Informationsintegration

Anwendungsszenarien

Ulf Leser

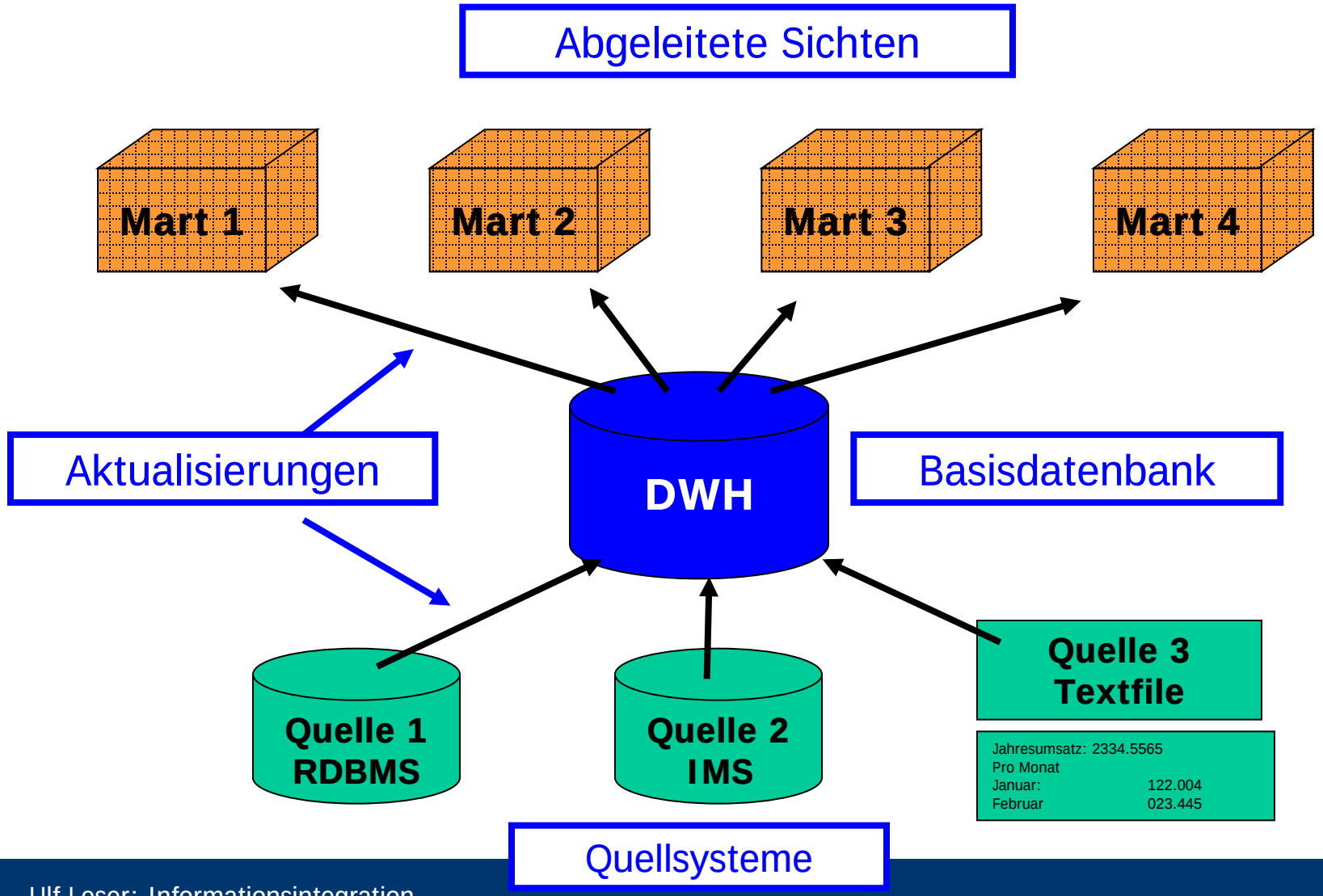
Inhalt dieser Vorlesung

- Zwei Anwendungsszenarien
  - Data Warehouse
  - Föderierte Datenbanken
- Abgrenzung
- Daten versus Schema

Data Warehouse (siehe eigene Vorlesung)

- Integrierte Datenbank innerhalb eines Unternehmens
- Ziel: Zugriff auf alle Informationen, um strategische Entscheidungen zu unterstützen
- Charakteristisch: Aufbau einer integrierten Datenbasis
  - Daten werden aus Quellen kopiert
  - Zur Anfragezeit wird nur das Data Warehouse benutzt
  - Problem: Aktualisierung
- Homogenisierung beim Import
  - Einheiten, Attributnamen, Formate, ...
- Integrationsprozess im DWH heißt ETL
  - Extraction, Transformation, Load

Übersicht


Abgeleitete Sichten
Mart 1
Mart 2
Mart 3
Mart 4
Aktualisierungen
DWH
Basisdatenbank
Quelle 1
RDBMS
Quelle 2
IMS
Quelle 3
Textfile
Jahresumsatz: 2334.5565
Pro Monat
Januar: 122.004
Februar 023.445
Quellsysteme

Frage eines Biologen

Finde alle menschlichen Sequenzen, die zu mindestens 60% identisch sind mit „channel“ Proteinen, die im Gewebe des zentralen Nervensystems der Maus exprimiert werden

Quelle: *A Practitioner's Guide to Data Management and Data Integration in Bioinformatics,* Barbara A. Eckman in Bioinformatics by Zoe Lacroix and Terence Critchlow, 2003, Morgan Kaufmann.

