



Informationsintegration

Einleitung Teil 2

Ulf Leser

-
- Informationsintegration am Beispiel
 - Quellen
 - Daten versus Schemata

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

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



- Alle Quellen sind frei verfügbar

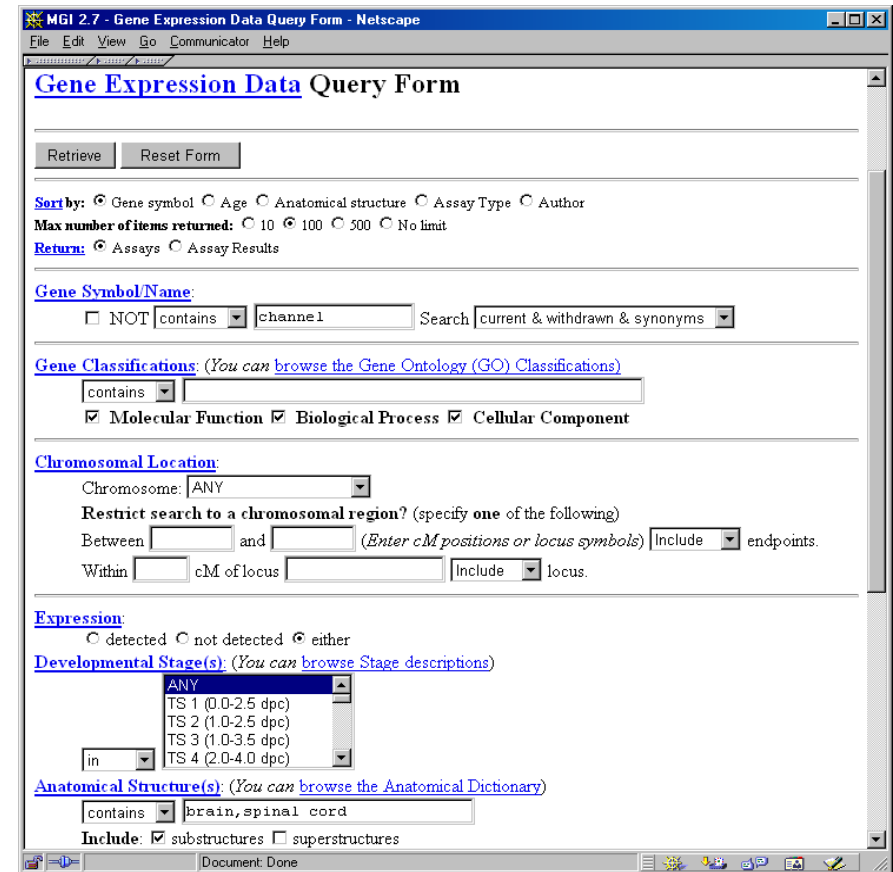
Zusammenhang

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



The screenshot shows the MGI 2.7 Gene Expression Data Query Form in a Netscape browser window. The form is titled "Gene Expression Data Query Form" and includes several sections for searching gene expression data. The "Sort by" section has radio buttons for "Gene symbol", "Age", "Anatomical structure", "Assay Type", and "Author". The "Max number of items returned" section has radio buttons for "10", "100", "500", and "No limit". The "Return" section has radio buttons for "Assays" and "Assay Results". The "Gene Symbol/Name" section has a dropdown menu for "NOT", "contains", and "channel", a text input field with "channel", and a "Search" button. The "Gene Classifications" section has a dropdown menu for "contains" and three checked checkboxes: "Molecular Function", "Biological Process", and "Cellular Component". The "Chromosomal Location" section has a dropdown menu for "Chromosome" set to "ANY", a text input field for "Between" and "Within", and a dropdown menu for "Include" set to "endpoints". The "Expression" section has radio buttons for "detected", "not detected", and "either". The "Developmental Stage(s)" section has a dropdown menu for "ANY" and a list of stages: "TS 1 (0.0-2.5 dpc)", "TS 2 (1.0-2.5 dpc)", "TS 3 (1.0-3.5 dpc)", and "TS 4 (2.0-4.0 dpc)". The "Anatomical Structure(s)" section has a dropdown menu for "contains" and a text input field with "brain,spinal cord". The "Include" section has checked checkboxes for "substructures" and "superstructures".

