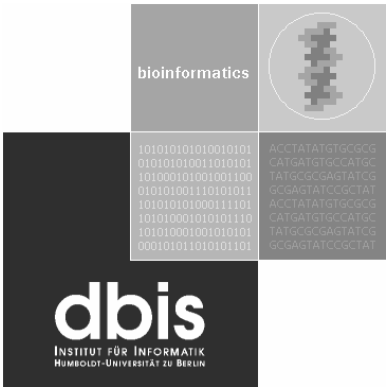
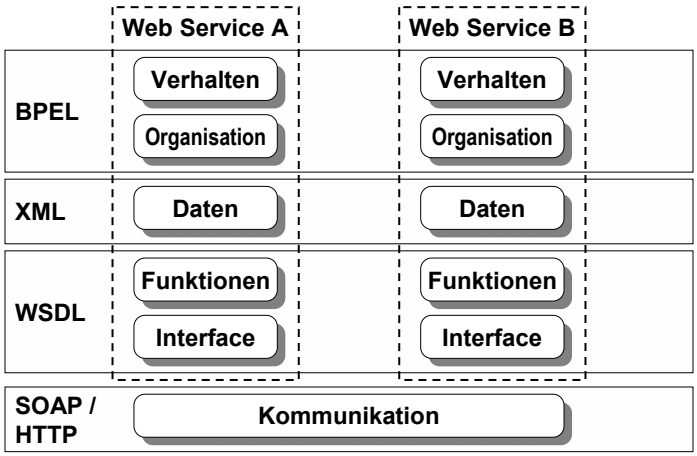
Datenintegration in der Bioinformatik

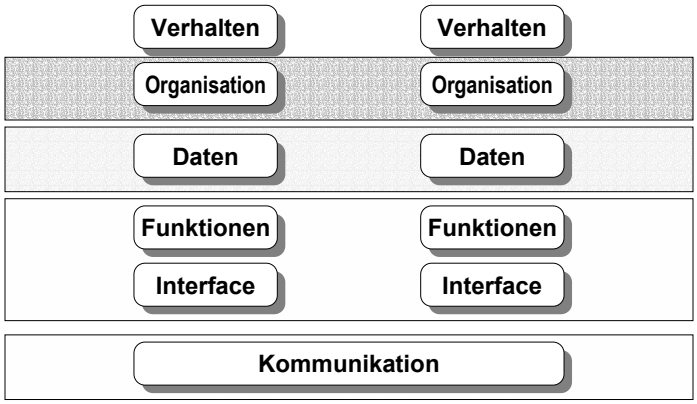


Eine Anleihe bei Prof. Reisig

(Vorangegangene RingVL)



Gegenstand heute



Ursprüngliche Vortragsankündigung:

R. Eckstein, S. Heymann

XML , (Modellierung) & Datenintegration - Fortschritt im Life Science Bereich

(Schwerpunkt "Große Datenräume in Web-basierten Umgebungen")

XML hat auch in der Bioinformatik Einzug gehalten - zum Datenaustausch, aber auch zur Repräsentation der komplexen Informationen. Im ersten Teil werden Aspekte von XML sowie weitergehende Entwicklungen vorgestellt, die für den Life Science Bereich von besonderem Interesse sind. Dazu gehören die konzeptionelle Modellierung von Dokumentschemata sowie für semantische Informationen über die Biodaten. Im zweiten Teil wird ein Überblick für XML-Anwendungen im Life Science Bereich gegeben und ein typisches Anwendungsbeispiel aus dem Forschungsgebiet Biodiversität & Ökologie erläutert: Die Erfassung und Darstellung von Taxonomie-Daten, den Ubergang von klassischen Morphologie-basierten botanischen Schulen zu genbasierter Kladistik. Es wird demonstriert, wie die Konflikte behandelt und die Daten in ein navigierbares Graphenformat transponiert werden. Dabei kommt der GeneViator zum Einsatz.

Herkömmliche Dokumentkonventionen

ACTGATGCGATCGCTAATATCTACGGC

Die Vortragsankündigung hätte man auch *formularbasiert* schreiben können:

Referenten:	R. Eckstein, S. Heymann
Titel:	XML , (Modellierung) & Datenintegration - Fortschritt im Life Science Bereich
Untertitel:	(Schwerpunkt "Große Datenräume in Web-basierten Umgebungen")
Zusammenfassung:	... Die Erfassung und Darstellung von Taxonomie-Daten, den Ubergang von klassischen Morphologie-basierten botanischen Schulen zu genbasierter Kladistik. Es wird ...

