



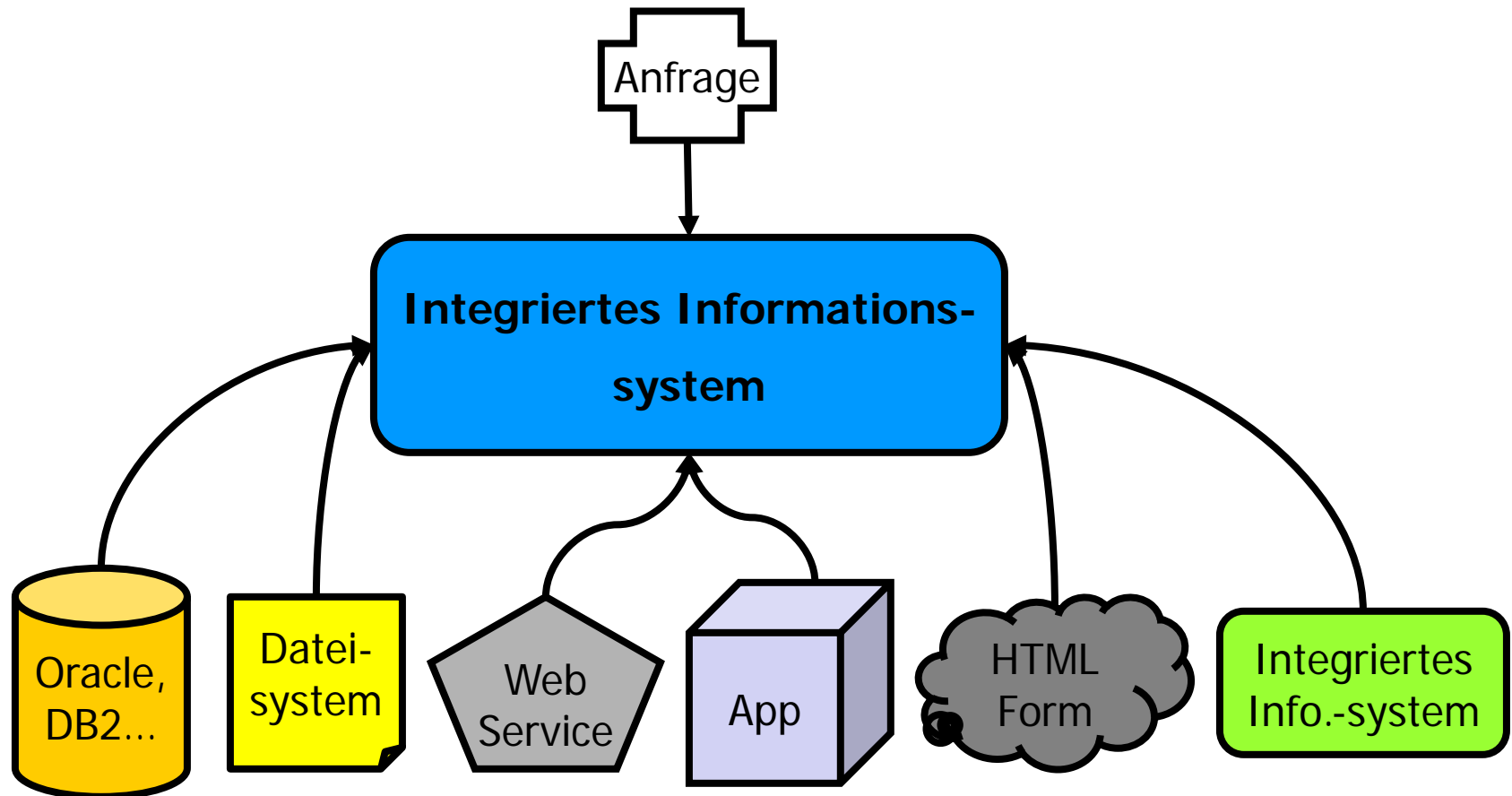
Integration genomischer Daten

Ulf Leser

Topics Today

- Data Integration
- Data Integration for the Life Sciences
- Integration within the project

Informationsintegration



Was ist Informationsintegration?

- Kurz: Homogener Zugriff auf den Inhalt verschiedener Datenquellen
- Lang: Informationsintegration bezeichnet die **korrekte, vollständige und effiziente** Bereitstellung des Inhalts verschiedener, **verteilter, autonomer und heterogener** Quellen **an einer Stelle** in Form einer **einheitlichen und strukturierten** Informationsmenge mit dem Ziel, eine **effektiven Nutzung** durch Nutzer und Anwendungen zu ermöglichen

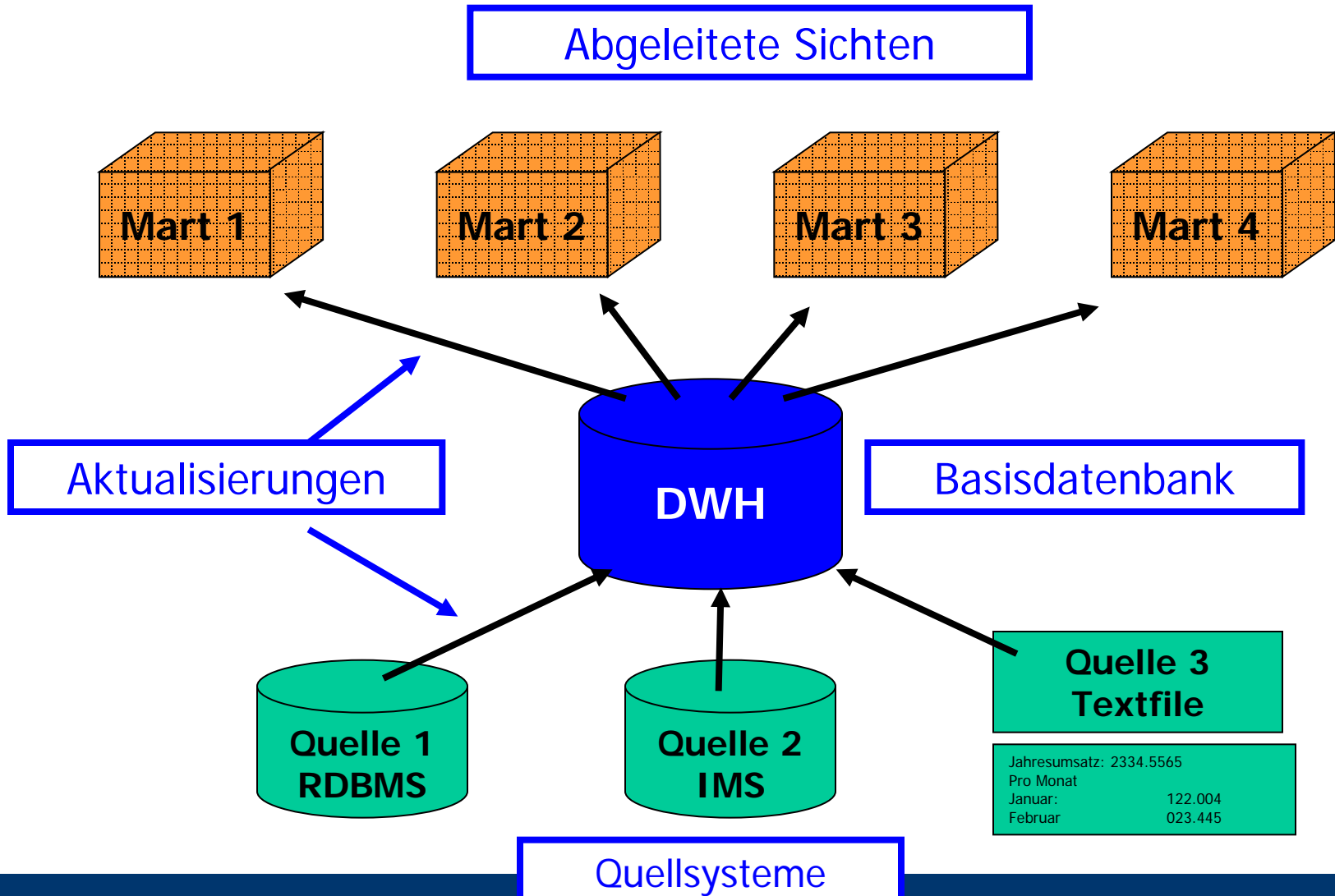
Ein uraltes Problem

- Seit 50 Jahren auf der Forschungsagenda
- Wird immer schwieriger und immer wichtiger
 - Web, Internet, Vernetzung
 - Viele, viele Quellen
 - Neue Formate und Datenmodelle (EXCEL, XML, GIS, OO,...)
 - Neue Arten von Anfragen (Ranking, Spatial, Text, Web, Mining ...)
 - Neue Arten von Nutzern (Laien (Web), Manager, ...)
 - Neue Anforderungen (24x7x365, schnell, Ad-Hoc, Online)
 - Neue Anwendungen
 - Marktplätze, eCommerce, eProcurement
 - Virtual enterprise, Web services, SOA
 - Data Markets, Mashups, Web

Warum ist es schwer?