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

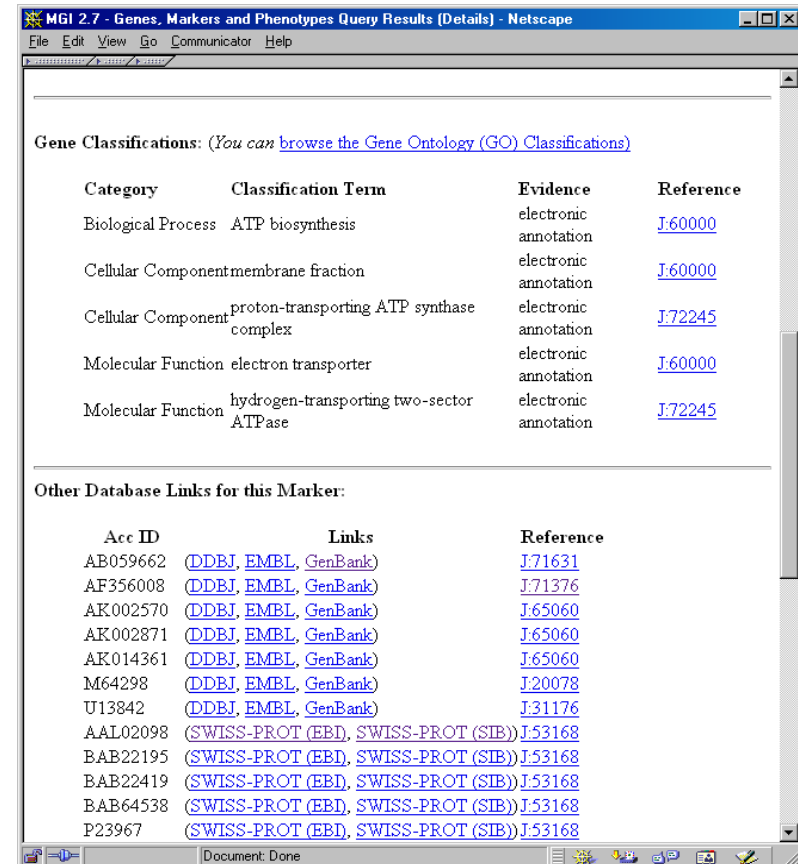
17 matching assays displayed

Gene	Assay Type	Assay	RefID	Reference
Atp6l	Northern blot	MGI2150866	J71376	Nishi T, J Biol Chem 2001 Sep 7;276(36):34122-30
Cacnb3	RT-PCR	MGI1205020	J46439	Freeman TC, MGI Direct Data Submission 1998,0;
Gja1	Immunohistochemistry	MGI1338492	J31725	Yancey SB, Development 1992 Jan;114(1):203-12
Gja1	Immunohistochemistry	MGI1338557	J31725	Yancey SB, Development 1992 Jan;114(1):203-12
Kcna4	Immunohistochemistry	MGI1335744	J41027	Zhong W, Development 1997 May;124(10):1887-97
Kcnab2	RT-PCR	MGI1204928	J46439	Freeman TC, MGI Direct Data Submission 1998,0;
Kcnh1	RT-PCR	MGI1205795	J46439	Freeman TC, MGI Direct Data Submission 1998,0;
Kcni12	RT-PCR	MGI1204727	J46439	Freeman TC, MGI Direct Data Submission 1998,0;
Kcni2	RT-PCR	MGI1205781	J46439	Freeman TC, MGI Direct Data Submission 1998,0;
Kcni3	RT-PCR	MGI1205497	J46439	Freeman TC, MGI Direct Data Submission 1998,0;
Kcni4	RT-PCR	MGI1204196	J46439	Freeman TC, MGI Direct Data Submission 1998,0;
Kcni4	RT-PCR	MGI1204198	J46439	Freeman TC, MGI Direct Data Submission 1998,0;
Kcni5	RT-PCR	MGI1205098	J46439	Freeman TC, MGI Direct Data Submission 1998,0;
Kcni6	RT-PCR	MGI1204201	J46439	Freeman TC, MGI Direct Data Submission 1998,0;
Kcni9	RT-PCR	MGI1204204	J46439	Freeman TC, MGI Direct Data Submission 1998,0;
Kcnma1	RT-PCR	MGI1205940	J46439	Freeman TC, MGI Direct Data Submission 1998,0;
Kcnma1	RT-PCR	MGI1205942	J46439	Freeman TC, MGI Direct Data Submission 1998,0;

Document: Done

Herkömmlicher Ansatz: Browsing

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



Gene Classifications: (You can [browse the Gene Ontology \(GO\) Classifications](#))

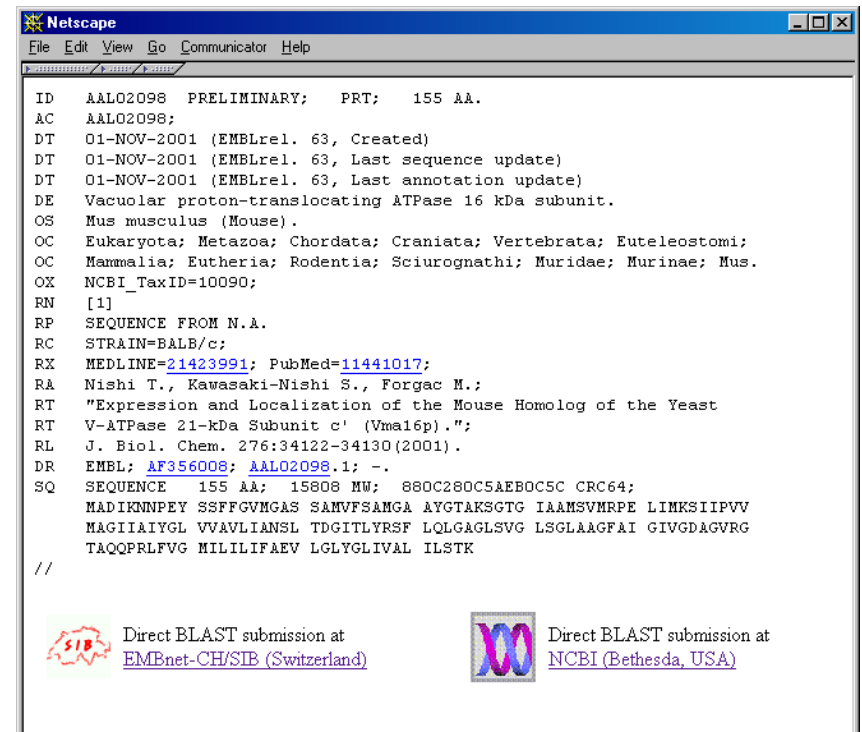
Category	Classification Term	Evidence	Reference
Biological Process	ATP biosynthesis	electronic annotation	J:60000
Cellular Component	membrane fraction	electronic annotation	J:60000
Cellular Component	proton-transporting ATP synthase complex	electronic annotation	J:72245
Molecular Function	electron transporter	electronic annotation	J:60000
Molecular Function	hydrogen-transporting two-sector ATPase	electronic annotation	J:72245

Other Database Links for this Marker:

Acc ID	Links	Reference
AB059662	(DDBJ, EMBL, GenBank)	J:71631
AF356008	(DDBJ, EMBL, GenBank)	J:71376
AK002570	(DDBJ, EMBL, GenBank)	J:65060
AK002871	(DDBJ, EMBL, GenBank)	J:65060
AK014361	(DDBJ, EMBL, GenBank)	J:65060
M64298	(DDBJ, EMBL, GenBank)	J:20078
U13842	(DDBJ, EMBL, GenBank)	J:31176
AAL02098	(SWISS-PROT (EBI), SWISS-PROT (SIB))	J:53168
BAB22195	(SWISS-PROT (EBI), SWISS-PROT (SIB))	J:53168
BAB22419	(SWISS-PROT (EBI), SWISS-PROT (SIB))	J:53168
BAB64538	(SWISS-PROT (EBI), SWISS-PROT (SIB))	J:53168
P23967	(SWISS-PROT (EBI), SWISS-PROT (SIB))	J:53168

Herkömmlicher Ansatz: Browsing

- Betrachtung jedes SwissProt Eintrags
- Durch Klick BLAST anwerfen



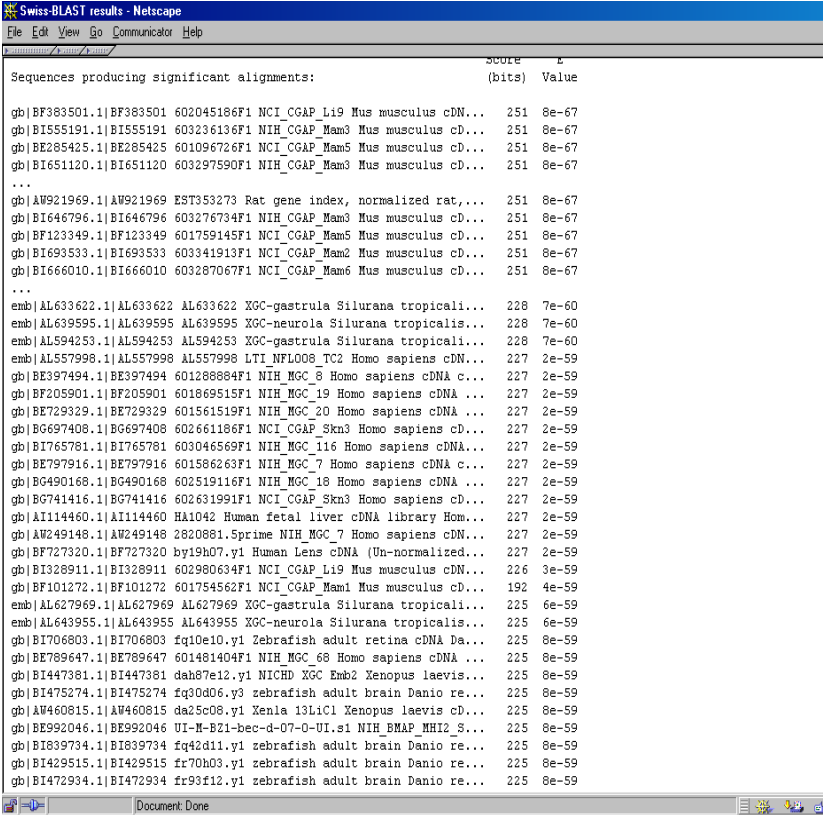
```
ID AAL02098 PRELIMINARY; PRT; 155 AA.
AC AAL02098;
DT 01-NOV-2001 (EMBLrel. 63, Created)
DT 01-NOV-2001 (EMBLrel. 63, Last sequence update)
DT 01-NOV-2001 (EMBLrel. 63, Last annotation update)
DE Vacuolar proton-translocating ATPase 16 kDa subunit.
OS Mus musculus (Mouse).
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX NCBI_TaxID=10090;
RN [1]
RP SEQUENCE FROM N.A.
RC STRAIN=BALB/c;
RX MEDLINE=21423991; PubMed=11441017;
RA Nishi T., Kawasaki-Nishi S., Forgac M.;
RT "Expression and Localization of the Mouse Homolog of the Yeast
RT V-ATPase 21-kDa Subunit c' (Vma16p).";
RL J. Biol. Chem. 276:34122-34130(2001).
DR EMBL; AF356008; AAL02098.1; -.
SQ SEQUENCE 155 AA; 15808 MW; 880C280C5AEBOC5C CRC64;
MADIKNNPEY SSFFGVMGAS SAMVFSAMGA AYGTAQSGTG IAAMSVMRPE LINKSIIPVV
MAGIIAIYGL VVAVLIANSL TDGITLYRSF LQLGAGLSVG LSGLAAGFAI GIVGDAGVRG
TAQQPRLFVG MILILIFAEV LGLYGLIVAL ILSTK
//
```