//

Herkömmliche Dokumentkonventionen

ACTGATGCGATCGCTAATATCTACGGC

Die Vortragsankündigung hätte man auch *formularbasiert* schreiben können:

Referenten:	R. Eckstein, S. Heymann
Titel:	XML , (Modellierung) & Datenintegration - Fortschritt im Life Science Bereich
Untertitel:	(Schwerpunkt "Große Datenräume in Web-basierten Umgebungen")
Zusammenfassung:	... Die Erfassung und Darstellung von Taxonomie-Daten, den Ubergang von klassischen Morphologie-basierten botanischen Schulen zu genbasierter Kladistik. Es wird ...

//

(Karteikartenprinzip)

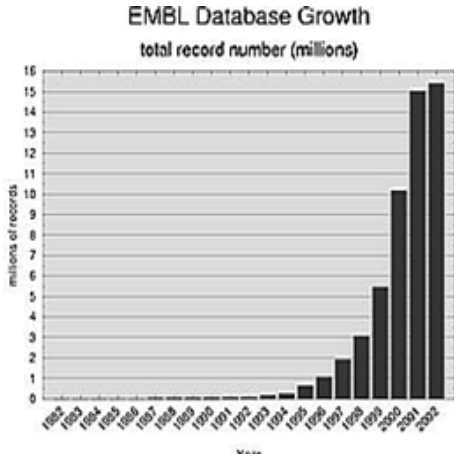
Gensequenzen: Beispiel SARS

ACTGATGCGATCGCTAATATCTACGC

```
LOCUS                   AY274119          29736 bp    RNA     linear   VRL 14-APR-2003
DEFINITION              SARS coronavirus TOR2, complete genome.
ACCESSION               AY274119
VERSION                 AY274119.1  GI:29826276
KEYWORDS    .
ORGANISM                SARS Coronavirus Tor2
                  .
BASE COUNT     8475 a    5940 c    6186 g    9135 t
ORIGIN
        1 ctacccagga aaagccaacc aacctcgatc tcttgtagat ctgttctcta aacgaacttt
       61 aaaatctgtg tagctgtcgc tcggctgcat gcctagtgca cctacgcagt ataaacaata
      ...
    29641 agccctaatg tgtaaaatta attttagtag tgctatcccc atgtgatttt aatagcttct
    29701 taggagaatg acaaaaaaaa aaaaaaaaaa aaaaaa
//
```


Grösse der Datensammlungen

ACTGATGCGATCGCTAATATCTACGC



- EMBL July 2002
  - > 150 Gbytes
- Microarray
  - 1 Petabyte p.A.
- Sanger Centre
  - 20 TB an Daten
- Genome Sequenzen wachsen p.A. um das Vierfache

Struktur – Verarbeitbarkeit

<vortrag>

<autor> R. Eckstein </autor>

<autor>S. Heymann </autor>

<titel>XML , (Modellierung) & Datenintegration - Fortschritt im Life Science Bereich</titel>

<untertitel>Schwerpunkt "Große Datenräume in Web-basierten Umgebungen„</untertitel>

<zusammenfassung>
... Die Erfassung und Darstellung von Taxonomie-Daten, den Ubergang von klassischen Morphologie-basierten botanischen Schulen zu genbasierter <term ref="http://...">Kladistik</term>. Es wird ...
</zusammenfassung>

</vortrag>

Struktur – Verarbeitbarkeit

<vortrag>

<autor> R. Eckstein </autor>

<autor>S. Heymann </autor>

<titel>XML , (Modellierung) & Datenintegration - Fortschritt im Life Science Bereich</titel>

<untertitel>Schwerpunkt "Große Datenräume in Web-basierten Umgebungen„</untertitel>

<zusammenfassung>
... Die Erfassung und Darstellung von Taxonomie-Daten, den Ubergang von klassischen Morphologie-basierten botanischen Schulen zu genbasierter <term ref="http://...">Kladistik</term>. Es wird ...
</zusammenfassung>

</vortrag>

Lesbarkeit ←→ Struktur ←→ Verarbeitbarkeit

Strukturierter Text

ACTGATGCGATCGCTAATATCTACGCG

<vortrag>

<autor> R. Eckstein </autor>

<autor>S. Heymann </autor>

<titel>XML , (Modellierung) & Datenintegration - Fortschritt im Life Science Bereich</titel>