Verschiedene Informationsquellen

- Beteiligte Informationsquellen
  - Mouse Genome Database (MGD) @ Jackson Labs
  - SwissProt @ EBI
  - BLAST tool @ NCBI
  - GenBank nucleotide sequence database @ NCBI

- Alle Quellen sind frei verfügbar

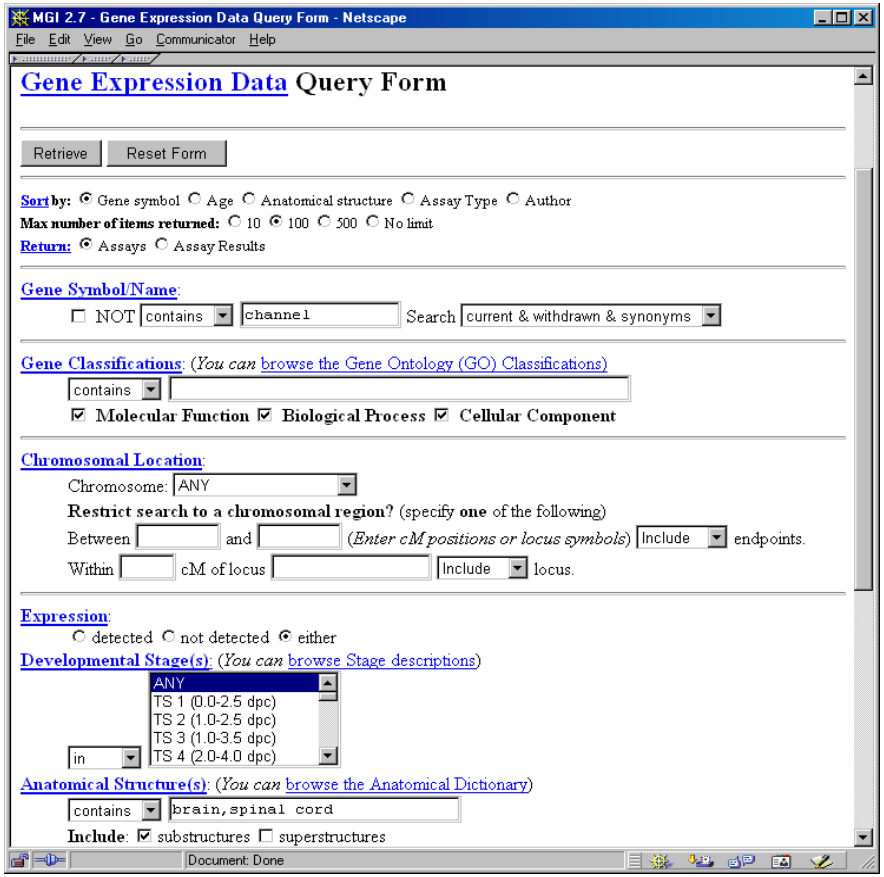
Zusammenhang

Finde alle menschlichen Sequenzen, die zu mindestens 60% identisch sind mit „channel“ Proteinen, die im Gewebe des zentralen Nervensystems der Maus exprimiert werden

- MGD speichert Aktivität und Sequenz von Mausgenen
  - Was sind „channel“ Proteine? Wo werden die exprimiert?
- Swiss-Prot speichert die Proteinsequenz eines Gens
  - Proteinsequenz zu exprimierten Genen
- BLAST findet ähnliche Proteinsequenzen
  - Suche nach ähnlichen Sequenzen
- Genbank enthält DNA-Sequenz von Proteinen
  - Spezies zu einem Protein? DNA-Sequenz zum Protein
- Keine Datenbank enthält alles