 Direct BLAST submission at
[EMBNET-CH/SIB \(Switzerland\)](#)

 Direct BLAST submission at
[NCBI \(Bethesda, USA\)](#)

Herkömmlicher Ansatz: Browsing

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



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_Man3 Mus musculus cD...	251	8e-67
gb BE285425.1 BE285425 601096726F1 NCI_CGAP_Man5 Mus musculus cD...	251	8e-67
gb BI651120.1 BI651120 603297590F1 NIH_CGAP_Man3 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_Man3 Mus musculus cD...	251	8e-67
gb BF123349.1 BF123349 601759145F1 NCI_CGAP_Man5 Mus musculus cD...	251	8e-67
gb BI693533.1 BI693533 603341913F1 NCI_CGAP_Man2 Mus musculus cD...	251	8e-67
gb BI666010.1 BI666010 603287067F1 NCI_CGAP_Man6 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.Sprime NIH_MGC_7 Homo sapiens cDN...	227	2e-59
gb BF727320.1 BF727320 by19h07.y1 Human Lens cDNA (Un-normalized...	226	3e-59
gb BI328911.1 BI328911 602980634F1 NCI_CGAP_Li9 Mus musculus cDN...	226	3e-59
gb BF101272.1 BF101272 601754562F1 NCI_CGAP_Man1 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 13LiC1 Xenopus laevis cD...	225	8e-59
gb BE992046.1 BE992046 UI-M-B21-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

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)