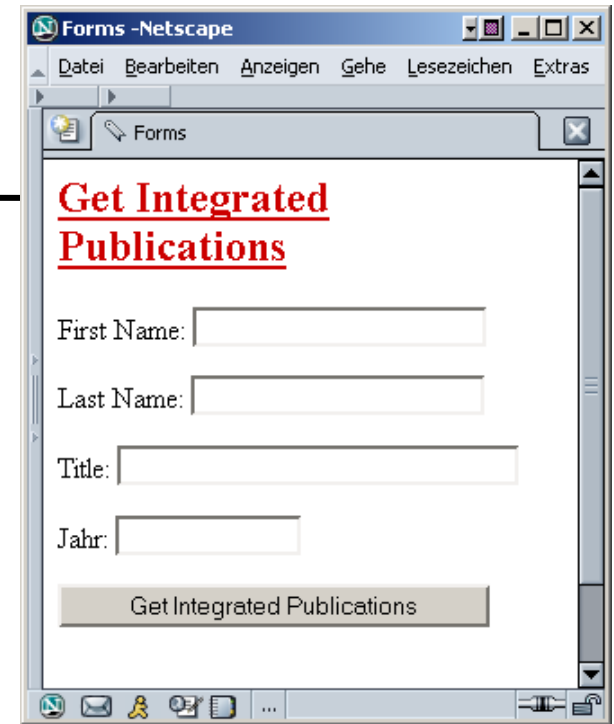
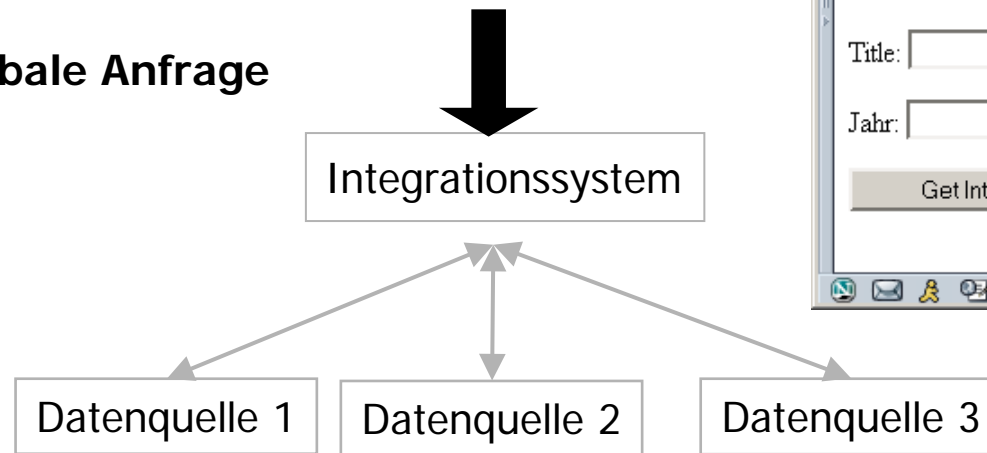
- Technische Gründe
 - Verschiedene Plattformen, Anfragesprachen, Policies, ...
 - Verteilung, Anfragebearbeitung über mehrere Systeme
- Semantische Gründe
 - Heterogenität auf allen Ebenen (Daten, Schema, Sprachen)
 - Semantik von Begriffen ist **kontextabhängig**
 - **Semantik** ist schwer zu beschreiben
- **Soziologische/psychologische Gründe**
 - Einblick in „fremde“ Datenbanken muss gestattet werden
 - Menschen zur Zusammenarbeit überreden
 - Einhalten von Verabredungen und **Standards**

Materialisierte Integration

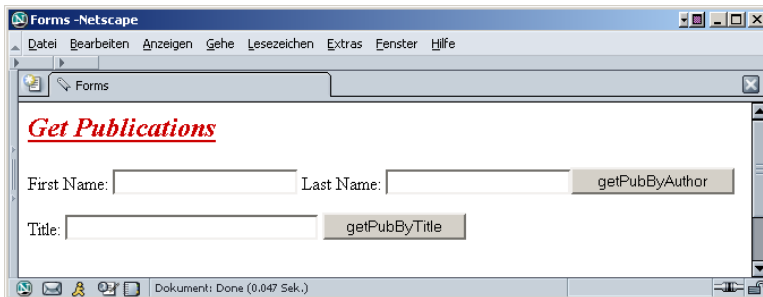
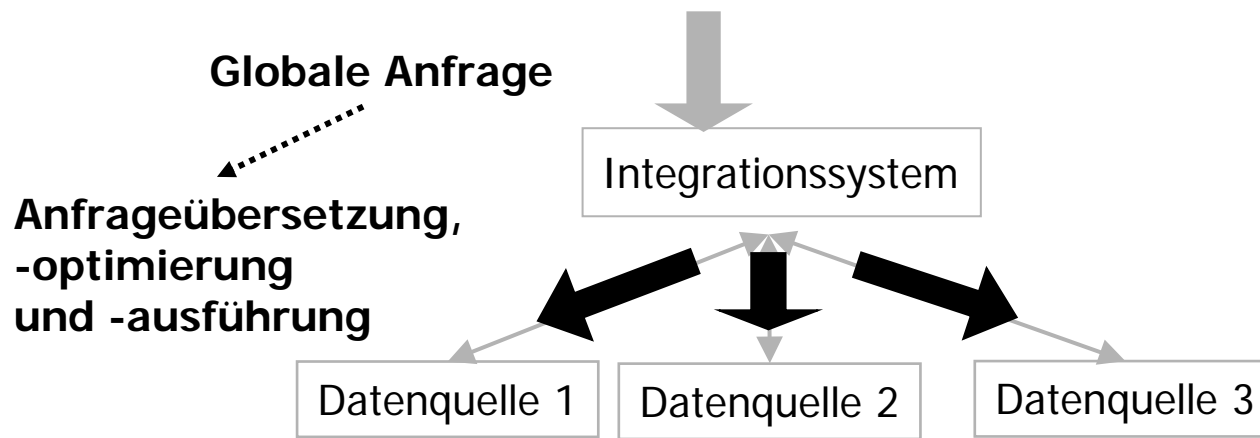