<untertitel>Schwerpunkt "Große Datenräume in Web-basierten Umgebungen„</untertitel>

<zusammenfassung>
... Die Erfassung und Darstellung von Taxonomie-Daten, den Ubergang von klassischen Morphologie-basierten botanischen Schulen zu genbasierter <term ref= http://www.visualgenomics.ca/gordonp/xml/ > Kladistik</term>. Es wird ...
</zusammenfassung>

</vortrag>

→ **EXTENSIBLE MARKUP LANGUAGE**

EXTENSIBLE MARKUP LANGUAGE

ACTGATGCGATCGCTAATATCTACGCG

<vortrag>

<autor> R. Eckstein </autor>

<autor>S. Heymann </autor>

<titel>XML , (Modellierung) & Datenintegration - Fortschritt im Life Science Bereich</titel>

<untertitel>Schwerpunkt "Große Datenräume in Web-basierten Umgebungen„</untertitel>

<zusammenfassung>
... Die Erfassung und Darstellung von Taxonomie-Daten, den Ubergang von klassischen Morphologie-basierten botanischen Schulen zu genbasierter <term ref= http://www.visualgenomics.ca/gordonp/xml/ > Kladistik</term>. Es wird ...
</zusammenfassung>

</vortrag>

Beschreibungssprachen im Bio-Bereich:
http://www.visualgenomics.ca/gordonp/xml/

EXTENSIBLE MARKUP LANGUAGE


```
<vortrag>
        <autor> R. Eckstein </autor>
        <autor>S. Heymann </autor>
        <titel>XML , (Modellierung) & Datenintegration - Fortschritt im Life Science Bereich</titel>
        <untertitel>Schwerpunkt "Große Datenräume in Web-basierten Umgebungen„</untertitel>
        <zusammenfassung>
        ... Die Erfassung und Darstellung von Taxonomie-Daten, den Ubergang von klassischen Morphologie-
        basierten botanischen Schulen zu genbasierter <term ref= http://www.visualgenomics.ca/gordonp/xml/ >
        Kladistik</term>. Es wird ...
        </zusammenfassung>
</vortrag>
```

Beschreibungssprachen im Bio-Bereich:
http://www.visualgenomics.ca/gordonp/xml/

Handbücher **Dokumenttypdefinitionen**

Kompendium aller Erkenntnisse

DTD:

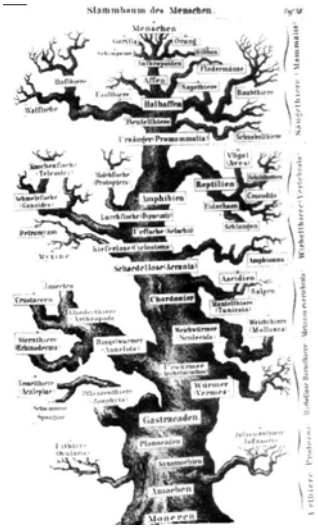
```
<!ENTITY % local.aa_type.value "">
<!ENTITY % aa_type   "type (A|B|C|D|E|F|G|H|I|K|L|M
                            |N|P|Q|R|S|T|V|W|X|Y|Z
                            %local.aa_type.value;)
                            #REQUIRED">
```

DTD's: Kompendium neuer Erkenntnisse

<!ENTITY % local.aa_type.value ""> (bis 2002)
<!ENTITY % aa_type "type (A|B|C|D|E|F|G|H|I|K|L|M
|N|P|Q|R|S|T|V|W|X|Y|Z
%local.aa_type.value;)
#REQUIRED">

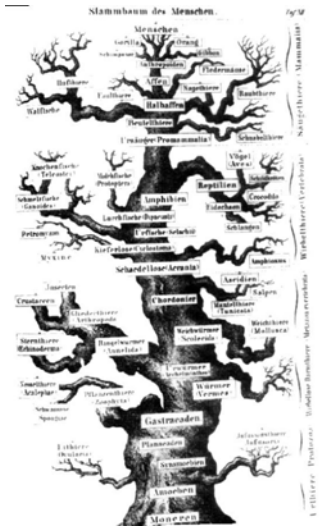
<!ENTITY % local.aa_type.value ""> (seit Feb. 2003)
<!ENTITY % aa_type "type (A|B|C|D|E|F|G|H|I|K|L|M
|N|P|Q|R|S|T|V|U|W|X|Y|Z
%local.aa_type.value;)
#REQUIRED">