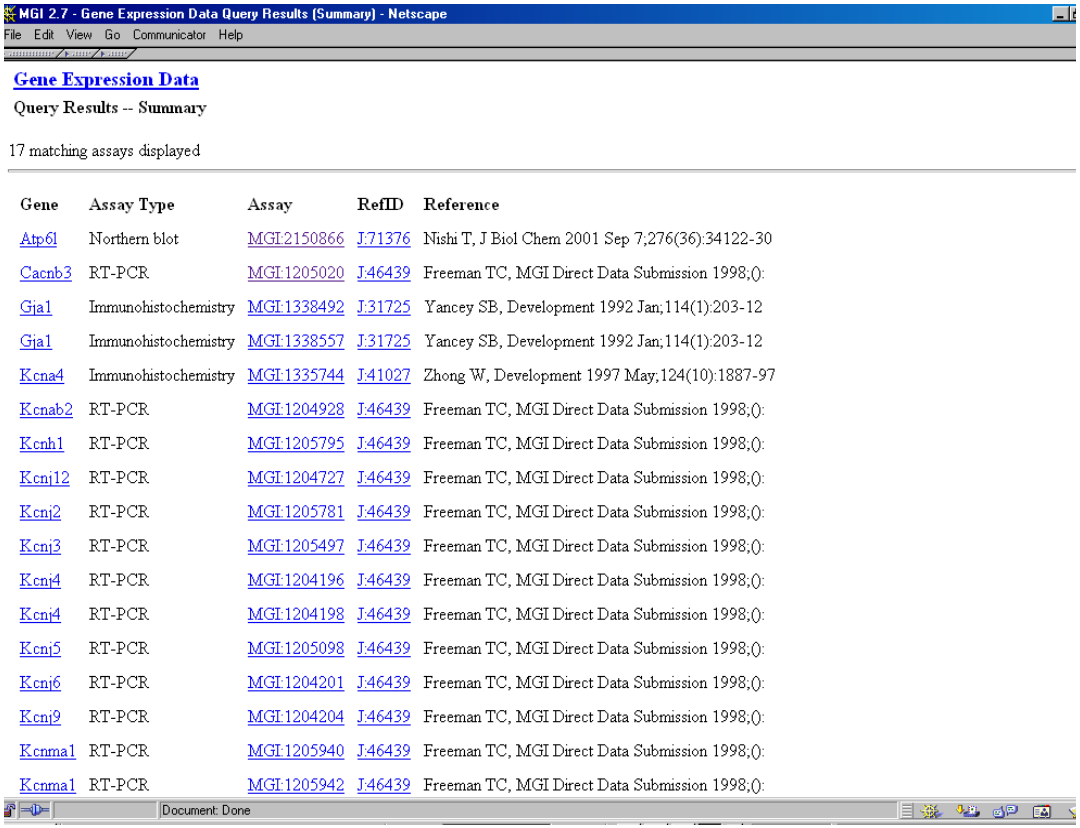
Herkömmlicher Ansatz: Browsing

1. Suche im zentralen Nervensystem aktive „channel“ Gene und deren Sequenzen in der MGD mittels HTML Formular



Herkömmlicher Ansatz: Browsing

- MGD Resultat
  - 14 Gene aus 17 Experimenten



MGI 2.7 - Gene Expression Data Query Results (Summary) - Netscape

File Edit View Go Communicator Help

Gene Expression Data

Query Results -- Summary

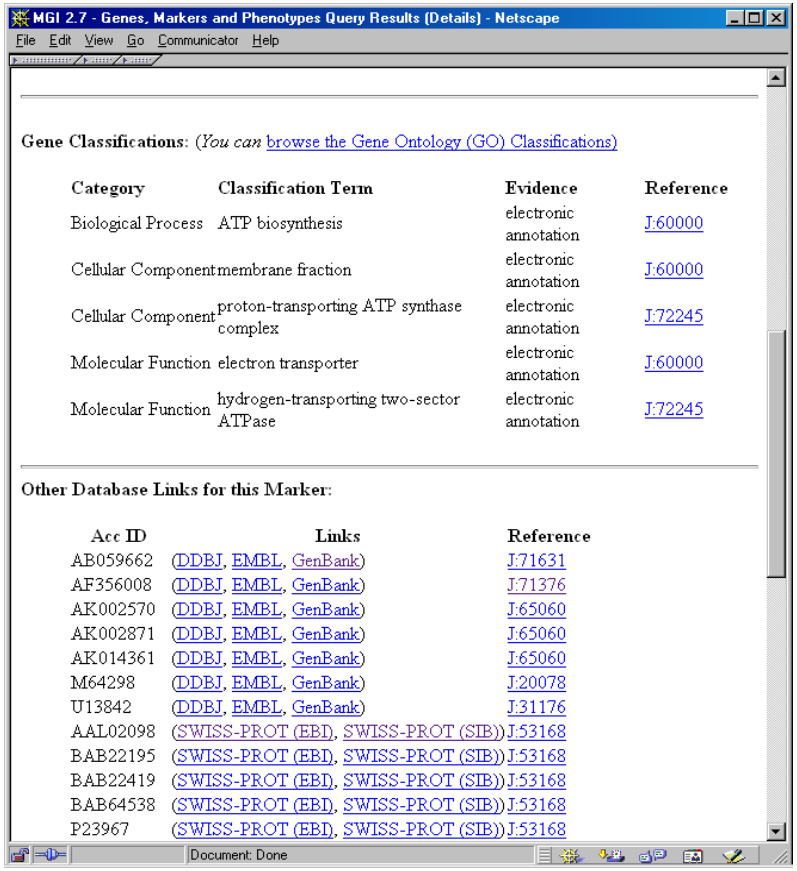
17 matching assays displayed

Gene	Assay Type	Assay	RefID	Reference
Atp6l	Northern blot	MGI:2150866	J:71376	Nishi T, J Biol Chem 2001 Sep 7;276(36):34122-30
Cacnb3	RT-PCR	MGI:1205020	J:46439	Freeman TC, MGI Direct Data Submission 1998;():
Gja1	Immunohistochemistry	MGI:1338492	J:31725	Yancey SB, Development 1992 Jan;114(1):203-12
Gja1	Immunohistochemistry	MGI:1338557	J:31725	Yancey SB, Development 1992 Jan;114(1):203-12
Kcna4	Immunohistochemistry	MGI:1335744	J:41027	Zhong W, Development 1997 May;124(10):1887-97
Kcnab2	RT-PCR	MGI:1204928	J:46439	Freeman TC, MGI Direct Data Submission 1998;():
Kcnh1	RT-PCR	MGI:1205795	J:46439	Freeman TC, MGI Direct Data Submission 1998;():
Kcnj12	RT-PCR	MGI:1204727	J:46439	Freeman TC, MGI Direct Data Submission 1998;():
Kcnj2	RT-PCR	MGI:1205781	J:46439	Freeman TC, MGI Direct Data Submission 1998;():
Kcnj3	RT-PCR	MGI:1205497	J:46439	Freeman TC, MGI Direct Data Submission 1998;():
Kcnj4	RT-PCR	MGI:1204196	J:46439	Freeman TC, MGI Direct Data Submission 1998;():
Kcnj4	RT-PCR	MGI:1204198	J:46439	Freeman TC, MGI Direct Data Submission 1998;():
Kcnj5	RT-PCR	MGI:1205098	J:46439	Freeman TC, MGI Direct Data Submission 1998;():
Kcnj6	RT-PCR	MGI:1204201	J:46439	Freeman TC, MGI Direct Data Submission 1998;():
Kcnj9	RT-PCR	MGI:1204204	J:46439	Freeman TC, MGI Direct Data Submission 1998;():
Kcnma1	RT-PCR	MGI:1205940	J:46439	Freeman TC, MGI Direct Data Submission 1998;():
Kcnma1	RT-PCR	MGI:1205942	J:46439	Freeman TC, MGI Direct Data Submission 1998;():