Virtualisierte Integration

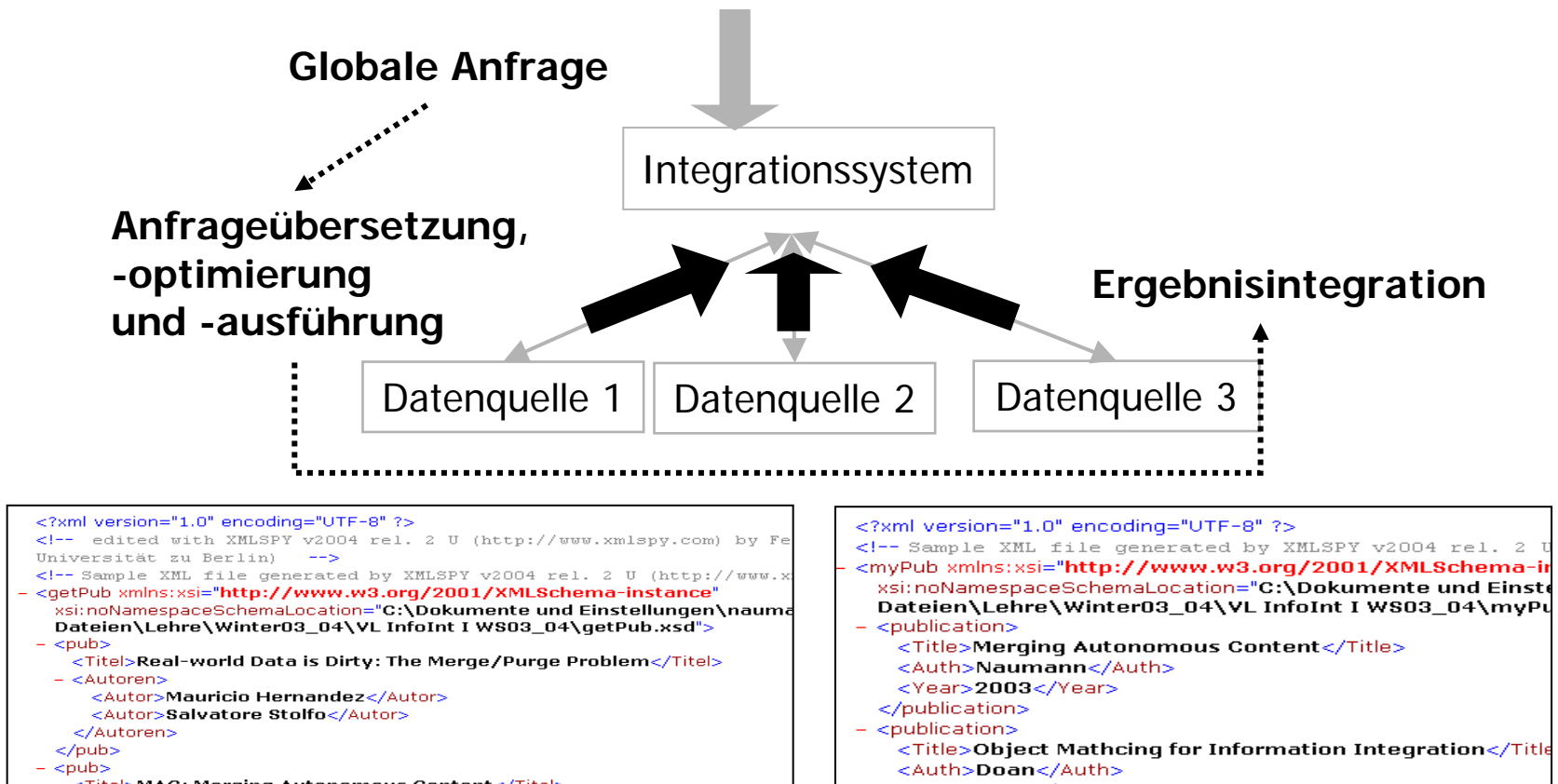
Globale Anfrage



Ablauf 2

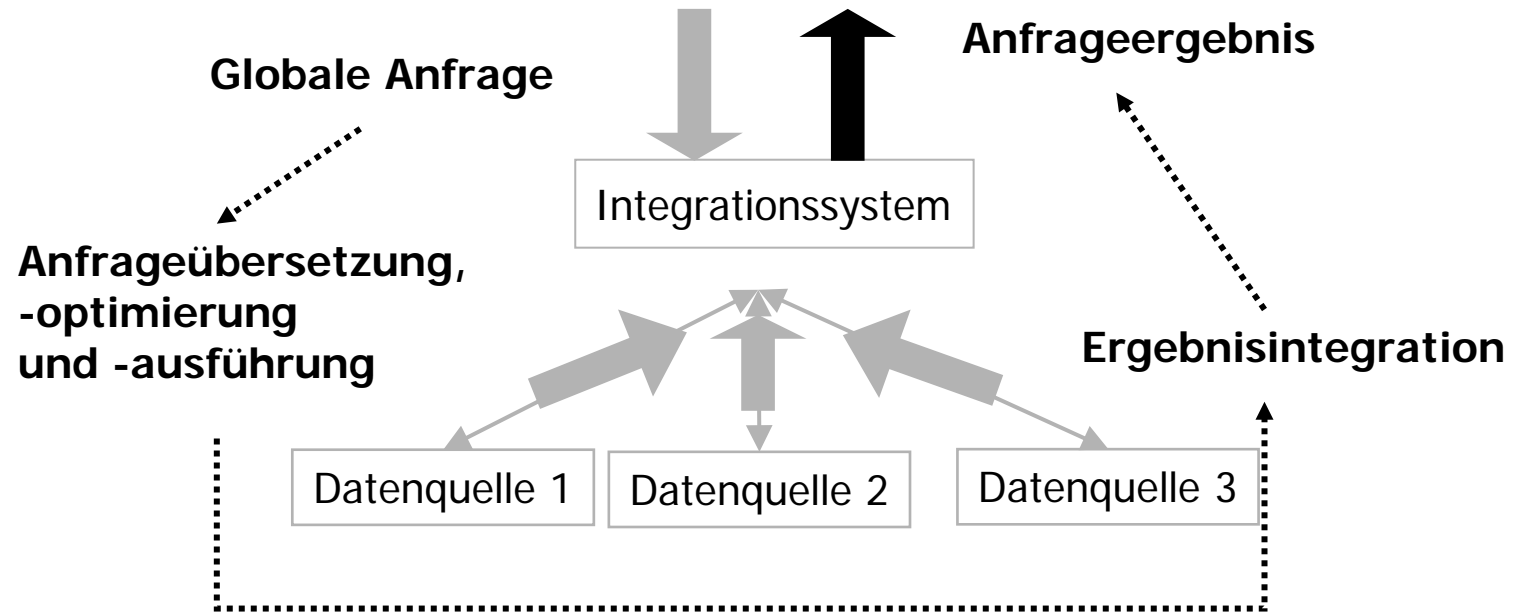


Ablauf 3



Ablauf 4

```
- <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>
```



Zwei Quellen, 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>
```

```
<?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>
```

Web Service B

„Was ist was?“ - Schema Matching

```
<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: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>
```



„Wer ist wer?“ - 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>
```

Intension und Extension einer Tabelle

- **Intension**

- Struktur einer Menge von Entitäten
- Semantik der Struktureinheiten
- Schema - statisch

- **Extension**

- Zustand der Tabelle
- Menge von Entitäten
- Zeilen - dynamisch

Buch		
ISBN	Titel	Autor
3442727316	Moby Dick	Herman Melville
3491960827	Robinson Crusoe	Daniel Defoe
3462032283	Zwölf	Nick McDonell
3883891606	Timbuktu	Paul Auster
...

Intensionale Redundanz

Quelle 1

ISBN	Author	Pages
3442727316	Herman Melville	1056
978-3491960824	Daniel Defoe	644
3462032283	Nick McDonell	240
3883891606	Paul Auster	227

Quelle 2

ISBN	Autorname	Year
3491960827	Daniel Defoe	1719
3442727316	H Melville	1851
3462026496	Saul Bellow	1992

Extensionale Redundanz

ISBN	Author	Pages
3442727316	Herman Melville	1056
978- 3491960824	Daniel Defoe	644

ISBN	Autorname	Year
3491960827	Daniel Defoe	1719
3442727316	H Melville	1851

- Extensionale Redundanz: Menge der von zwei Quellen **gemeinsam repräsentierten Objekte** ist nicht leer
- Voraussetzung dafür ist intensionale Redundanz
 - Gleiche Objekte müssen aus der gleichen „Klasse“ sein
- Oftmals schwer zu erkennen (Duplikaterkennung)

Redundanz und Komplementierung

- Redundanz (Überlappung)
 - In Extension und Intension möglich
 - Segen: Ohne **minimale Redundanz** ist Integration meist sinnlos
 - Was gehört zu was?
 - Fluch: Führt zu **Widersprüchen**, Doppelungen, ...
- Komplementierung
 - Informationen mehrerer Quellen werden zu einem **größeren Ganzen** integriert
 - Der eigentliche Sinn von Informationsintegration

Im Projekt

- **Intensionale Redundanz: Manuell prüfen**
 - Wesentlicher gemeinsamer Schlüssel ist die Genomkoordinate
 - Andere Werte dürfen nicht unter Synonymen auftauchen
 - Beispiel: Attribute „diseases“, „disorder“, „maliciency“, „dis.“, ...
- **Extensionale Redundanz: Wird nicht geprüft**
 - Wir gehen von Disjunktheit aller Quellen aus

Inhalt dieser Vorlesung

- Heterogenität
 - Technische Heterogenität
 - Syntaktische Heterogenität
 - Datenmodellheterogenität
 - Strukturelle Heterogenität
 - Schematische Heterogenität
 - Semantische Heterogenität

Technische Heterogenität

Ebene	Mögliche Ausprägungen
Anfragemöglichkeit	Anfragesprache, parametrisierte Funktionen, Formulare (engl. <i>canned queries</i>)
Anfragesprache	SQL, XQuery, Volltextsuche
Austauschformat	Binärdaten, XML, HTML, tabellarisch
Kommunikationsprotokoll	HTTP, JDBC, SOAP

Syntaktische Heterogenität

- Unterschiedliche Darstellung desselben Sachverhalts
 - Dezimalpunkt oder –komma
 - Euro oder €
 - Comma-separated oder tab-separated
 - HTML oder ASCII oder Unicode oder ...
 - Notenskala 1-6 oder „sehr gut“, „gut“, ...
 - Binärcodierung oder Zeichen
 - Datumsformate (12. September 2006, 12.9.2006, 9/12/2006, ...)
- Überwindung in der Regel **nicht problematisch**
 - Umrechnung, Übersetzungstabellen, ...