Beispiele aus der Abstammungslehre



http://www.visualgenomics.ca/gordonp/xml/

Beispiele aus der Abstammungslehre



Homo sapiens:

Other names:

man[common name]

Lineage(full)

cellular organisms; Eukaryota; Fungi/Metazoa group; Metazoa; Eumetazoa; Bilateria; Coelomata; Deuterostomia; Chordata; Craniata;Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Primates; Catarrhini; Hominidae; Homo/Pan/Gorilla group; Homo

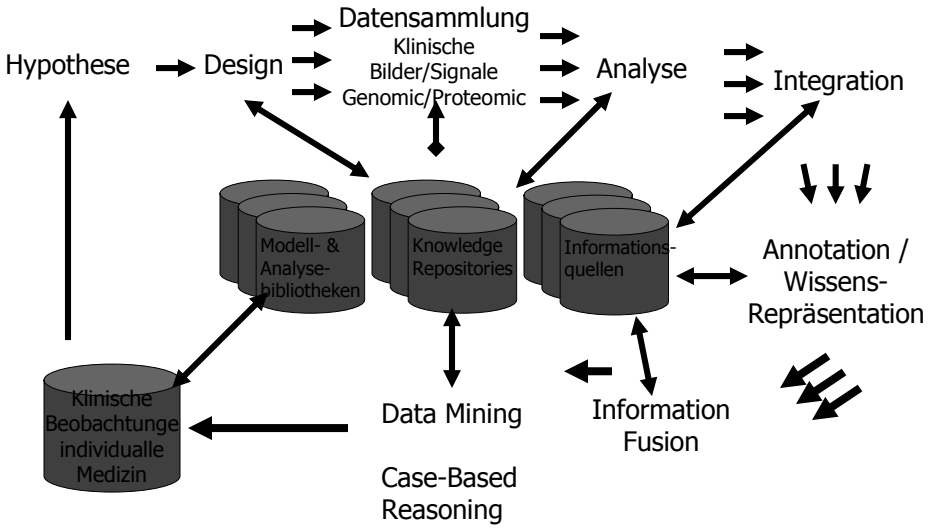
Beispiele aus der Abstammungslehre

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="HTMLOutput.xsl"?>

<!DOCTYPE MultipleClassifications SYSTEM "XMLOutput.dtd">

<MultipleClassifications Order="Alpha">
<!--#NEXUS [18-Dec-2001 16:59:36]-->
<ranks>
        <rank RankID="Family"><rankName>Family</rankName><rankValue>10</rankValue></rank>
        <rank RankID="Sub-Family"><rankName>Sub-
Family</rankName><rankValue>15</rankValue></rank>
        <rank RankID="Legion"><rankName>Legion</rankName><rankValue>17</rankValue></rank>
        <rank RankID="Tribe"><rankName>Tribe</rankName><rankValue>20</rankValue></rank>