Idealvorstellung: Integriertes System

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 b.hit = b.query(s.seq)
AND         b.hit = g.accnum
AND         b.percentid >= 60 AND b.species="HS";
```

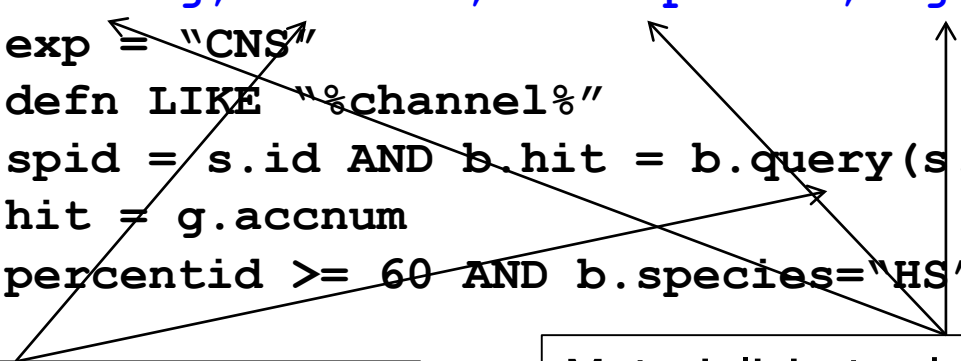


Table Function – Tabelle wird on-the-fly berechnet

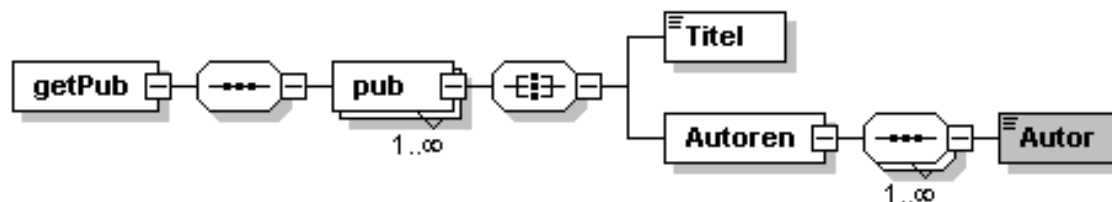
Materialisiert oder virtuell (wenn die **Suchprädikate** ausgewertet werden können)

Aber ...

- Nur theoretisch so einfach
- Tatsächlich sind Daten und Systeme heterogen
- Zweites Beispiel: Schritt für Schritt
 - Zwei Bibliotheken mit [Web Service Schnittstelle](#)
 - Virtuelle Integration
 - Wir suchen Bücher über deren Autornamen
 - Ziel: Nur [eine Anfrage](#), integrierte Ergebnisse

Beispiel – Web Service A

- Web Service A
 - Standort: Trier
 - Operation:
 - `getPubByAuthor( firstName, lastName)`
 - `getPubByTitle( title)`
 - Output-Format



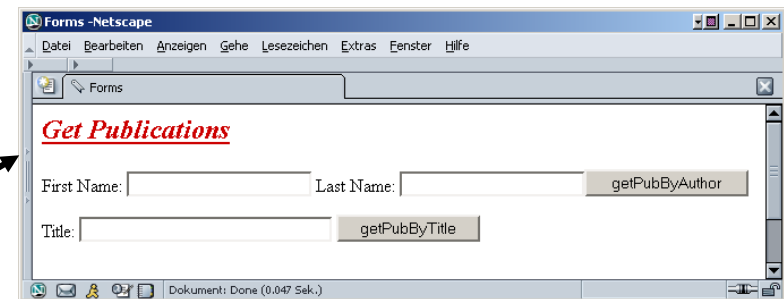
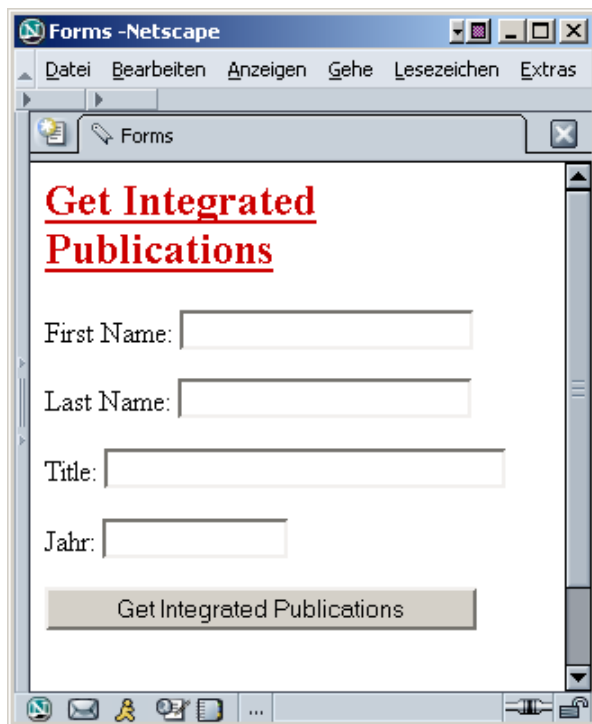
Beispiel – Web Service B

- Web Service B
 - Standort: Humboldt-Universität
 - Operation: `myPubs ( author, year )`
 - Output-Format



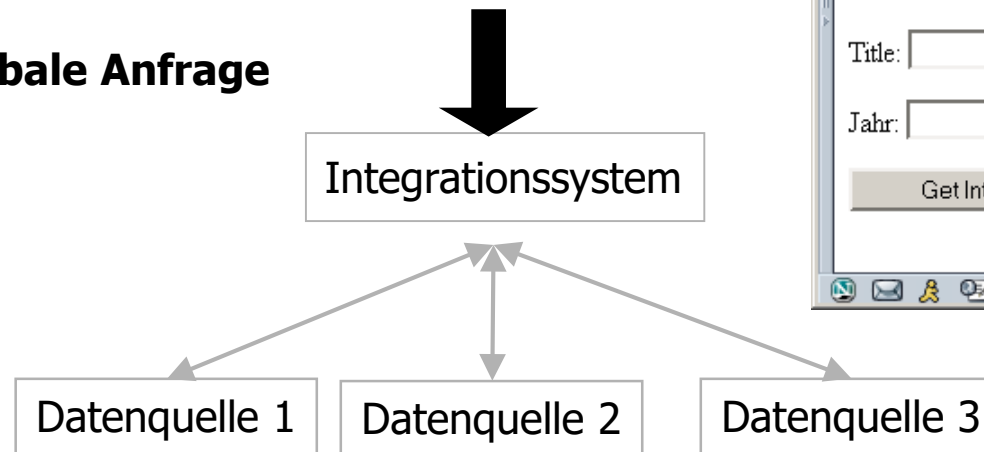
Eine Schnittstelle

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



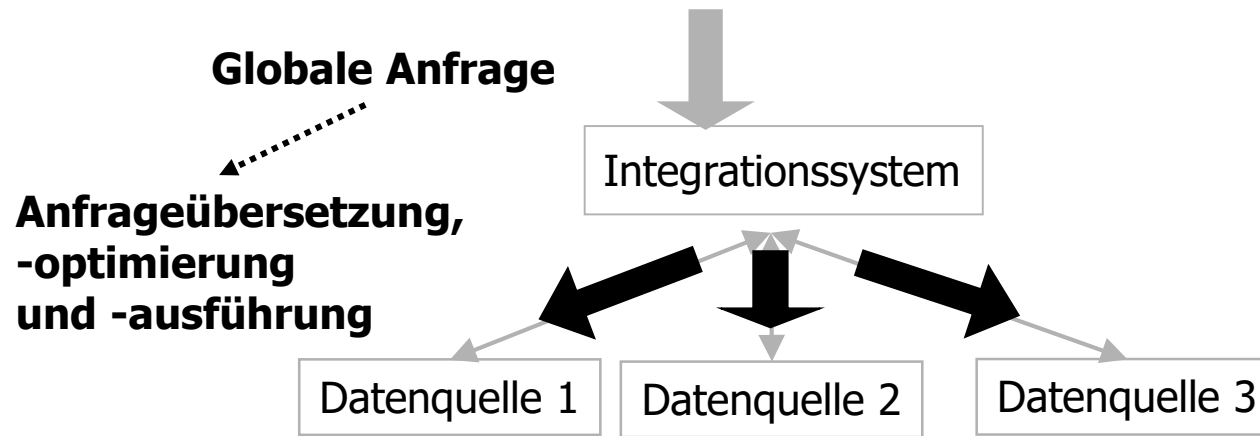
Ablauf 1

Globale Anfrage



The screenshot shows a Netscape browser window with the title "Forms - Netscape". The address bar displays "Forms". The main content area contains a form titled "Get Integrated Publications" in red, underlined text. The form includes four input fields: "First Name:", "Last Name:", "Title:", and "Jahr:". Below these fields is a button labeled "Get Integrated Publications". The browser's menu bar includes "Datei", "Bearbeiten", "Anzeigen", "Gehe", "Lesezeichen", and "Extras". The status bar at the bottom shows various icons, including a globe, a mail icon, a person icon, a document icon, and a lock icon.

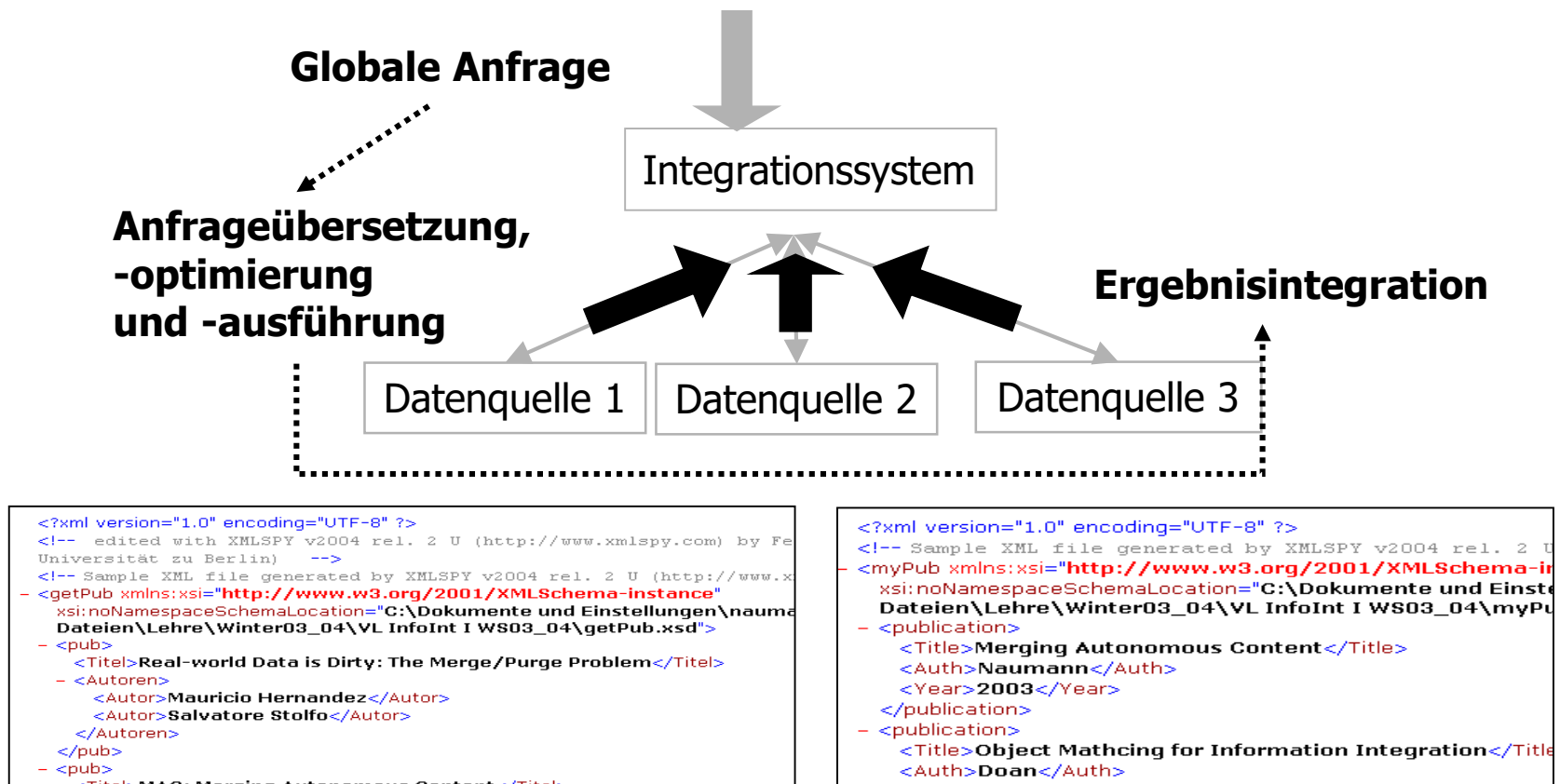
Ablauf 2



The screenshot shows a Netscape browser window with the title 'Forms - Netscape'. The address bar shows 'Forms'. The main content area displays the heading 'Get Publications' in red. Below the heading, there are two input fields: 'First Name:' and 'Last Name:', followed by a button labeled 'getPubByAuthor'. Below these, there is a 'Title:' input field followed by a button labeled 'getPubByTitle'. The status bar at the bottom indicates 'Dokument: Done (0,047 Sek.)'.

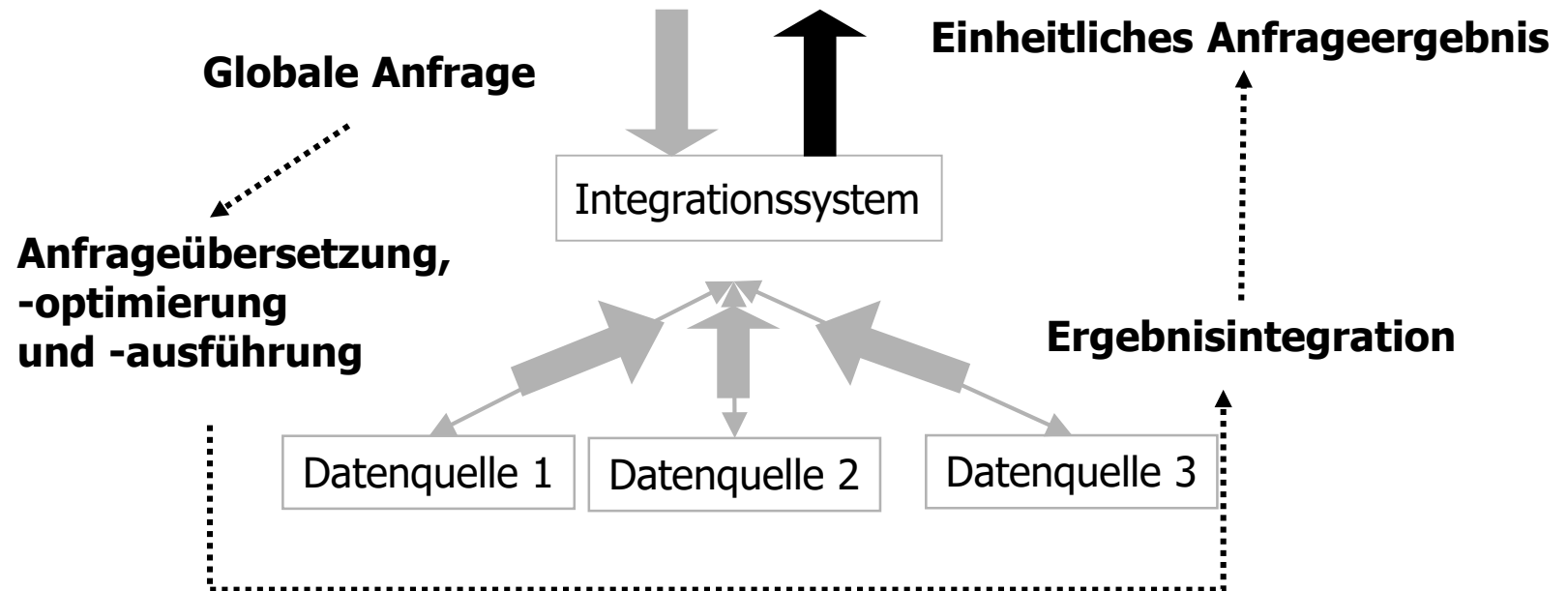
The screenshot shows a Netscape browser window with the title 'Forms - Netscape'. The address bar shows 'Forms'. The main content area displays the heading 'The great myPubs Webservice interface'. Below the heading, there are two input fields: 'Autor:' and 'Jahr:', followed by a button labeled 'myPubs'. The status bar at the bottom indicates 'Dokument: Done (0,046 Sek.)'.

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>
```