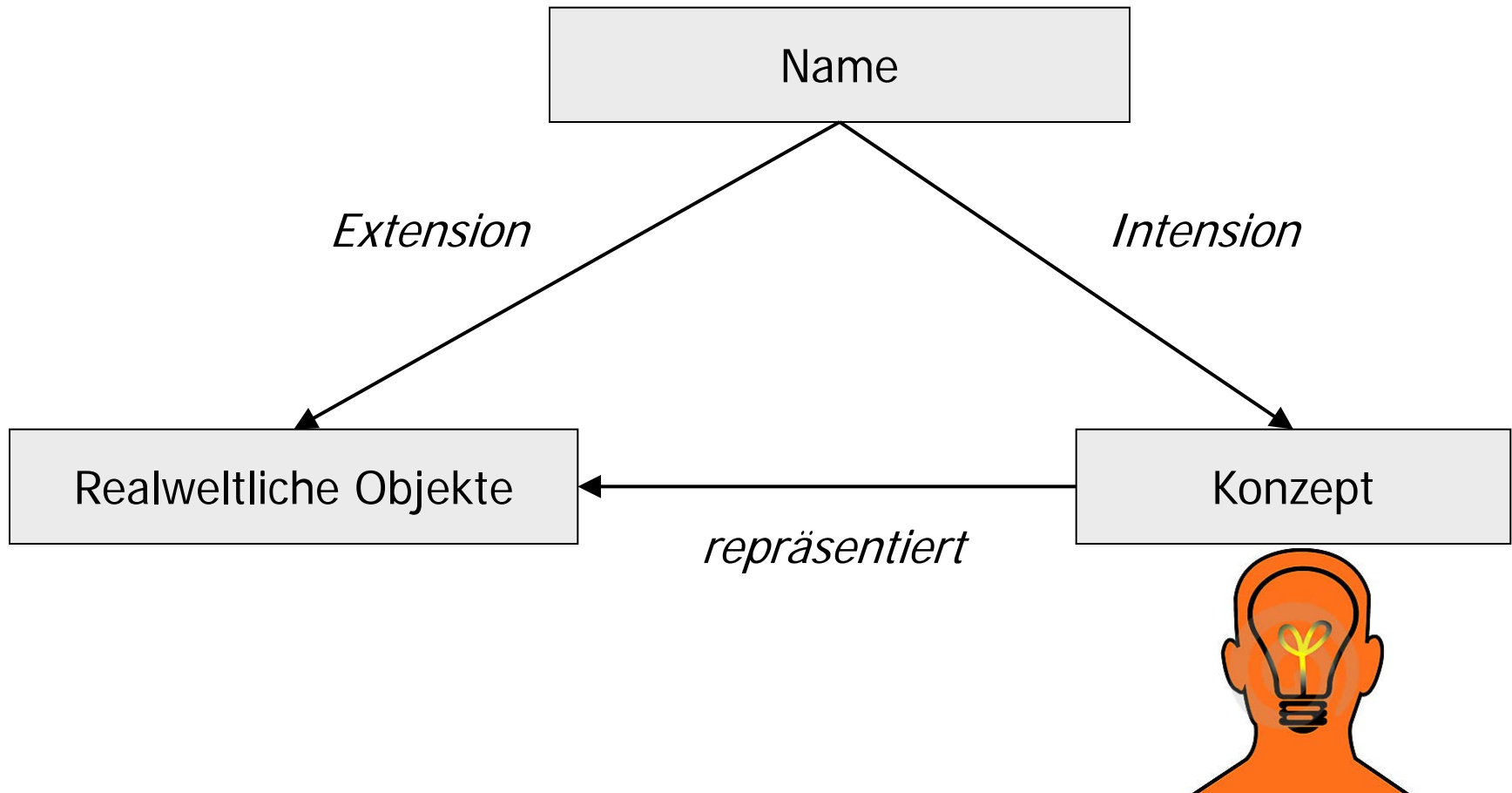
Datenmodellheterogenität

- Typische Datenmodelle
 - Einfach (CSV, EXCEL, ...)
 - Relational (Tupel)
 - XML (Hierarchisch)
 - Domänenspezifisch (EXPRESS, OPEN-GIS, ...)
- Zum **Austausch oder zur Speicherung**
 - **Black-Box-Sicht**
 - Entscheidend ist, was die Quelle liefert (also Austauschformat)
- Erfordert Konvertierung

Strukturelle Heterogenität

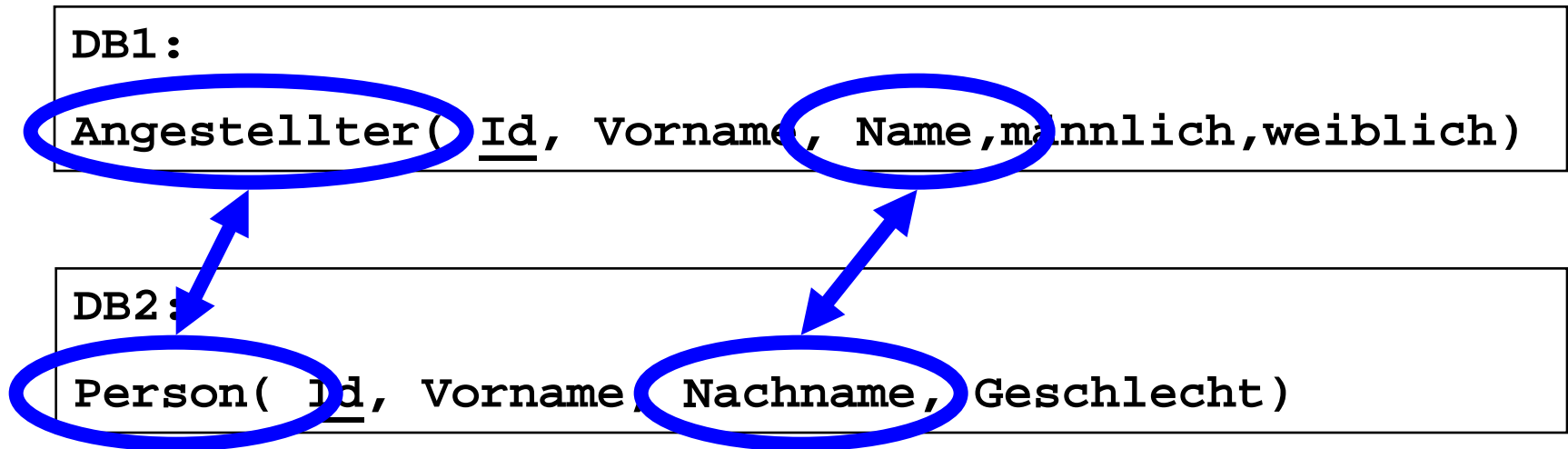
- Gleiche Dinge in **unterschiedlichen Schemata** ausdrücken
 - Andere Aufteilung von Attributen auf Tabellen
 - Fehlende / neue Attribute (wenn Intension nicht betroffen ist)
- Sehr oft mit semantischer Heterogenität verbunden
- Spezialfall: Schematische Heterogenität
 - Verwendung **anderer Elemente** eines Datenmodells
 - Kann meist nicht durch Anfragesprachen überwunden werden

Semantik von was?