        <rank RankID="Sub-Tribe"><rankName>Sub-
Tribe</rankName><rankValue>25</rankValue></rank>
        <rank RankID="Grex"><rankName>Grex</rankName><rankValue>27</rankValue></rank>
        <rank RankID="Genus"><rankName>Genus</rankName><rankValue>30</rankValue></rank>
</ranks>
<taxa>
        <taxon RankIDREF="Family" TaxonID="TApiaceae">Apiaceae</taxon>
        <taxon RankIDREF="Sub-Family" TaxonID="TPauciiugatae">Pauciiugatae</taxon>
..
```

Beispiel: Genetik & Medizin



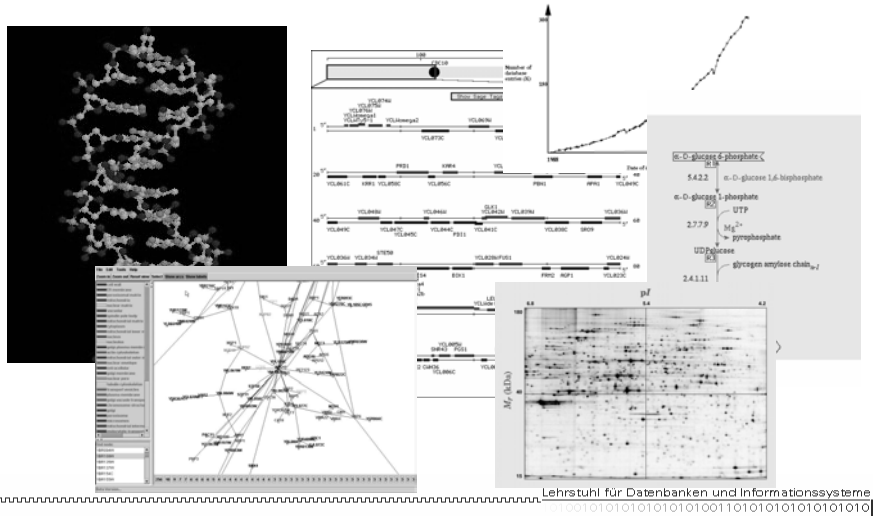
Herausforderungen...

- Formatheterogenität
- Datenheterogenität / Anzahl der Datenquellen
- Umfang der Daten / Grösse der Datensammlungen
- Zugriffsheterogenität

Formatheterogenität

ACTGATGCGATCGCTAATATCTACGCG

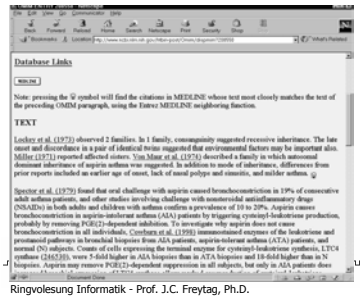
- Multimedia: Bilder & Video (e.g. microarrays, 3D, ...)



Formatheterogenität (cont.)

ACTGATGCGATCGCTAATATCTACGCG

- Text "Annotationen" & Literatur
  - strukturiert vs. semistrukturiert vs. unstrukturiert
- Unterschiedliche Formate, Strukturen, Schemata, Umfänge, …