Document: Done

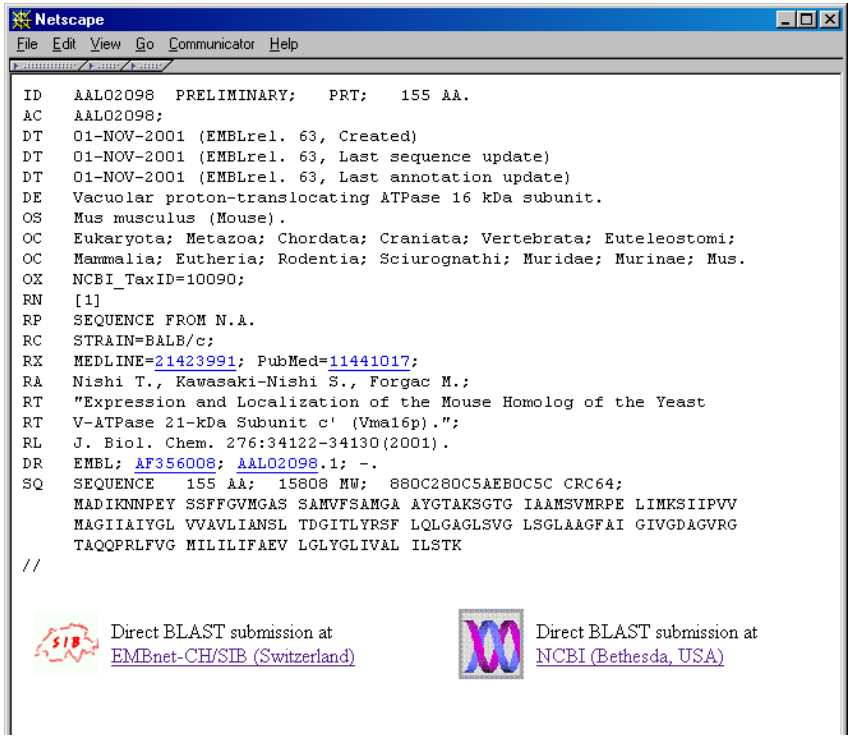
Herkömmlicher Ansatz: Browsing

- In MGD Details zu jedem der 14 Gene ansehen
- Durchschnittlich fünf SwissProt Links pro Gen



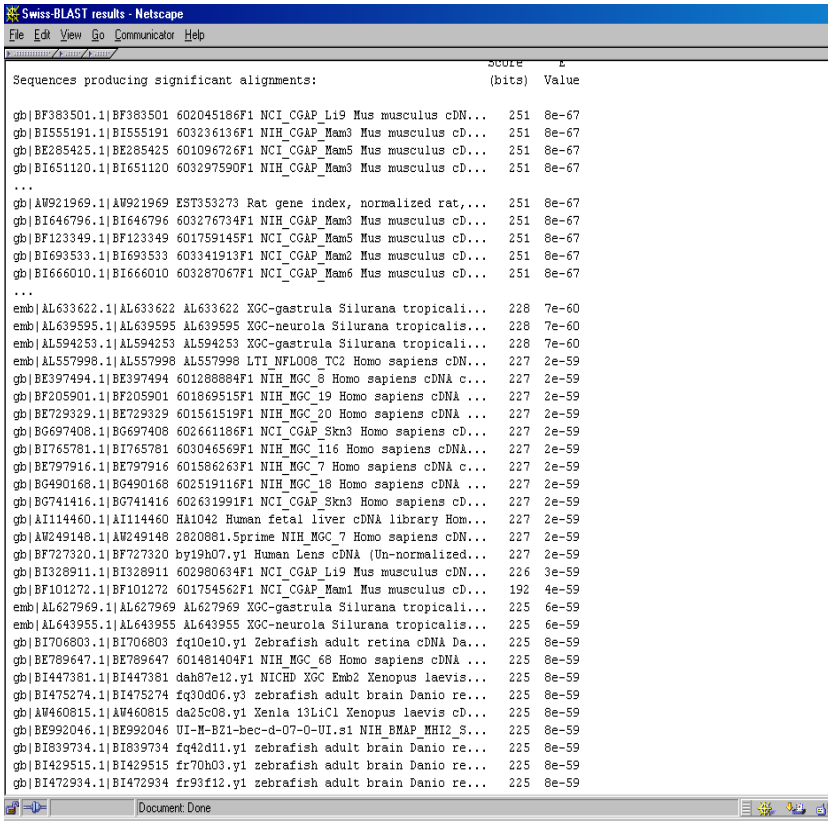
Herkömmlicher Ansatz: Browsing

- Betrachtung jedes SwissProt Eintrags
- Durch Klick BLAST anwerfen



Herkömmlicher Ansatz: Browsing

- Betrachtung jedes BLAST Resultats
  - nicht-menschliche Treffer eliminieren,
  - 60% Identität sicherstellen



