication>
      <Title>Object Mathcing for Information Integration</Title>
      <Auth>Doan</Auth>
      <Year>1999</Year>
    </publication>
</myPub>
```

Schema Matching



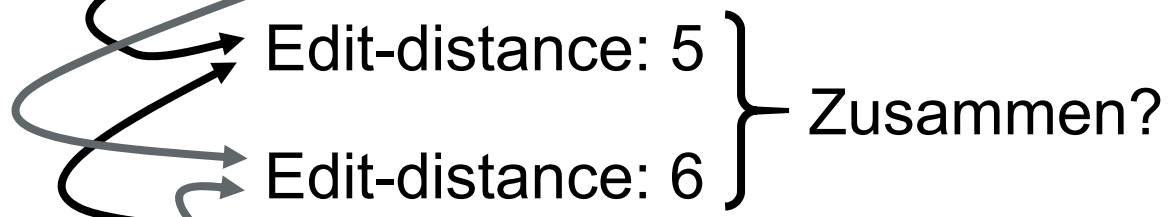
Objektidentifikation


```
- <pub>
    <Titel>Real-world Data is Dirty: The Merge/Purge Problem</Titel>
  - <Autoren>
      <Autor>Mauricio Hernandez</Autor>
      <Autor>Salvatore Stolfo</Autor>
    </Autoren>
  </pub>
- <pub>
    <Titel>MAC: Merging Autonomous Content</Titel>
  - <Autoren>
      <Autor>Felix Naumann</Autor>
      <Autor>Jens Bleiholder</Autor>
    </Autoren>
  </pub>
</getPub>
```

```
- <publication>
    <Title>Merging Autonomous Content</Title>
    <Auth>Naumann</Auth>
    <Year>2003</Year>
  </publication>
- <publication>
    <Title>Object Mathcing for Information Integration</Title>
    <Auth>Doan</Auth>
    <Year>1999</Year>
  </publication>
</myPub>
```

Objektidentifikation


```
- <pub>
    <Titel>MAC: Merging Autonomous Content</Titel>
  - <Autoren>
      <Autor>Felix Naumann</Autor>
      <Autor>Jens Bleiholder</Autor>
    </Autoren>
  </pub>
</getPub>
```

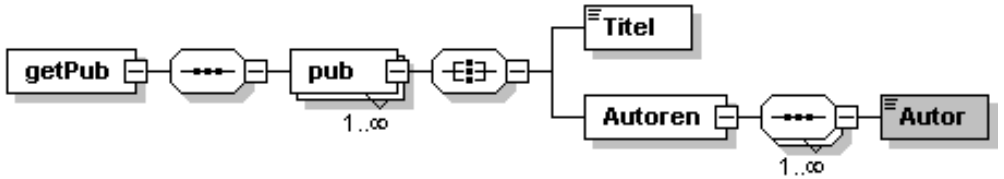



```
- <publication>
    <Title>Merging Autonomous Content</Title>
    <Auth>Naumann</Auth>
    <Year>2003</Year>
  </publication>
```

Stand der Dinge

- Wir haben die heterogenen Informationen.
- Wir wissen, was wir integrieren wollen.