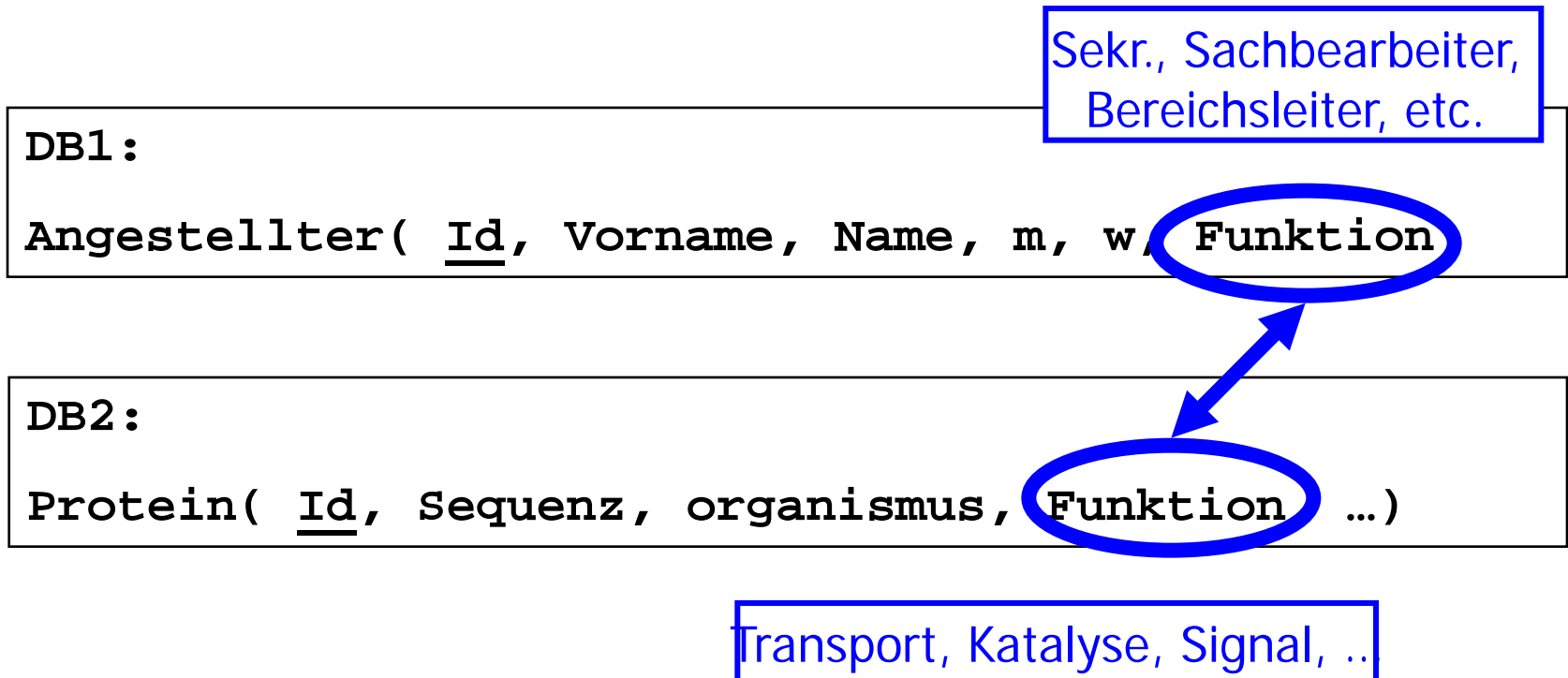
Synonyme

- Verschiedene Namen für dieselbe Menge
 - Immer im Kontext der Anwendung



Homonyme

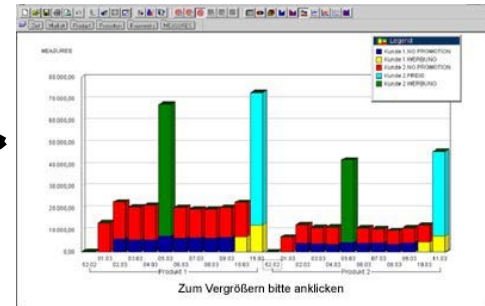
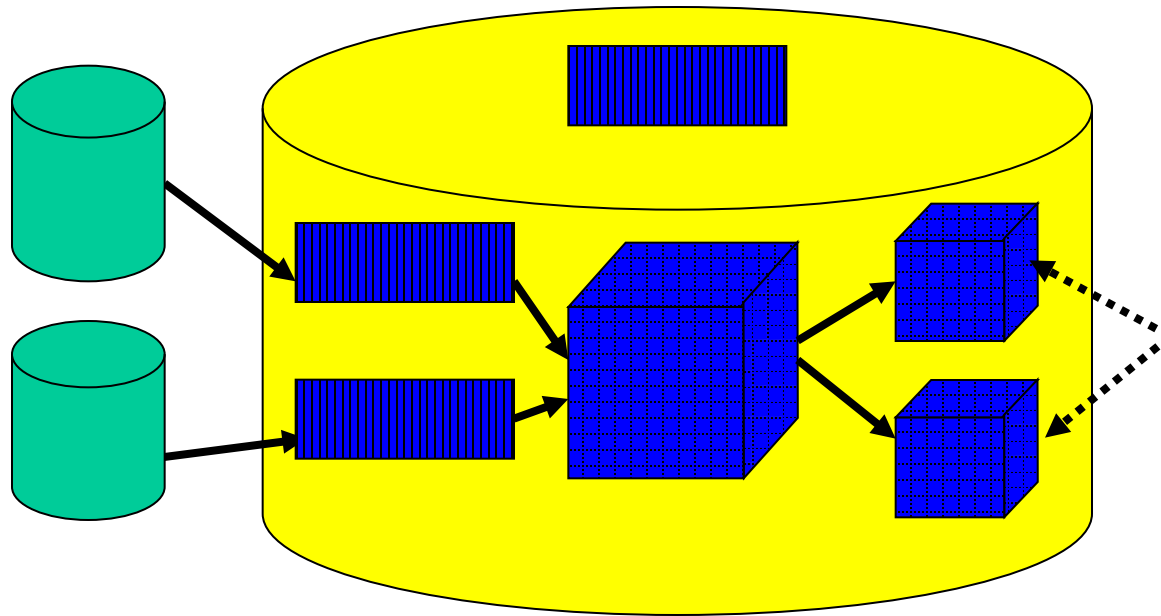
- Gleiche Namen für verschiedene Mengen
 - Treten oft bei Überschreitung von Domänengrenzen auf



Bedeutung: Woher nehmen?

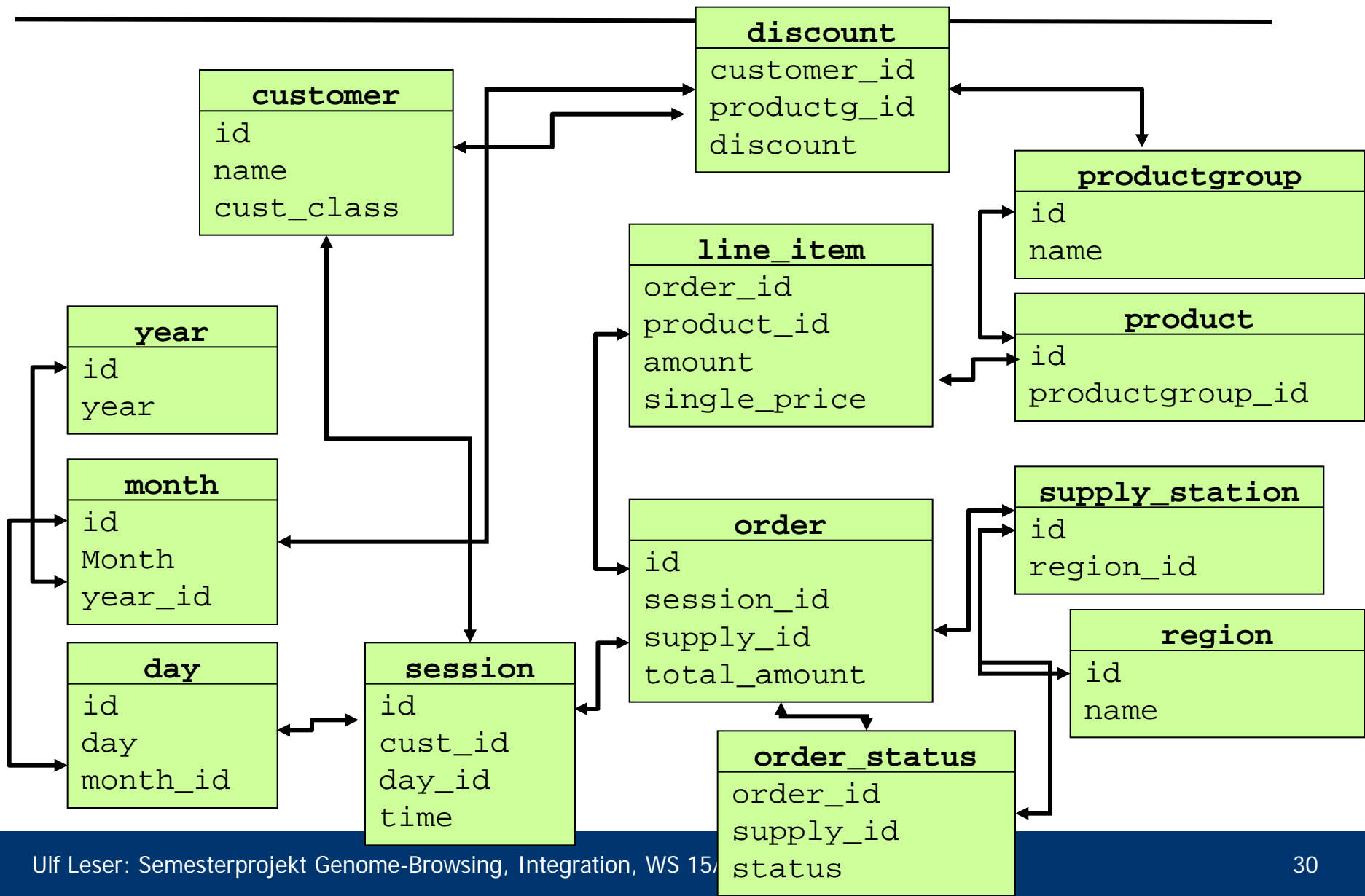
- Schemaelemente sind nur Namen
- Was bestimmt die **Semantik eines Namens?**
- Für Attributnamen
 - Datentyp
 - Constraints (Schlüssel, FK, unique, CHECK, ...)
 - Zugehörigkeit zu einer Relation
 - Andere Attribute dieser Relation
 - Beziehung der Relation zu anderen Relationen
 - Dokumentation
 - Vorhandene Werte
 - Wissen über den Anwendungsbereich
 - **Der Kontext**