- Web-Schnittstellen, Verteilung als Dateien, Datenbank-Dumps, XML-Dokumente, …



Database Links

Note: pressing the ☤ symbol will find the citations in MEDLINE whose text most closely matches the text of the preceding OMIM paragraph, using the Entrez MEDLINE neighboring function.

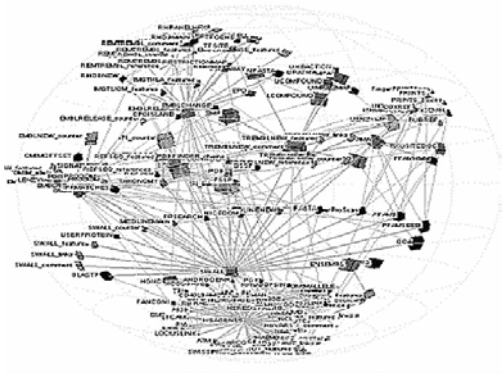
TEXT

Lockey et al. (1973) observed 2 families. In 1 family, consanguinity suggested recessive inheritance. The late onset and discordance in a pair of identical twins suggested that environmental factors may be important also. Miller (1971) reported affected sisters. Von Maur et al. (1974) described a family in which autosomal dominant inheritance of aspirin asthma was suggested. In addition to mode of inheritance, differences from prior reports included an earlier age of onset, lack of nasal polyps and sinusitis, and milder asthma.

Spector et al. (1979) found that oral challenge with aspirin caused bronchoconstriction in 19% of consecutive adult asthma patients, and other studies involving challenge with nonsteroidal antiinflammatory drugs (NSAIDs) in both adults and children with