- Aber noch nicht wie:
  - Integriertes Schema
  - Integrierte Daten

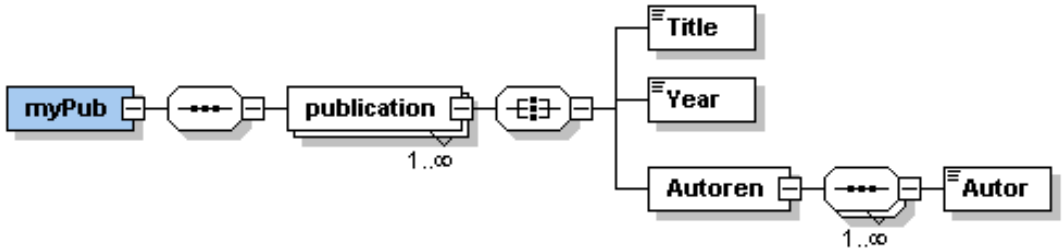
Angestrebtes Integrationsergebnis



+

=

Integriertes Schema:

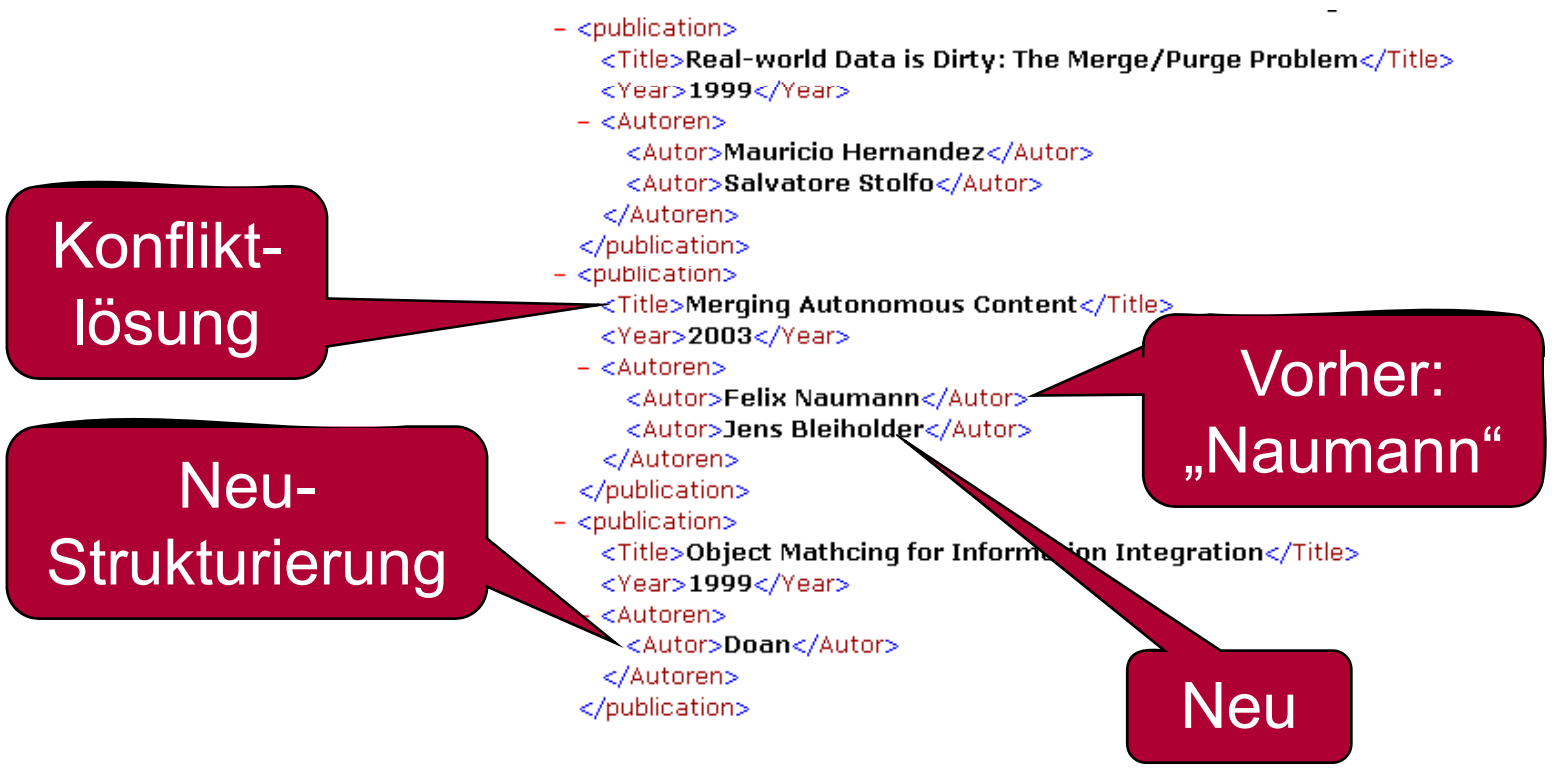


Angestrebtes Integrationsergebnis

Integrierte Daten:

```
- <publication>
    <Title>Real-world Data is Dirty: The Merge/Purge Problem</Title>
    <Year>1999</Year>
  - <Autoren>
      <Autor>Mauricio Hernandez</Autor>
      <Autor>Salvatore Stolfo</Autor>
    </Autoren>
  </publication>
- <publication>
    <Title>Merging Autonomous Content</Title>
    <Year>2003</Year>
  - <Autoren>
      <Autor>Felix Naumann</Autor>
      <Autor>Jens Bleiholder</Autor>
    </Autoren>
  </publication>
- <publication>
    <Title>Object Mathcing for Information Integration</Title>
    <Year>1999</Year>
  - <Autoren>
      <Autor>Doan</Autor>
    </Autoren>
  </publication>
```

Integrierte Daten – was ist passiert?