Swiss-BLAST results - Netscape

File Edit View Go Communicator Help

Sequences producing significant alignments:	Score (bits)	E Value
gb|BF383501.1|BF383501 602045186F1 NCI_CGAP_Li9 Mus musculus cDN...	251	8e-67
gb|BI555191.1|BI555191 603236136F1 NIH_CGAP_Mam3 Mus musculus cD...	251	8e-67
gb|BE285425.1|BE285425 601096726F1 NCI_CGAP_Mam5 Mus musculus cD...	251	8e-67
gb|BI651120.1|BI651120 603297590F1 NIH_CGAP_Mam3 Mus musculus cD...	251	8e-67
...		
gb|AW921969.1|AW921969 EST353273 Rat gene index, normalized rat,...	251	8e-67
gb|BI646796.1|BI646796 603276734F1 NIH_CGAP_Mam3 Mus musculus cD...	251	8e-67
gb|BF123349.1|BF123349 601759145F1 NCI_CGAP_Mam5 Mus musculus cD...	251	8e-67
gb|BI693533.1|BI693533 603341913F1 NCI_CGAP_Mam2 Mus musculus cD...	251	8e-67
gb|BI666010.1|BI666010 603287067F1 NCI_CGAP_Mam6 Mus musculus cD...	251	8e-67
...		
emb|AL633622.1|AL633622 AL633622 XGC-gastrula Silurana tropicali...	228	7e-60
emb|AL639595.1|AL639595 AL639595 XGC-neurola Silurana tropicalis...	228	7e-60
emb|AL594253.1|AL594253 AL594253 XGC-gastrula Silurana tropicali...	228	7e-60
emb|AL557998.1|AL557998 AL557998 LTI_NFL008_TC2 Homo sapiens cDN...	227	2e-59
gb|BE397494.1|BE397494 601288884F1 NIH_MGC_8 Homo sapiens cDNA c...	227	2e-59
gb|BF205901.1|BF205901 601869515F1 NIH_MGC_19 Homo sapiens cDNA ...	227	2e-59
gb|BE729329.1|BE729329 601561519F1 NIH_MGC_20 Homo sapiens cDNA ...	227	2e-59
gb|BG697408.1|BG697408 602661186F1 NCI_CGAP_Skn3 Homo sapiens cD...	227	2e-59
gb|BI765781.1|BI765781 603046569F1 NIH_MGC_116 Homo sapiens cDNA...	227	2e-59
gb|BE797916.1|BE797916 601586263F1 NIH_MGC_7 Homo sapiens cDNA c...	227	2e-59
gb|BG490168.1|BG490168 602519116F1 NIH_MGC_18 Homo sapiens cDNA ...	227	2e-59
gb|BG741416.1|BG741416 602631991F1 NCI_CGAP_Skn3 Homo sapiens cD...	227	2e-59
gb|AI114460.1|AI114460 HA1042 Human fetal liver cDNA library Hom...	227	2e-59
gb|AW249148.1|AW249148 2820881.5prime NIH_MGC_7 Homo sapiens cDN...	227	2e-59
gb|BF727320.1|BF727320 by19h07.y1 Human Lens cDNA (Un-normalized...	227	2e-59
gb|BI328911.1|BI328911 602980634F1 NCI_CGAP_Li9 Mus musculus cDN...	226	3e-59
gb|BF101272.1|BF101272 601754562F1 NCI_CGAP_Mam1 Mus musculus cD...	192	4e-59
emb|AL627969.1|AL627969 AL627969 XGC-gastrula Silurana tropicali...	225	6e-59
emb|AL643955.1|AL643955 AL643955 XGC-neurola Silurana tropicalis...	225	6e-59
gb|BI706803.1|BI706803 fq10e10.y1 Zebrafish adult retina cDNA Da...	225	8e-59
gb|BE789647.1|BE789647 601481404F1 NIH_MGC_68 Homo sapiens cDNA ...	225	8e-59
gb|BI447381.1|BI447381 dah87e12.y1 NICHD XGC Emb2 Xenopus laevis...	225	8e-59
gb|BI475274.1|BI475274 fq30d06.y3 zebrafish adult brain Danio re...	225	8e-59
gb|AW460815.1|AW460815 da25c08.y1 Xenla 13LiCl Xenopus laevis cD...	225	8e-59
gb|BE992046.1|BE992046 UI-M-BZ1-bec-d-07-0-UI.s1 NIH_BMAP_MHI2_S...	225	8e-59
gb|BI839734.1|BI839734 fq42d11.y1 zebrafish adult brain Danio re...	225	8e-59
gb|BI429515.1|BI429515 fr70h03.y1 zebrafish adult brain Danio re...	225	8e-59
gb|BI472934.1|BI472934 fr93f12.y1 zebrafish adult brain Danio re...	225	8e-59

Document: Done

Herkömmlicher Ansatz: Browsing

- Für jede verbleibende Sequenz
  - Komplette Proteinsequenz bei GenBank holen
- Wir haben oft geklickt
  - 1 + (Suche)
  - 14 + (Gendetails)
  - 14*5 + (SwissProt)
  - 14*5 + (Blast)
  - 14*5*X (Genbank)

Föderierter DBMS Ansatz

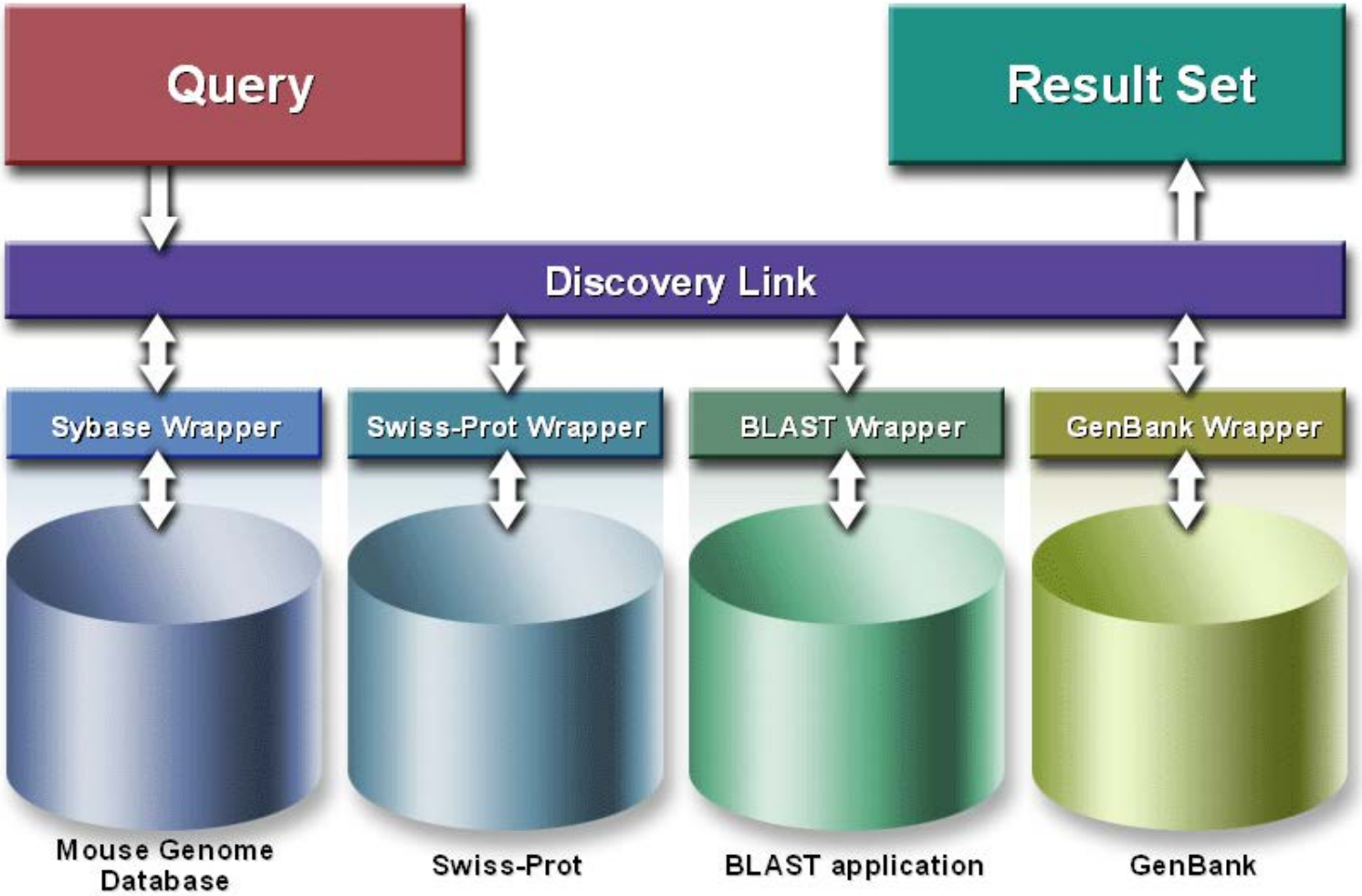
Finde alle menschlichen Sequenzen, die zu mindestens 60% identisch sind mit „channel“ Proteinen, die im Gewebe des zentralen Nervensystems der Maus exprimiert werden