asthma confirm a prevalence of 10 to 20%. Aspirin causes bronchoconstriction in aspirin-intolerant asthma (AIA) patients by triggering cysteinyl-leukotriene production, probably by removing PGE(2)-dependent inhibition. To investigate why aspirin does not cause bronchoconstriction in all individuals, Cowburn et al. (1998) immunostained enzymes of the leukotriene and prostanoid pathways in bronchial biopsies from AIA patients, aspirin-tolerant asthma (ATA) patients, and normal (N) subjects. Counts of cells expressing the terminal enzyme for cysteinyl-leukotriene synthesis, LTC4 synthase (246318), were 5-fold higher in AIA biopsies than in ATA biopsies and 18-fold higher than in N biopsies. Aspirin may remove PGE(2)-dependent suppression in all subjects, but only in AIA patients does

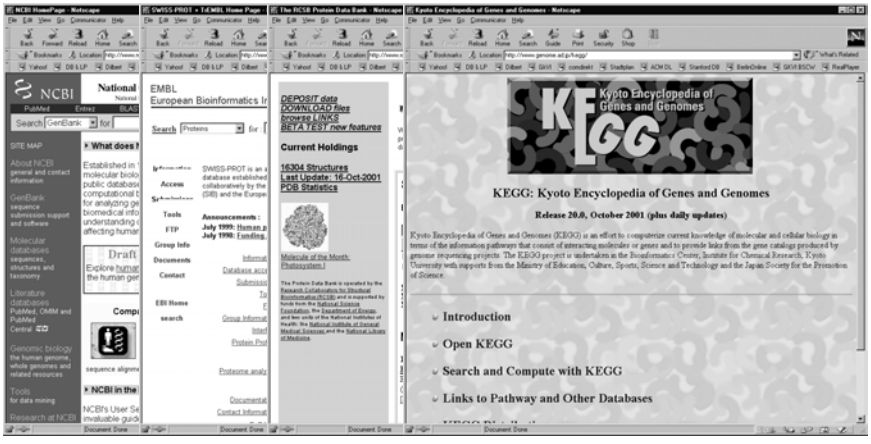
```
ID   TRBG361    standard; RNA; PLN; 1859 BP.
XX
AC   X56734; S46826;
XX
SV   X56734.1
XX
DT   12-SEP-1991 (Rel. 29, Created)
DT   15-MAR-1999 (Rel. 59, Last updated, Version 9)
XX
DE   Trifolium repens mRNA for non-cyanogenic beta-glucosidase