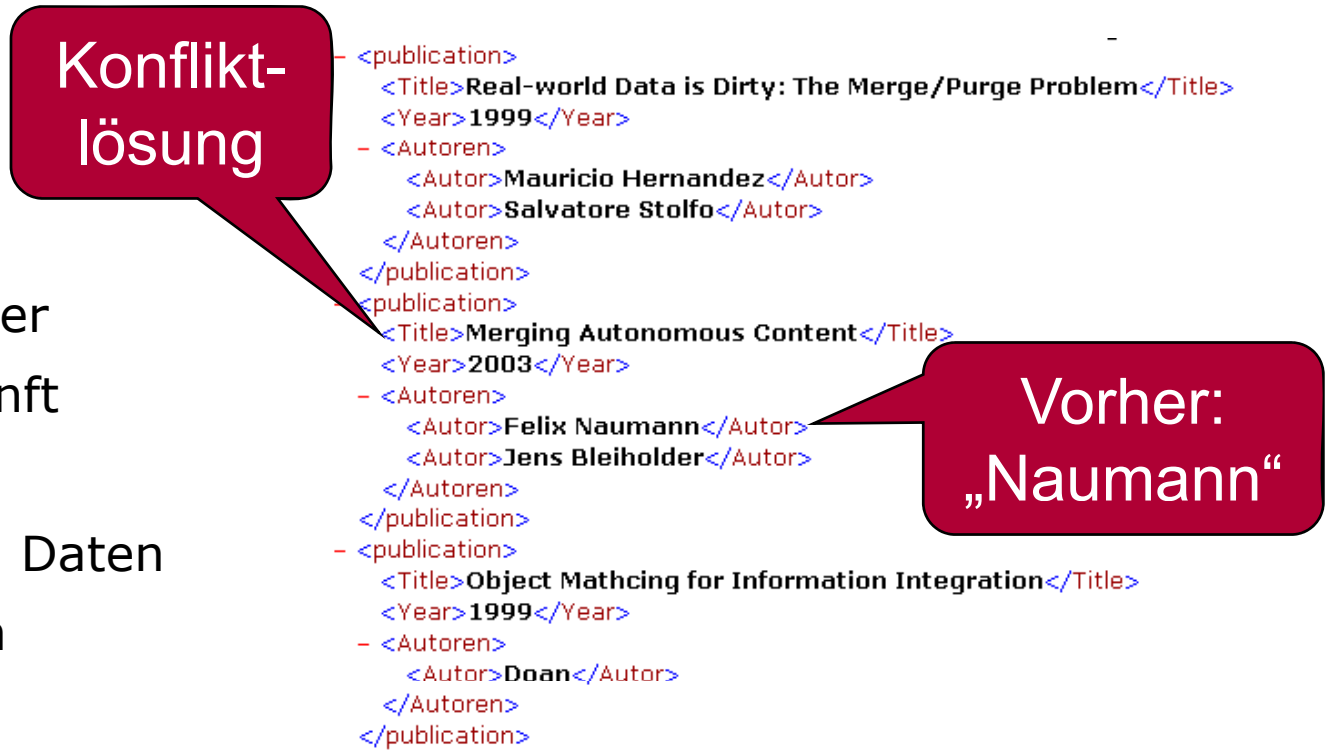
Implementierung

- Auf Folien ist alles klar, aber wie implementieren?
- Deklarativ?
  - SQL, XQuery, XSLT
  - Oft nicht alles möglich
  - Langsam
- Prozedural?