Quellenauswahl

- **Geschwindigkeit**

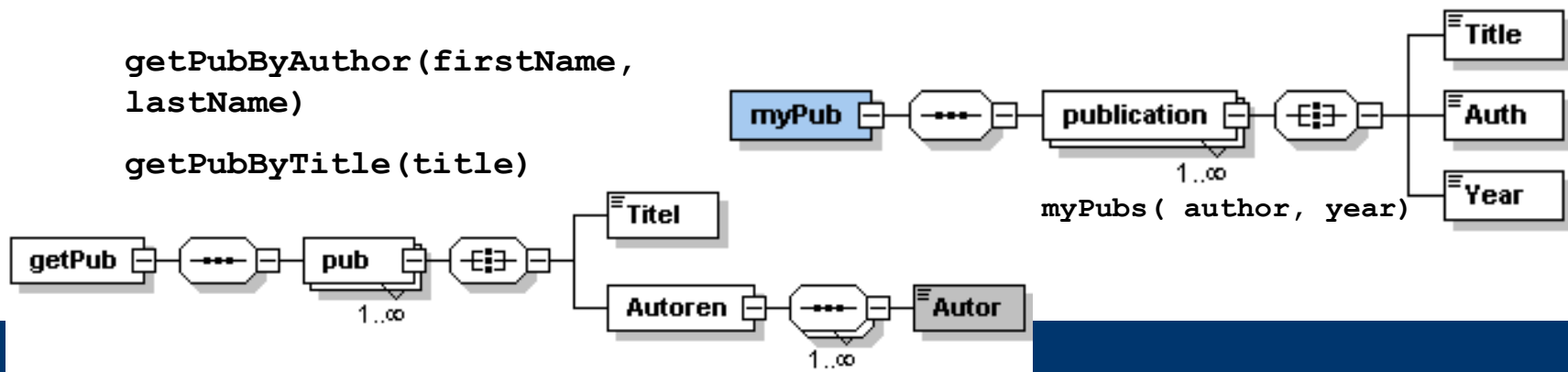
- Welches System ist schneller (bei gleicher Ergebnisgröße)?
- Web Service A in Trier (remote)
- Web Service B in Adlershof (local)