- Ziel: Eine SQL-Anfrage

```
SELECT    g.accnum, g.sequence
FROM      genbank g, blast b, swissprot s, mgd m
WHERE     m.exp = “CNS”
AND       m.defn LIKE “%channel%”
AND       m.spid = s.id AND s.seq = b.query
AND       b.hit = g.accnum
AND       b.percentid > 60 AND b.species=“HS”;
```

- ohne alle Daten an einer Stelle zu speichern
  - On-the-fly integration

Beispiel: DiscoveryLink



Vergleich

DWH – Materialisierung	Föderation – Virtuelle Anfragen
Schnelle Anfragen (lokale Ausführung)	Langsame Anfragen (verteilte Ausführung)
Alle Anfragen	Nur beschränkte Anfragen (je nach API der Quellen)
Hoher Speicherbedarf	Kaum Speicherbedarf auf der Integrationsebene
Möglichkeit zur lokalen Änderung	Nur on-the-fly Korrekturen
Gefahr veralteter Daten, Divergenz zwischen Quelle und DWH	Stets aktuellste Daten
Daten müssen komplett verfügbar sein	Daten müssen über API zugreifbar sein
Prozedurale Integration	Deklarative Integration

Virtuelle Integration ist manchmal ein „Muss“

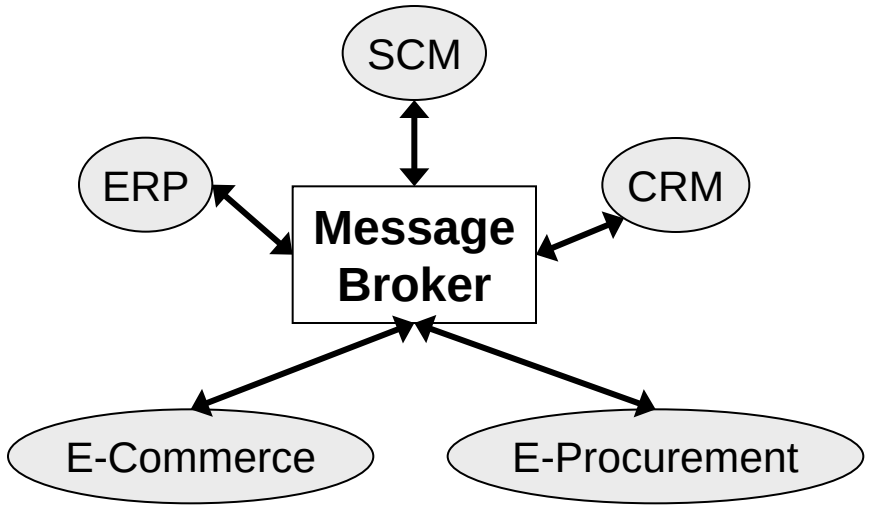
- Datenquellen zu groß (Web)
- Datenquellen nicht als Ganzes zugreifbar
  - Zugriffsbeschränkung, Copyrights, ...
- Inhalt der Datenquellen ändern sich sehr schnell
  - Börsenkurse, Newsticker, Preise, ...

Inhalt dieser Vorlesung

- Zwei Anwendungsszenarien
- Abgrenzung
  - Enterprise Application Integration
  - Objektorientierte Middleware
- Daten versus Schema

Enterprise Application Integration

- „Integration ist ein Produkt, kein Projekt“
- Viele kommerzielle Produkte und Anbieter
- Grundprinzip
  - Geschäftsvorfälle erzeugen Nachrichten
  - Diese werden an einen Message Broker gesendet
  - Der erkennt den Inhalt und wählt interessierte Quellen aus
  - Transformation der Nachrichten
  - Transaktionale Sicherheit („exactly once“)