  - Java, C++
  - Schlecht wartbar
  - Schnell

Anzeige beim Nutzer

Visualisierung der

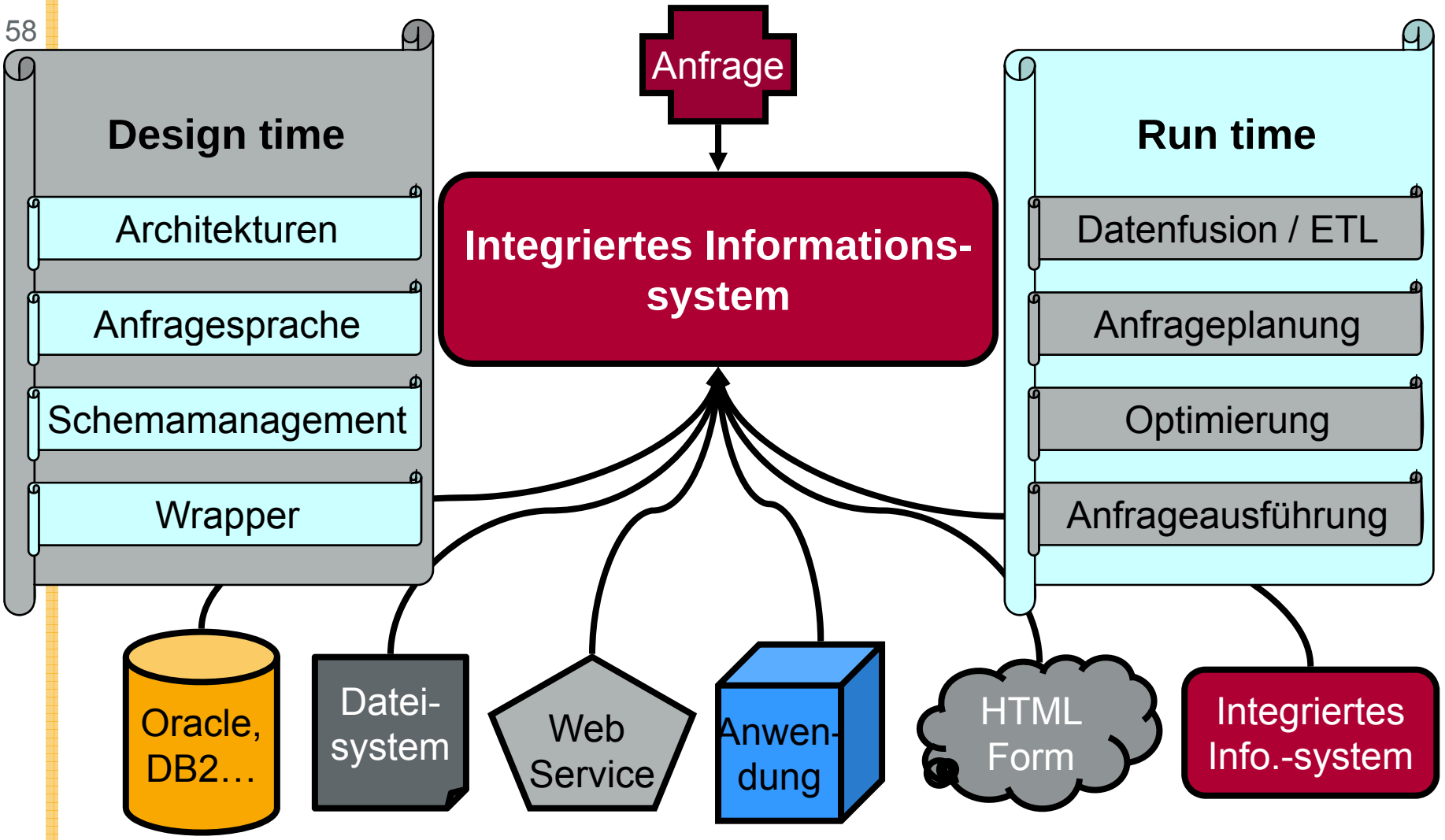
- Datenherkunft
- Qualität
- veränderten Daten
- Operationen



Überblick

- Vorstellung der Arbeitsgruppe
- Organisatorisches
- Informationssysteme
- Informationsintegration am Beispiel
- Ausblick auf das Semester

Integrierte Informationssysteme



Erste Hälfte

- Einführung in die Informationsintegration
- Szenarien der Informationsintegration
- Verteilung und Autonomie
- Heterogenität

Problemstellung

- Materialisierte und virtuelle Integration
- 5-Schichten-Architektur
- Mediator/Wrapper-Architektur / PDMS

Architekturen

- Schema Mapping
- Schema Matching
- SchemaSQL

Mapping

- Global-as-View und Lokal-as-View Modellierung
- Global-as-View Anfragebearbeitung

Modellierung

Zweite Hälfte

- Containment & Local-as-View Anfragebearbeitung
- Bucket Algorithmus
- Verteilte Anfragebearbeitung

Anfragen

- Duplikaterkennung
- Datenfusion – Union & Co.
- DWH, ETL & Data lineage
- Informationsqualität

Datenintegration

- Hidden Web
- Semantic Web

Anwendungen

Fragen, Wünsche und Vorstellungen

- Jetzt, oder...
- Raum: A.1-13
- Sprechstunden: Dienstags 15-16 Uhr
  oder n.V.
- Email: naumann@hpi.uni-potsdam.de
- Telefon: (0331) 5509 280

The end.