ETL

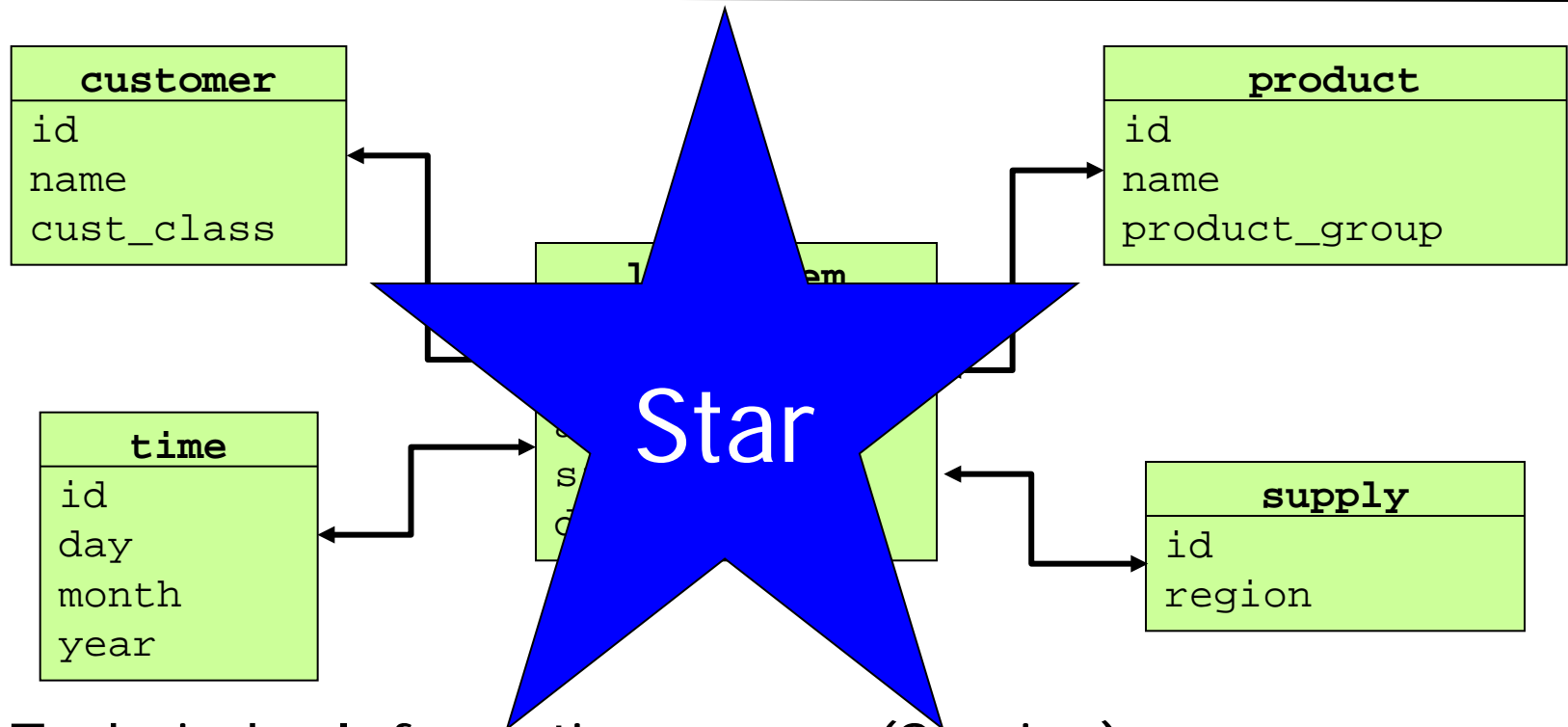


- Extraction
- Transformation
- Load

Normalisiertes Schema



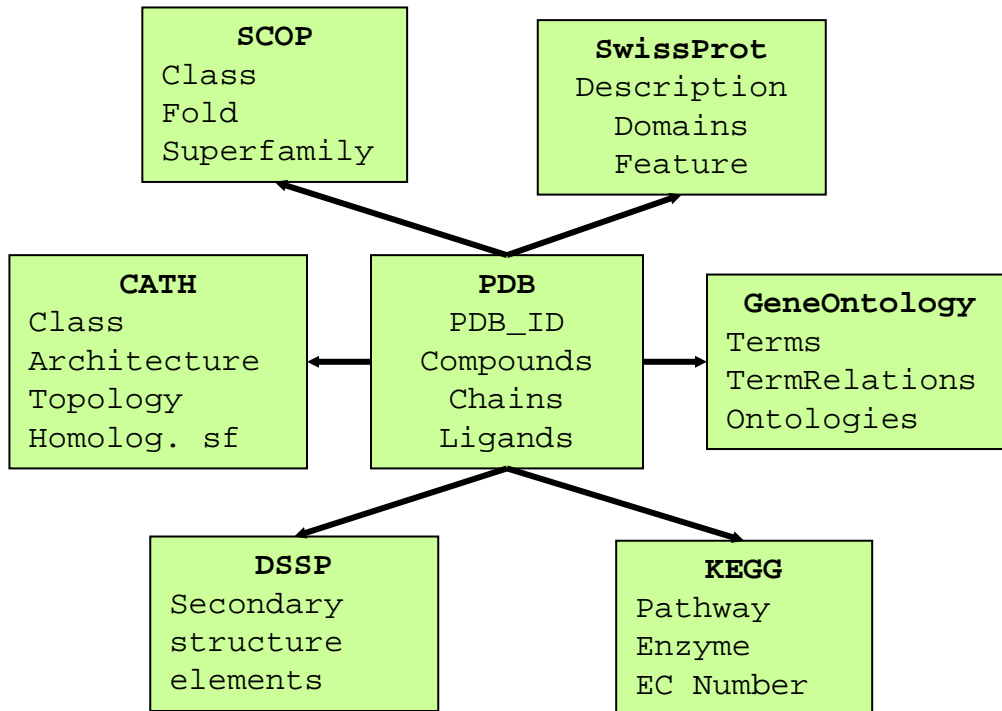
Multidimensionales Schema



- Technische Informationen raus (Session)
- Nur abgeschlossene Bestellungen aufnehmen (Orderstatus)
- Zusammenfassen (discount_rate)
- Denormalisieren (überall)

Multidimensional Integration

- Sources are dimensions for PDB entries
- There is no “data” integration



Advantages

- Clear data provenance
- Simplified maintenance
- Users recognize **their sources**
 - Quality, trust, ...
- **Intuitive query concept**

Disadvantage

- **No semantic integration**
- Redundancy, duplicates, data conflicts

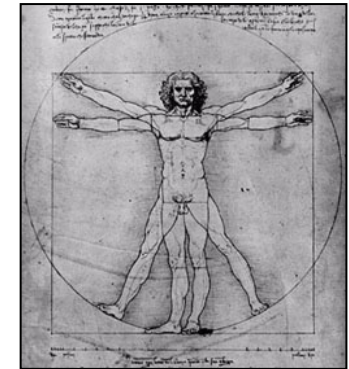
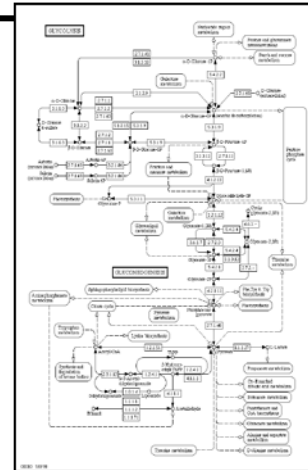
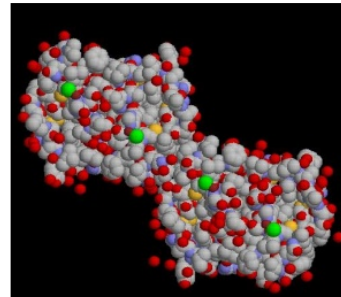
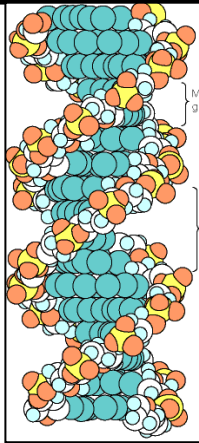
Topics Today

- Data Integration
- Data Integration for the Life Sciences
- Integration within the project

Data Integration for the Life Sciences, 1993

- Robbins, R. J. (1994). "Report of the invitational DOE Workshop on Genome Informatics I: Community Databases." [Rob94a]
 - DOE funded large parts of the HGP starting end of the 80ties
- *"Continued HGP progress will depend in part upon the ability of genome databases to answer increasingly **complex queries that span multiple community databases**. Some examples of such queries are given in this appendix."*

Database Perspective



Genomics

Sequence DBs
Gene DBs
Taxonomic DBs
TFBS-DBs
Epigenetic DBs
miRNA DBs
mRNA DBS

...

Proteomics

Structure DBs
Protein DBs
Small molecule DBs
Motive DBs
PPI DBs

...