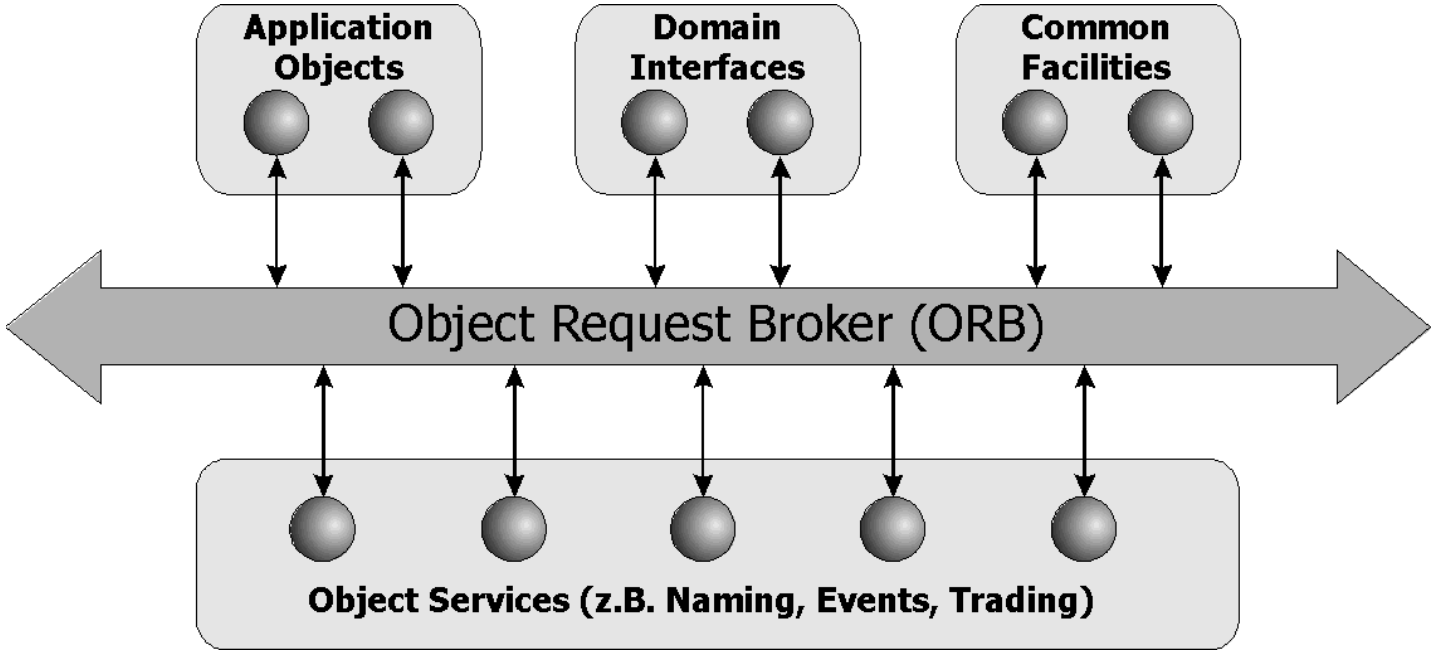
Unterschiede

- Enterprise Application Integration
  - Nachrichtenbasiert, keine Anfragen
  - Informationsverteilung
  - Aktion beim Eintreten eines Ereignisses
  - „Business Process Integration“
- Informationsintegration
  - Anfragebasiert
  - Annahme eines (praktisch) statischen Datenbestands
  - Aktion
    - Erst bei Anfrage (virtuelle Integration)
    - In regelmäßigen Zyklen (materialisierte Integration)
  - Datenbankintegration, Datenintegration

(OO) Middleware

- „Verteilung von Objekten wird vollkommen transparent“
- Viele Ansätze: DCOM, CORBA, OpenView, J2EE, ...
  - Weiterentwicklung von Remote Procedure Calls
- Grundidee
  - Objekte haben weltweit eindeutige ID und Interfaces
  - Clients programmieren gegen das Interface
  - Broker finden aufgerufene Objekte zur Laufzeit anhand ID
    - Marshalling, RPC, unmarshalling
  - Ziel: Plattformunabhängigkeit
    - Sprache, Betriebssystem, Kodierung, Protokoll

Beispiel: CORBA



- Verteilungstransparenz
- Domänenspezifische und generische Dienste
  - Namensauflösung, Nachrichtenübermittlung, Persistenz, ...

Unterschiede

- Middleware
  - Werkzeug für die Programmerstellung
  - Basiert auf Funktionsaufrufen
  - Versendung programmiersprachlicher Objekte
  - Vorwiegend transiente Objekte
  - Fokus auf Methodenaufrufen
- Informationsintegration
  - Ist eine Aufgabe (Projekt), kein Werkzeug
  - Basiert auf Anfragen
  - Arbeiten mit strukturierten Daten
  - Vorwiegend persistente Objekte
  - Fokus auf Objektmanipulation und -suche

Synergie

- Middleware: Plattform zur Entwicklung von integrierten Systemen
  - Lösung der Verteilungsproblematik
  - Auflösung von low-level Formatunterschieden (ASCII versus UniCode, Little Endian versus Big Endian, Datumsformate, ...)
  - Auflösung von Unterschieden im Betriebssystem, Programmiersprache, ...
- Middleware leistet i.A. nicht
  - Umgang mit strukturierten Daten / Schemata
  - Übersetzung von Anfragen
  - Semantische Integration
- EAI wird i.d.R. auf Middleware implementiert