- **Vollständigkeit**

- Web Service A hat weniger Attribute, aber mehr Publikationen
- Web Service B hat mehr Attribute, aber weniger Publikationen
- 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

`getPubByAuthor(firstName,
lastName)`

`getPubByTitle(title)`



Zwei Resultate

Web Service A

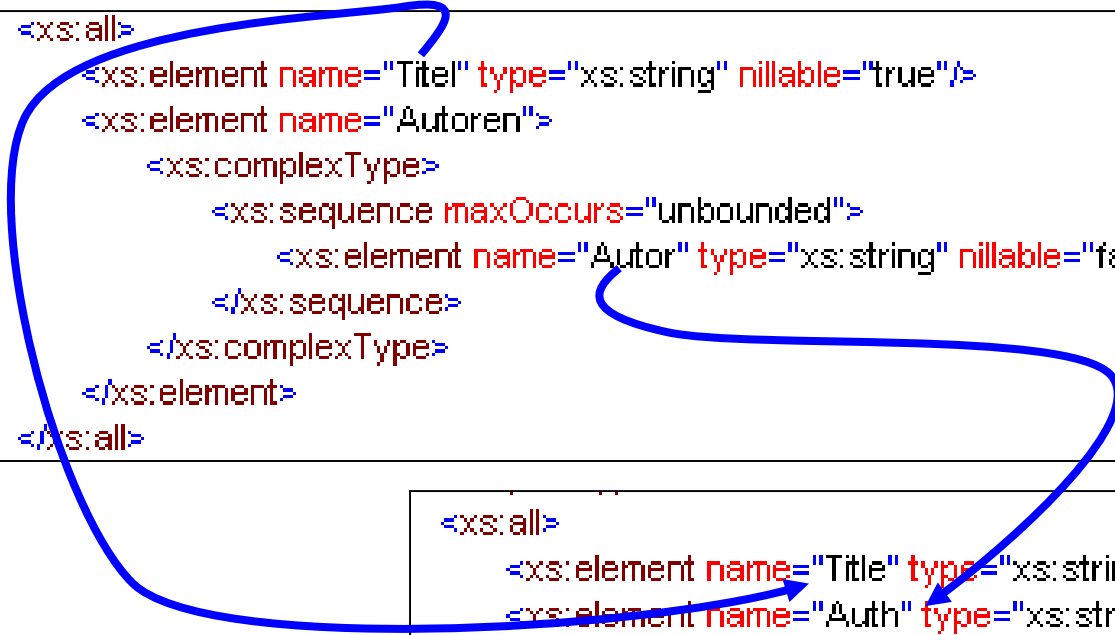
```
<?xml version="1.0" encoding="UTF-8" ?>
<!-- edited with XMLSPY v2004 rel. 2 U (http://www.xmlspy.com) by Felix Naumann
Universität zu Berlin) -->
<!-- Sample XML file generated by XMLSPY v2004 rel. 2 U (http://www.xmlspy.com)
- <getPub xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:noNamespaceSchemaLocation="C:\Dokumente 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

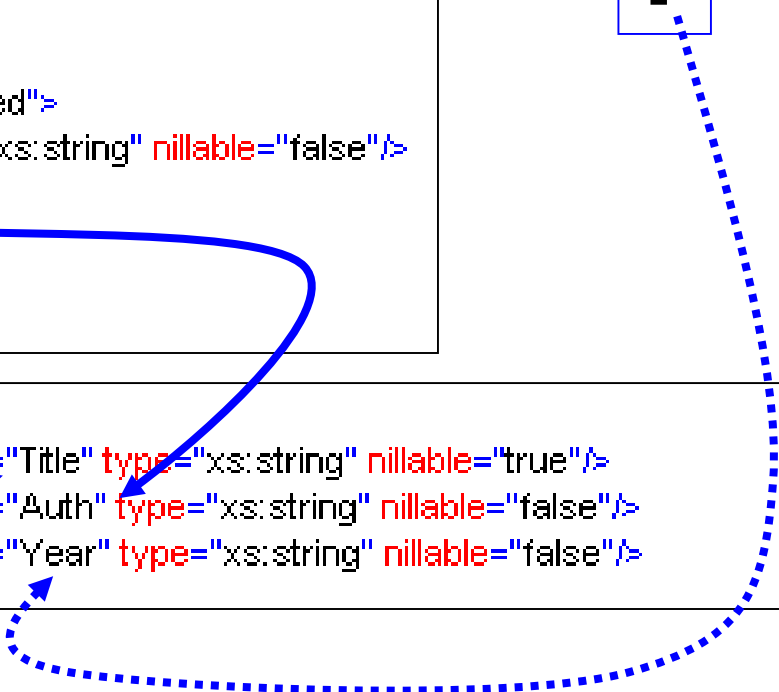
```
<?xml version="1.0" encoding="UTF-8" ?>
<!-- Sample XML file generated by XMLSPY v2004 rel. 2 U (http://www.xmlspy.com)
- <myPub xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:noNamespaceSchemaLocation="C:\Dokumente und Einstellungen\naumann\Eigene
  Dateien\Lehre\Winter03_04\VL InfoInt 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>
```

„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>
```



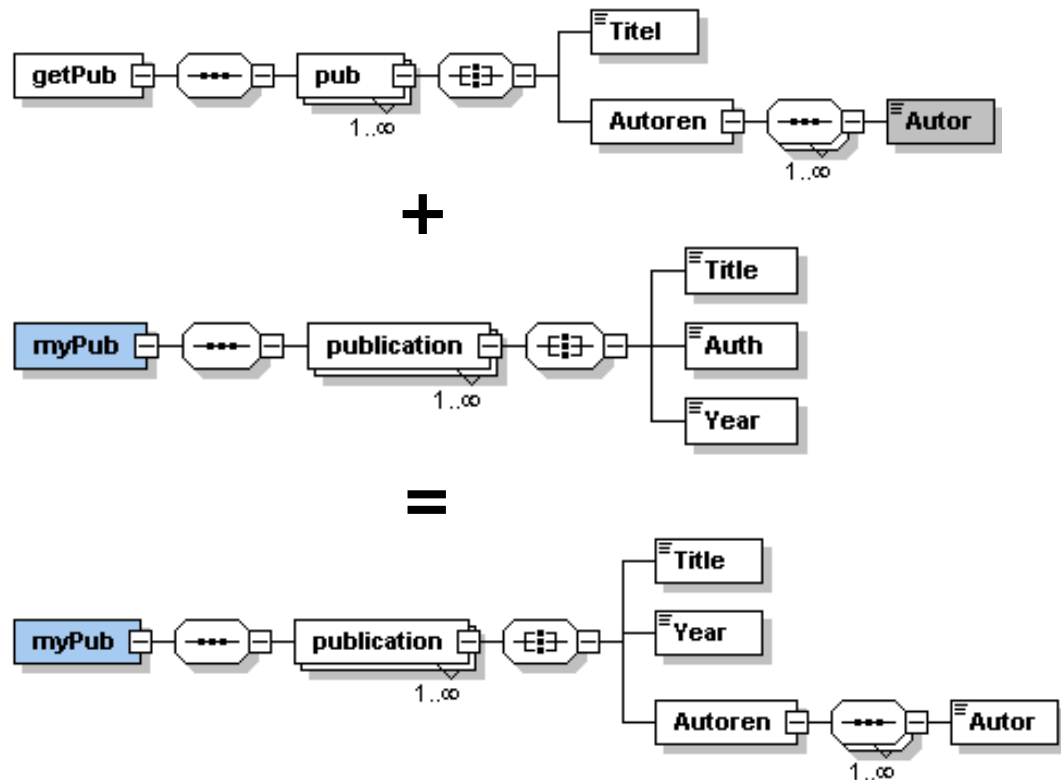
?