Systems Biology

Pathway DBs
Regulation DBs
Signaling DBs
Metabolic DBs
Model DBs
Kinetic DBs

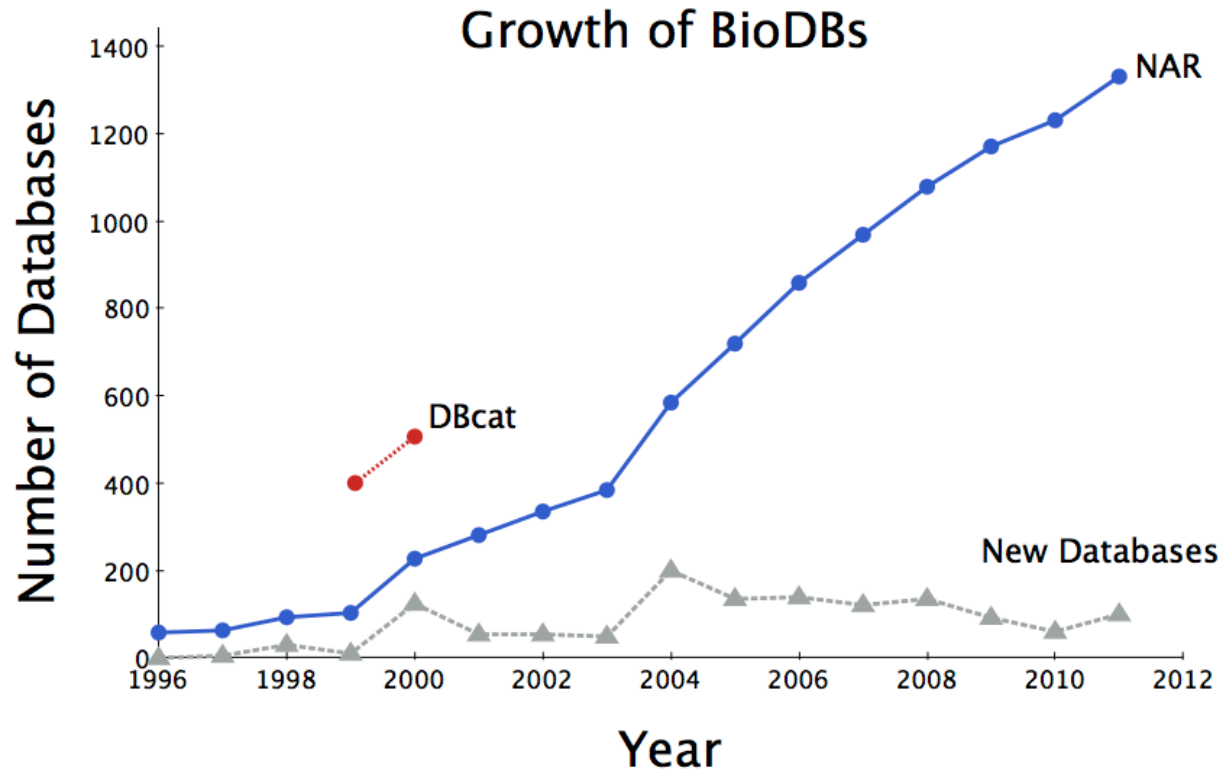
...

Medicine

Patient DBs
Biobanks
Drug DBs
Study DBs
Population DBs

...

There are 100reds of Them

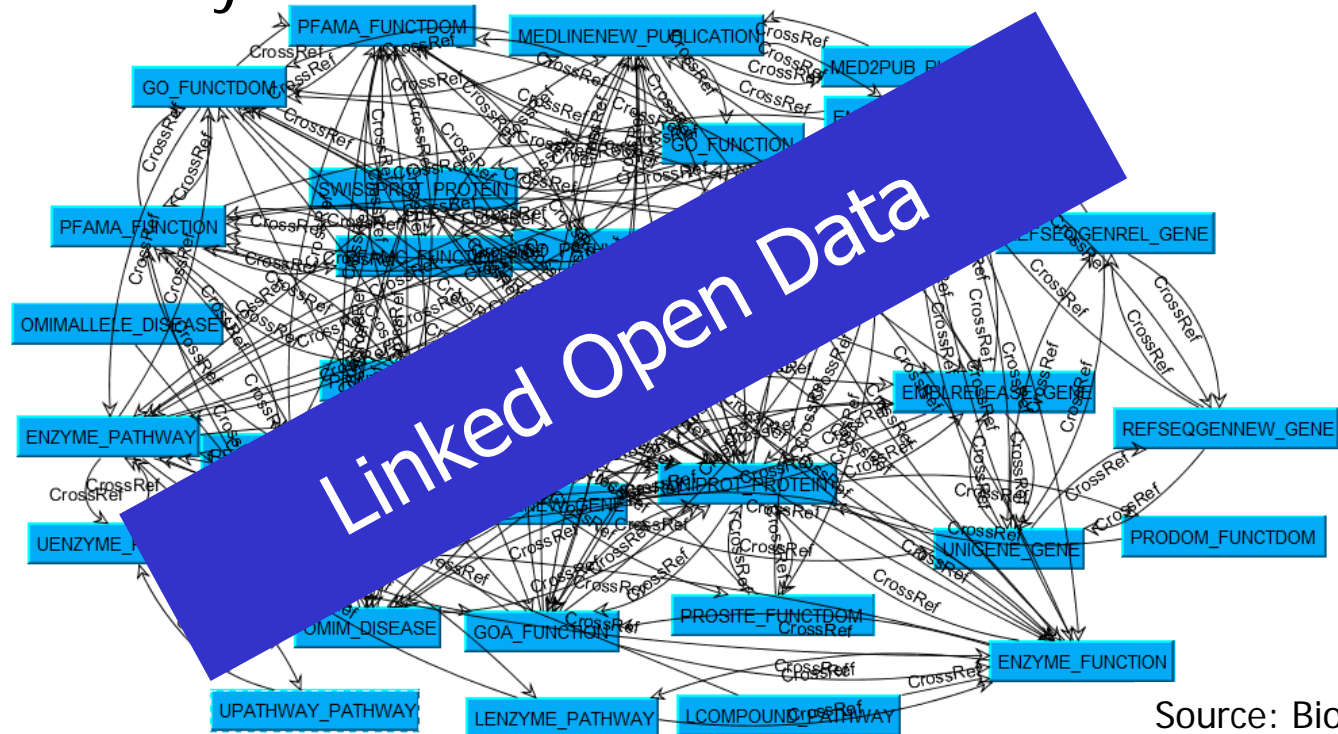


Number of existing (circles) and new databases (triangles) are plotted from 1996 to 2011. New databases are difference between the number of existing databases for each year. DBcat (red) is shown with NAR (blue) counts.

Copyright Geospiza 2011

Links

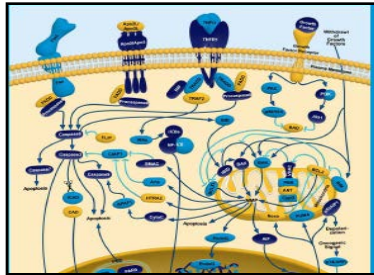
- BDB maintain links to many other BDBs
 - Instance level - external IDs, web browsing support
- No central authority for ID or links
- No consistency – “link hell”



Source: BioGuide

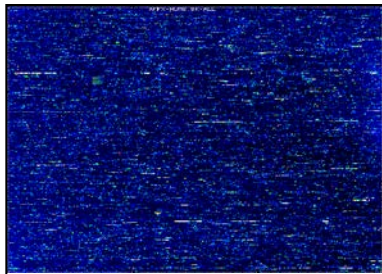
Types of "Data"

[Biocarta]



Feature key	Position(s)	Length	Description
Molecule processing			
<input type="checkbox"/> Chain	1 – 3685	3685	Dystrophin
Regions			
<input type="checkbox"/> Domain	1 – 240	240	Actin-binding
<input type="checkbox"/> Domain	15 – 119	105	CH 1
<input type="checkbox"/> Domain	134 – 237	104	CH 2
<input type="checkbox"/> Repeat	329 – 442	109	Spectrin 1
<input type="checkbox"/> Repeat	448 – 556	109	Spectrin 2
<input type="checkbox"/> Repeat	559 – 667	109	Spectrin 3
<input type="checkbox"/> Repeat	719 – 820	110	Spectrin 4
<input type="checkbox"/> Repeat	930 – 934	105	Spectrin 5
<input type="checkbox"/> Repeat	943 – 1045	103	Spectrin 6
<input type="checkbox"/> Repeat	1048 – 1154	107	Spectrin 7
<input type="checkbox"/> Repeat	1157 – 1263	107	Spectrin 8

[Affymetrics]



- Knowledge
 - Confirmed, abstract, condensed
 - Text, graphics
 - Publications
- Information
 - Interpreted, filtered, "data in context"
 - Objects, annotations
 - BDB – secondary databases
- Data
 - Measured - raw, noisy, context-free
 - Numbers, sequences, metadata
 - BDB – primary databases

The Presence

XML + Perl + MySQL

... or better ...