Inhalt dieser Vorlesung

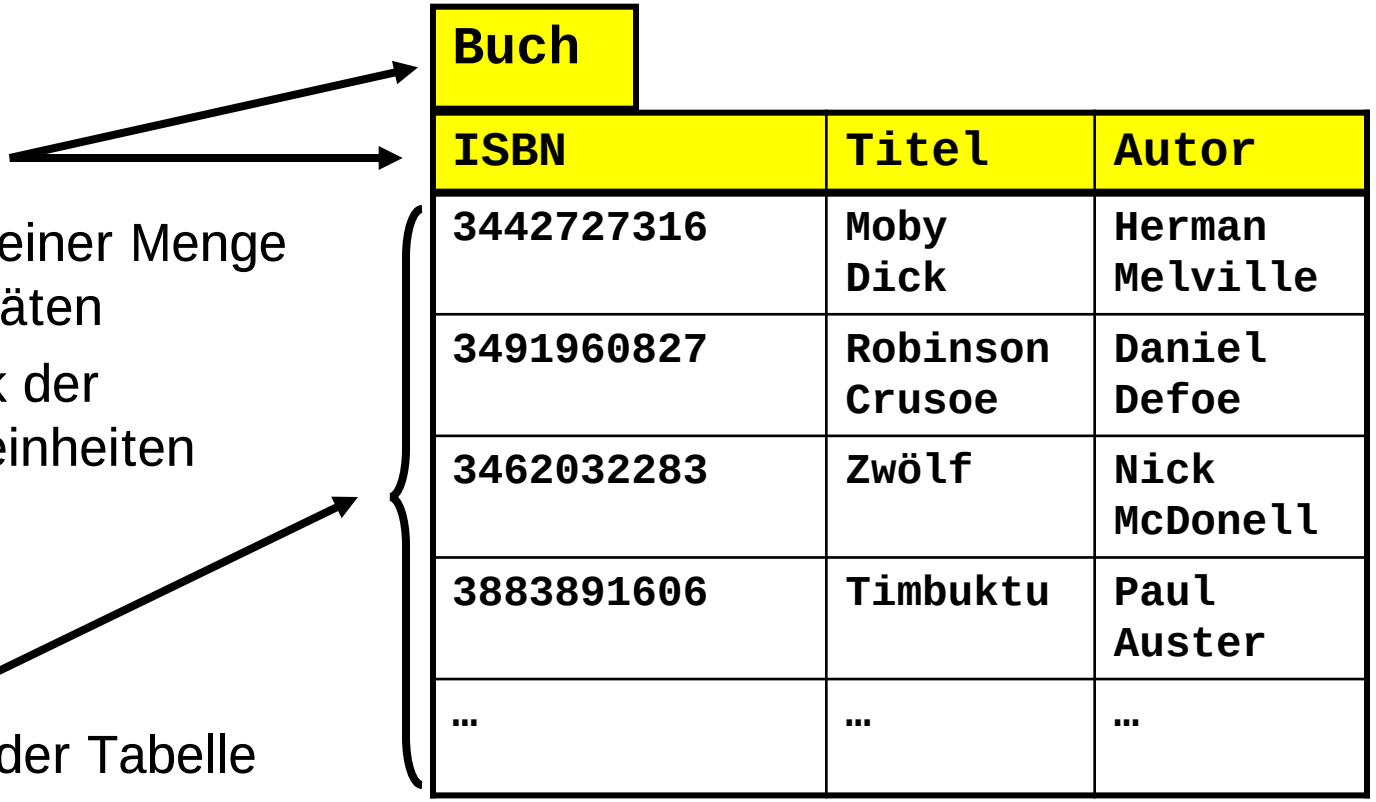
- Zwei Anwendungsszenarien
- Abgrenzung
- Daten versus Schema
  - Extension versus Intension
  - Redundanz versus Komplementierung

Intension & Extension

- Intension
  - Die Intension eines Informationssystems ist die Menge der Schemainformationen und deren Bedeutung (Semantik)
  - Schemata, Metadaten, Modelle, Klassen
- Extension
  - Die Extension eines Informationssystems ist die Menge aller in ihm gespeicherten Daten
  - Instanzen, Tupel, Daten, Objekte, Entitäten
- Die Intension gibt der Extension Struktur
- Informationsintegration muss beides betrachten
  - Intension: Schemaheterogenität
  - Extension: Widersprüchliche Daten

Intension und Extension einer Tabelle

- Intension
  - Struktur einer Menge von Entitäten
  - Semantik der Struktureinheiten
  - Statisch
- Extension
  - Zustand der Tabelle
  - Menge von Entitäten
  - Dynamisch



Buch

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

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

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

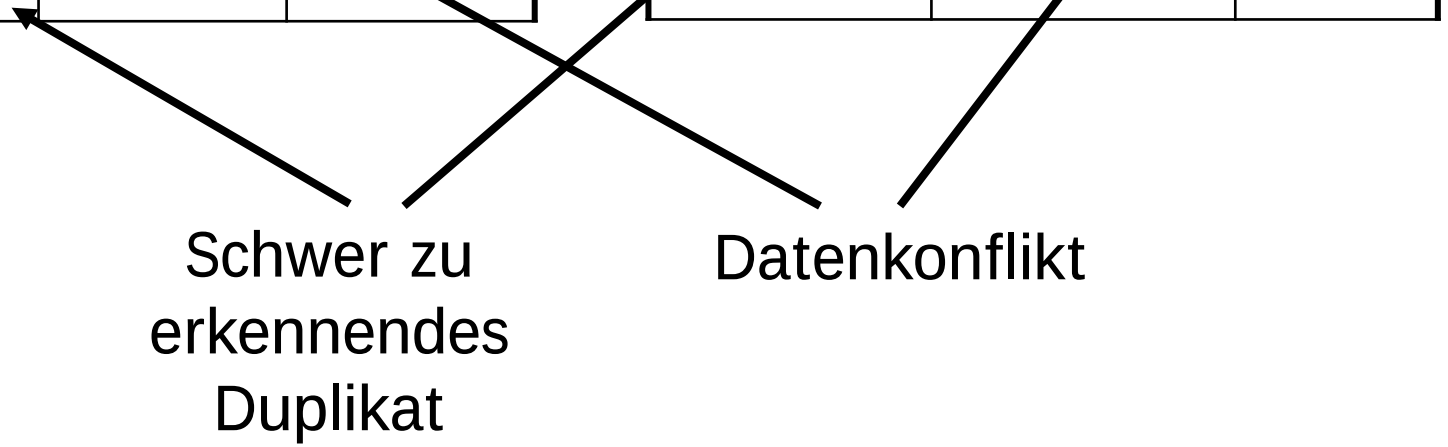
ISBN	Autorname	Year
3491960827	Daniel Defoe	1719
3442727316	H Melville	1851

- 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)

Probleme

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

ISBN	Autorname	Year
3491960827	Daniel Defoe	1719
3442727316	H Melville	1851



Komplementierung (nach Datenfusion)

Intensionale Komplementierung

Extensionale Komplementierung

ISBN	Author	Pages	Year
3442727316	Herman Melville	1056	1851
3491960824	Daniel Defoe	644	1719
3462032283	Nick McDonell	240	
3883891606	Paul Auster	227	
3462026496	Saul Bellow		1992

Zusammenfassung

- Intensionale Redundanz ermöglicht extensionale Komplementierung
  - Zwei Quellen mit teilweise „gleichem“ Schema können zu einer überdeckenderen Datenbasis integriert werden
  - Coverage
- Extensionale Redundanz ermöglicht intensionale Komplementierung
  - Zwei Quellen, die über gleiche Objekte sprechen, können zu einer dichteren Datenbasis integriert werden
  - Density
- Insgesamt ist das Ziel der Integration eine vollständigere Datenbasis (completeness)