„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 Matching for Information Integration</Title>
  <Auth>Doan</Auth>
  <Year>1999</Year>
</publication>
</myPub>
```

Ideales Ergebnis: Schema- und Datenintegration



Integriertes Schema

Integrierte Daten

Inhalt dieser Vorlesung

- Informationsintegration am Beispiel
- Quellen
 - Dateien, CSV, XML, JSON, HTML, Tupel, API, ...
- Daten versus Schemata

Dateien

- Datei
 - Informationseinheit: Zeile / Token
 - Anfrage: Parser
 - Struktur: Flach
- Position in Dateisystemen kann Bedeutung haben
 - Verzeichnisnamen als **Metainformation** (Klassen, Jahre, ...)
- Beispiele
 - Komma-delimited files (CSV)
 - DB/EXCEL Export
 - JAVA .property Dateien
 - Tausende proprietäre Formate

ID	RNGTPCHI	standard; RNA; ROD; 1016 BP.	Molecule type	
XX			Name	
DT	01-AUG-1991	(Rel. 28, Created)	Date of creation and last update	
DT	04-MAR-2000	(Rel. 63, Last updated, Version 2)		
XX				
DE	Rat GTP cyclohydrolase I mRNA, complete cds.		Free text description	
XX				
KW	GTP cyclohydrolase I.		Keywords describing the molecule	
XX				
OS	Rattus norvegicus (Norway rat)		Organism	
OC	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;			
OC	Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Rattus.			
XX				
RN	[1]		Article the sequence was published in	
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	Structural annotation (coding sequence)	
FT		/codon_start=1		
FT		/db_xref="GOA:P22288"	Link to functional annotation of resulting protein	
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="MEKPRGVRCITNGFPERELPRPGASRPAEKSRPPEAKGAQPADAWK	Translated protein sequence	
FT		AGRPRSEEDNELNLPNLAAAYSSILRSLGEDPQRQGLLKTPWRAATAMQFFTKGYQETI		
FT		SDVLNDAIFDEHDHDEMIVVKDIDMFSMCEHHLVPFVGRVHIGYLPNKQVLGLSKLARIV		
FT		EIYSRRLQVQERLTKQIAVAITEALQPAGVGVIIEATHMCMVMRGVQKMNSKTVTSTML		
FT		GVFREDPKTREEFLTLIRS"		
SQ	Sequence 1016 BP; 236 A; 279 C; 291 G; 210 T; 0 other;		Sequence of bases	
	gacttcgaac	ctcattcggt gcagaactcc tgtcccggtg acagccacag gtcacggcgcg		60
	ccggctaagc	cgagccgcag cgcttggttag caccttaggg tgtctcgga gcaatcgcgc		120
	cgggtccatg	gagaagccgc ggggtgtaag gtgcaccaat gggttccccg agcgggagct		180
	...			
	catcaggagc	tgaacttcgc tgtgcgagcc ccggtttgca gacccccgct gaggccagcg		900
	ttatctgtct	cgattgtaca ttccagttcc agttggtata cttgtcaact ttattttctca		960
	ccatgaattg	tatttaataa ttatttatag agatgtcaaa taaaggtgat caactt		1016

Markup-Dateien

- Markup Datei
 - Informationseinheit: **Tagged text**
 - Anfrage: Parser, Anfragesprache
 - Struktur: Meist hierarchisch geschachtelt
 - Semistrukturiert: Mengenwertige Tags, fehlende Tags, neue Tags
- XML, HTML, LaTeX, JSON, ...
- Einsatzgebiete heute
 - Webseiten
 - Datendateien
 - Web Services
 - Allg. **Nachrichtenaustausch**