XML +
(Perl | Java | Python) +
(MySQL | Oracle | PostGreSql)

The Presence

- “Data Warehouses” approaches everywhere
 - Virtual integration is mostly dead
 - Despite frequent papers stating the opposite
 - Survival in some niches: DAS, some mash-ups
- Semantic integration performed **manually** (wrappers)
 - No schema matching, little query rewriting
- Several systems up-and-running integrating **dozens of sources**
 - Freshness in the presence of data cleansing is the hardest problem

Open Challenges

Effort	Integrating dozens of data sources still requires considerable effort
Analysis	Interesting (from a LS perspective) DI problems require complex analysis processes
Provenance	Users want to know exactly where each piece of data comes from
Quality	Finding the right answer, not „finding any answer“ or “finding all answers”

Some Research Trends

Data Integration Workflows	<ul style="list-style-type: none">• Integration means analysis, and analysis means integration• No schemas, no explicit semantics• Scientific workflow systems	Effort Analysis Provenance Quality
Ranking	<ul style="list-style-type: none">• Report results in a biologically meaningful order• Stays with queries, adds ranking• Requires a DI system in place	Effort Analysis Provenance Quality
Semantic Web	<ul style="list-style-type: none">• Reduce upfront cost of DI• No schemas, explicit semantics• Semantic Web tech. (RDF, SPARQL)	Effort Analysis Provenance Quality

Topics Today

- Data Integration
- Data Integration for the Life Sciences
- Integration within the project

Einordnung

- Integration von Daten über Mutationen in Menschen
- Viele Tupel, eher wenig Attribute
 - Müssen untersucht, ausgesucht, gemapped werden
 - Gemeinsamer Schlüssel ist (nur) die Position: Chromosome+int+int
 - Punktmutationen oder ausgedehnte Mutationen
- Transparenz: Quelle einer Mutation wichtig
 - Zwei „Quellen“: Datenquelle (Datenbank), Patient / Sample
 - Datenquellen aus vielen Samples werden auf Quellebene aggregiert
 - Häufigkeit einer Mutationen in der Quelle
- Quellen mögen Querverweise zu Krankheiten etc. haben – nicht relevant als Selektionskriterium in Anfragen
 - TCGA: Quelle/Krankheit als „Quelle“ (Lane) modellieren

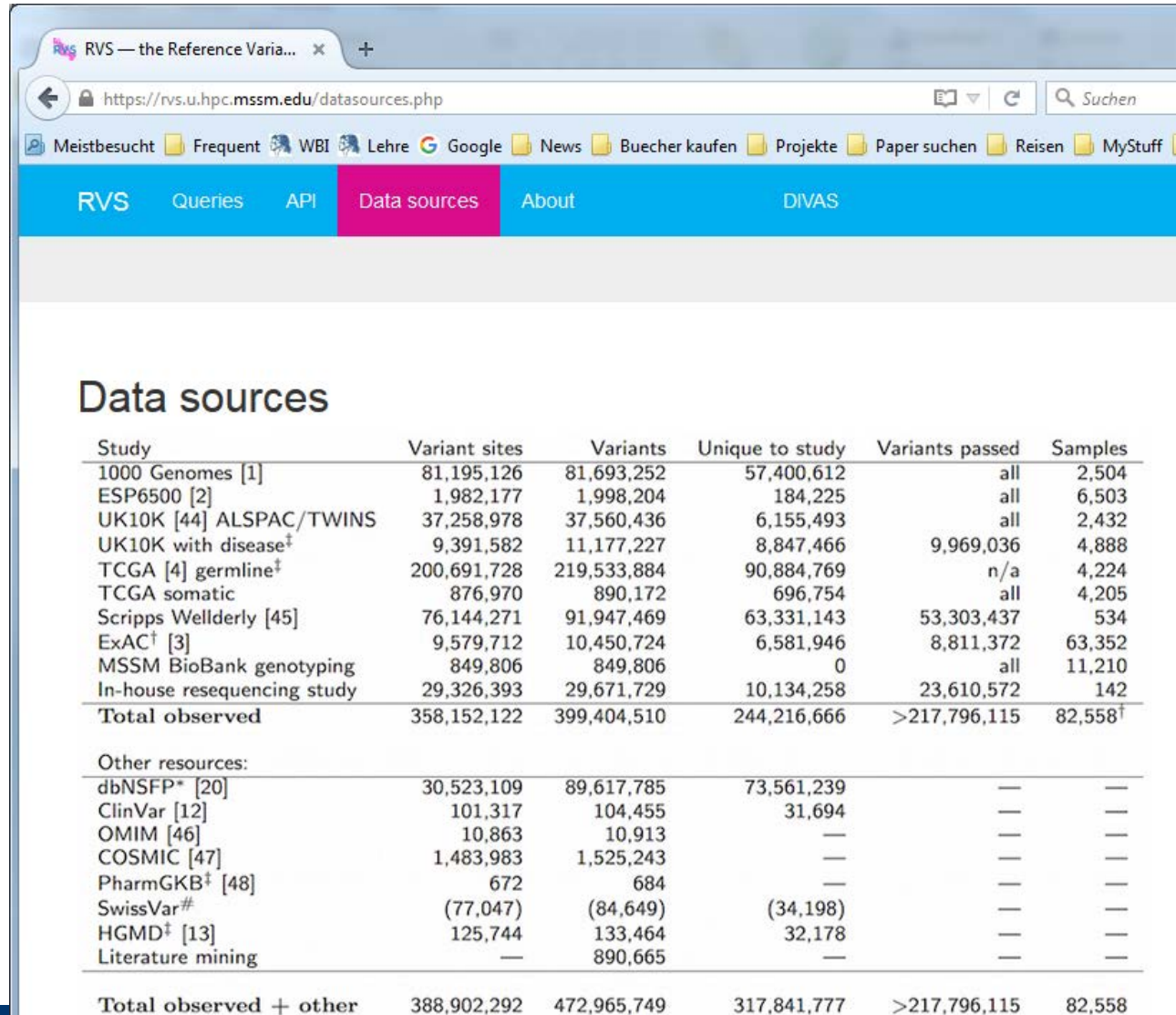
ETL

- Materialisierte Integration
 - Aktualisierung ist nicht Schwerpunkt
- ETL liest aus Flatfiles und schreibt in DB
 - Schnelles ETL: Bulk-Upload API verwenden
- Aggregation – in der Datenbank, beim Einlesen?
 - Ggf. staging area einführen
- Zielschema muss erstellt werden
 - Quellen als Dimensionen? Als Attribute?

Spezifische Probleme – Antworten notwendig

- Basis einer Häufigkeitsangabe einer Mutation
- Verschiedene Mutationen an derselben Position
- Mutationen mit Ausdehnung und Range-Queries
- Zugrundeliegendes Referenzgenom
- Präaggregation zur Performanzoptimierung
- Integration der fünften, sechsten, ... Quelle

Beispiel: Reference Variant Store



Data sources

Study	Variant sites	Variants	Unique to study	Variants passed	Samples
1000 Genomes [1]	81,195,126	81,693,252	57,400,612	all	2,504
ESP6500 [2]	1,982,177	1,998,204	184,225	all	6,503
UK10K [44] ALSPAC/TWINS	37,258,978	37,560,436	6,155,493	all	2,432
UK10K with disease [‡]	9,391,582	11,177,227	8,847,466	9,969,036	4,888
TCGA [4] germline [‡]	200,691,728	219,533,884	90,884,769	n/a	4,224
TCGA somatic	876,970	890,172	696,754	all	4,205
Scripps Wellderly [45]	76,144,271	91,947,469	63,331,143	53,303,437	534
ExAC [†] [3]	9,579,712	10,450,724	6,581,946	8,811,372	63,352
MSSM BioBank genotyping	849,806	849,806	0	all	11,210
In-house resequencing study	29,326,393	29,671,729	10,134,258	23,610,572	142
Total observed	358,152,122	399,404,510	244,216,666	>217,796,115	82,558[†]
Other resources:					
dbNSFP* [20]	30,523,109	89,617,785	73,561,239	—	—
ClinVar [12]	101,317	104,455	31,694	—	—
OMIM [46]	10,863	10,913	—	—	—
COSMIC [47]	1,483,983	1,525,243	—	—	—
PharmGKB [‡] [48]	672	684	—	—	—
SwissVar [#]	(77,047)	(84,649)	(34,198)	—	—
HGMD [‡] [13]	125,744	133,464	32,178	—	—
Literature mining	—	890,665	—	—	—
Total observed + other	388,902,292	472,965,749	317,841,777	>217,796,115	82,558