```
<?xml version="1.0" encoding="utf-8"?>
<Title>
  <LocalTitle>22 Bullets</LocalTitle>
  <OriginalTitle>L'immortel</OriginalTitle>
  <SortTitle>22 Bullets</SortTitle>
  <Added>16.9.2010 21:15:02</Added>
  <ProductionYear>2010</ProductionYear>
  <RunningTime>117</RunningTime>
  <IMDBrating>6,6</IMDBrating>
  <MPAARating>NR</MPAARating>
  <Description>Charly Mattei has
...</Description>
  <Type>XviD</Type>
  <AspectRatio />
  <LockData>False</LockData>
  <IMDB>tt1167638</IMDB>
  <TMDBId>37645</TMDBId>
  <Genres>
    <Genre>Action</Genre>
    <Genre>Crime</Genre>
    <Genre>Thriller</Genre>
  </Genres>
  <Persons>
    <Person>
      <Name>Jean Reno</Name>
      <Type>Actor</Type>
      <Role>Charly Mattei</Role>
    </Person>
    <Person>
```

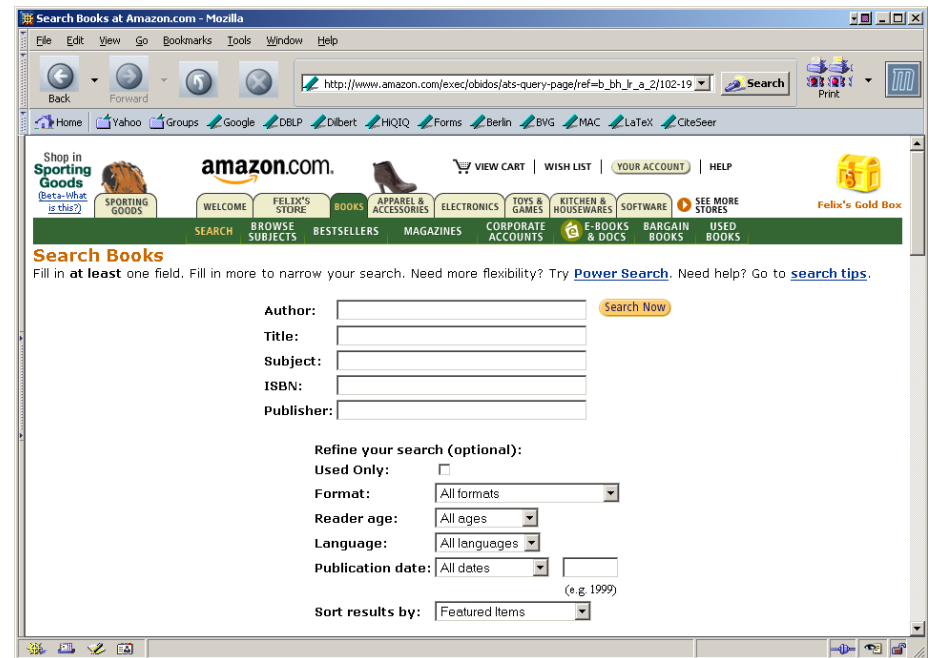
...

Datenbanken

- Strukturierte Datenbank
 - Informationseinheit: **Tupel / Attribut**, Objekt
 - **Anfrage**: (Komplexe) Anfragesprache, z.B. SQL, Xquery
- Vorherrschend: **Relational**
 - Aus der Mode: OO, Hierarchisch
 - In Mode (NoSQL): Key-Value stores, document, graphs
- Einsatzgebiete: Überall

Webformulare

- HTML Formulare, Webseiten
 - Informationseinheit: HTML Seite, Text
 - Anfrage: Suchworte, **Formulare** (Freetext, dropdown-list ...)
 - Verwaltungen verwenden noch oft **Dialogsysteme**
 - Struktur des Ergebnisses: Markup Datei (HTML, XML)
- **Deep Web**
- Einsatzgebiete: Überall



The screenshot shows the Amazon.com search interface in a Mozilla browser window. The address bar displays the URL: http://www.amazon.com/exec/obidos/ats-query-page/ref=bh_lr_a_2/102-19. The page features a navigation bar with links to various categories like Sporting Goods, Books, Electronics, etc. Below the navigation bar, the 'Search Books' section is prominent, with a prompt: 'Fill in at least one field. Fill in more to narrow your search. Need more flexibility? Try [Power Search](#). Need help? Go to [search tips](#).' The search form includes input fields for 'Author:', 'Title:', 'Subject:', 'ISBN:', and 'Publisher:', each followed by a 'Search Now' button. Below these fields, there is a 'Refine your search (optional):' section with several filters: 'Used Only' (checkbox), 'Format' (dropdown menu), 'Reader age' (dropdown menu), 'Language' (dropdown menu), 'Publication date' (dropdown menu with a text input for specific dates), and 'Sort results by' (dropdown menu). The page also includes a 'Shop in Sporting Goods' banner and a 'Felix's Gold Box' promotion.

„Legacy“ Anwendungen

- Anwendungen in „veralteten“ Programmiersprachen
- Beispiel: COBOL
 - Für kaufmännische Anwendungen
 - Operiert auf formatierten Dateien
 - Mischung aus DBMS und Programmiersprache
 - Zugriff über **spezielle Anwendungen** (Reports)
- Schätzungen: **1/2 aller Anwendungen** in Banken/ Versicherungen/ ... sind Legacy Anwendungen

Inhalt dieser Vorlesung

- Informationsintegration am Beispiel
- Quellen
- Daten versus Schemata
 - 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, qualitativ besseren Ganzen integriert
 - Das ist der 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

- Quellen haben Tabellen / Attribute mit **gleicher Bedeutung**
 - Muss nicht auch gleichen Namen bedeuten
- Oftmals schwer zu erkennen (Schema Matching)

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

- Menge der Quellen **repräsentierten Objekte überlappt**
- 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	844

ISBN	Autorname	Year
3491960827	Daniel Defoe	1719
3442727316	H Melville	1851

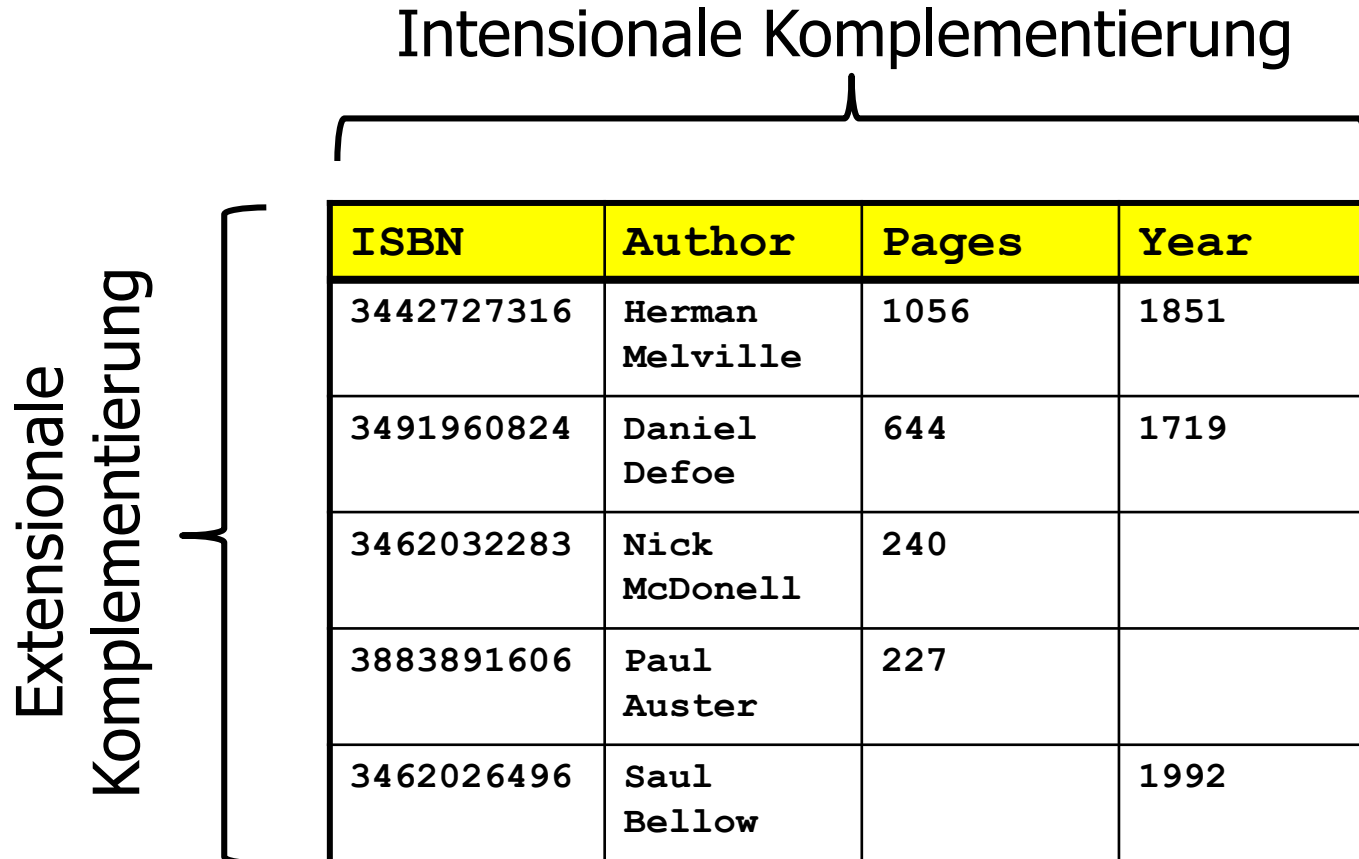
Schwer zu
erkennendes
Duplikat

Datenkonflikt

Komplementierung (nach Integration)

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
- Ziel von Integration ist eine **vollständigere** Datenbasis
 - Completeness