XX
KW   beta-glucosidase.
XX
OS   Trifolium repens (white clover)
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta.
OC   Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots.
OC   eurosids I; Fabales; Fabaceae; Papilionoideae; Trifolieae.
XX
```

Daten-/Inhaltsheterogenität

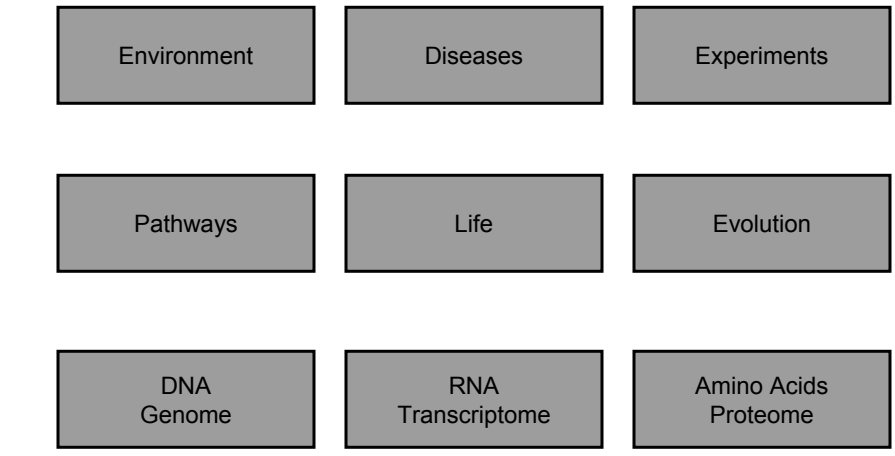


- Genomische, proteomische, transcriptomische, metabalomische, Protein-Protein Interactionen, regulatorische Bio-Netzwerke, Alinierungen, Krankheiten, Patterns & Motifs, Proteine Structuren, Proteinklassifikationen und -familien, spezielle Proteine (Enzyme, Receptoren), …

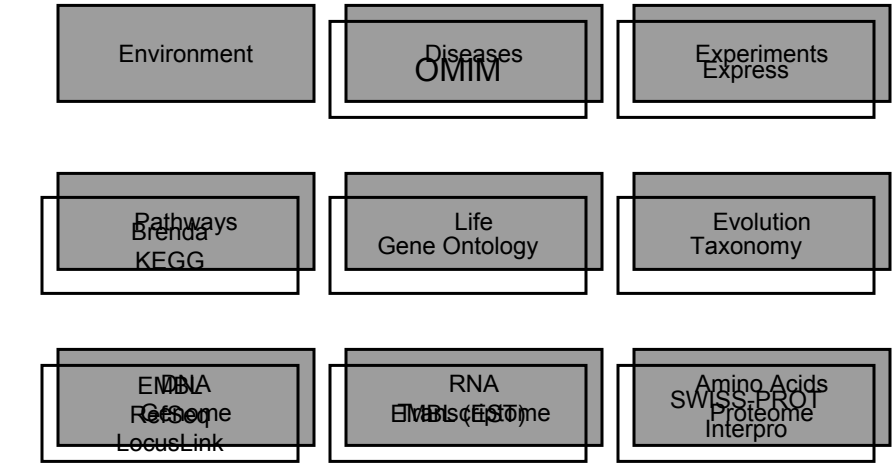
Zugriffsheterogenität



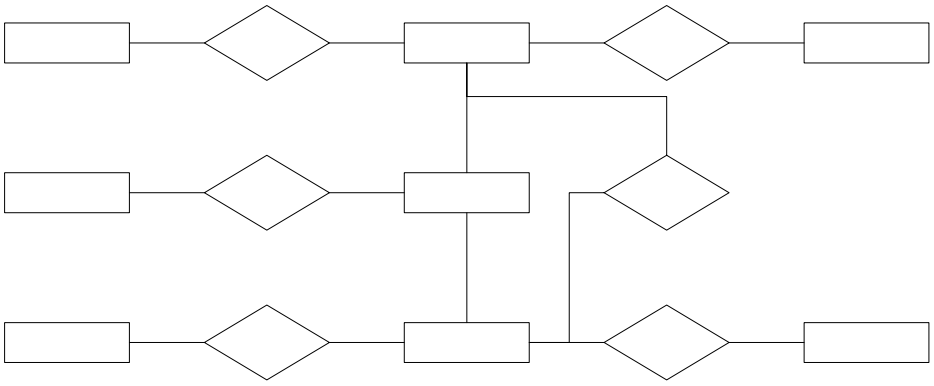
Genomisch relevante Bereiche



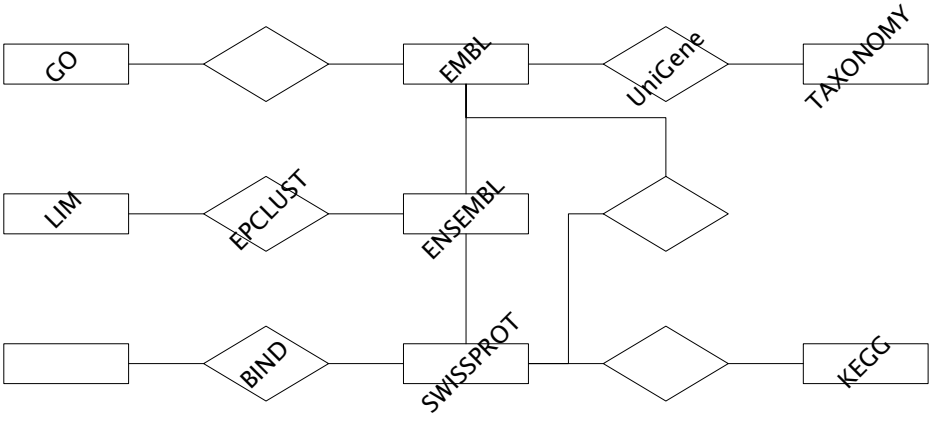
Inhalt von Datenquellen



Datenmodellierung



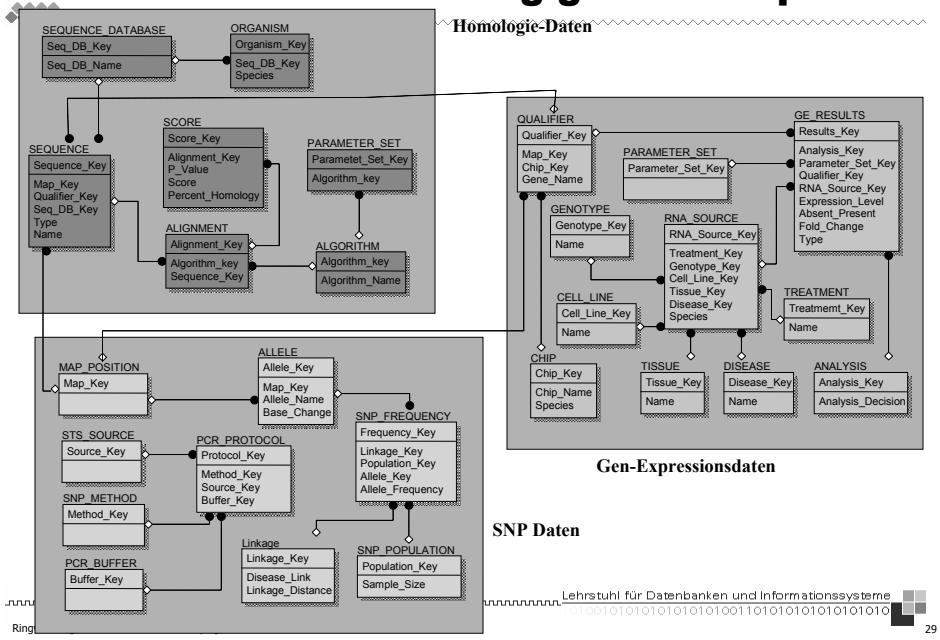
Datenmodellierung



OGY

GE-EXPERIMENT

Daten aus unabhängigen Datenquellen

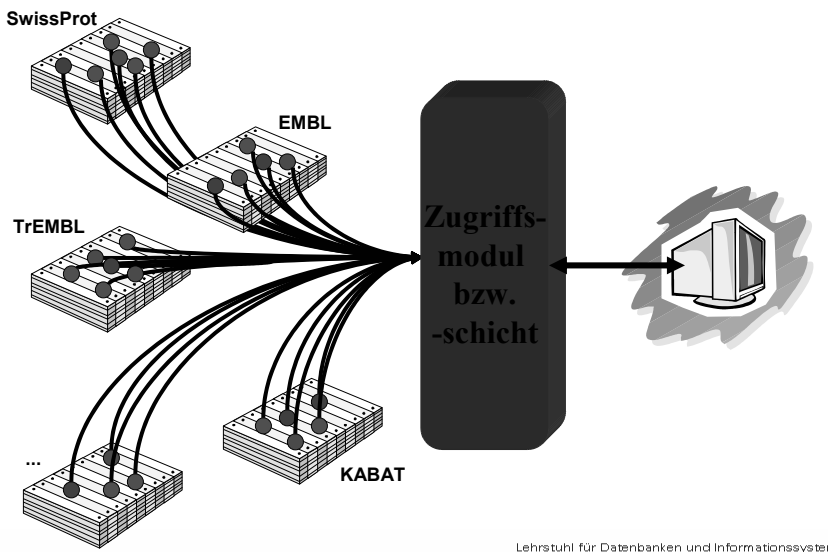


Homologie-Daten

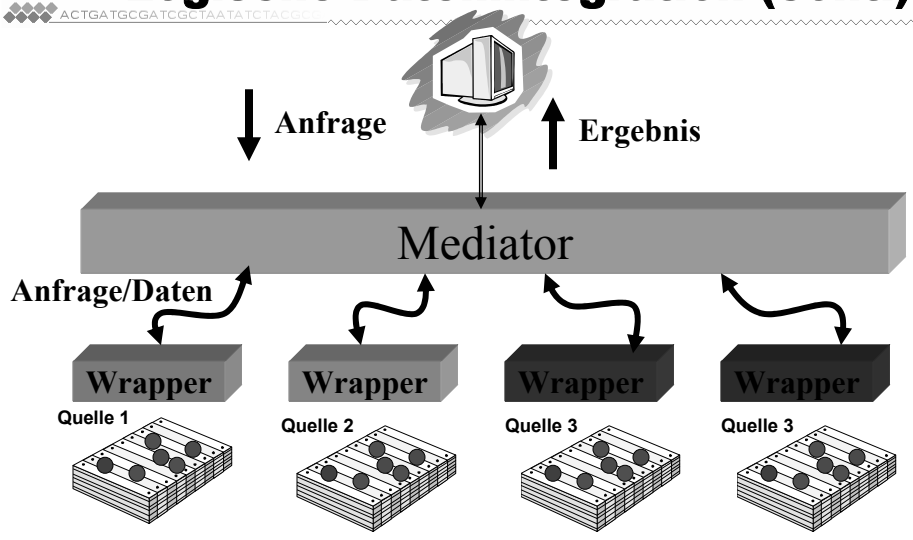
Gen-Expressionsdaten

SNP Daten

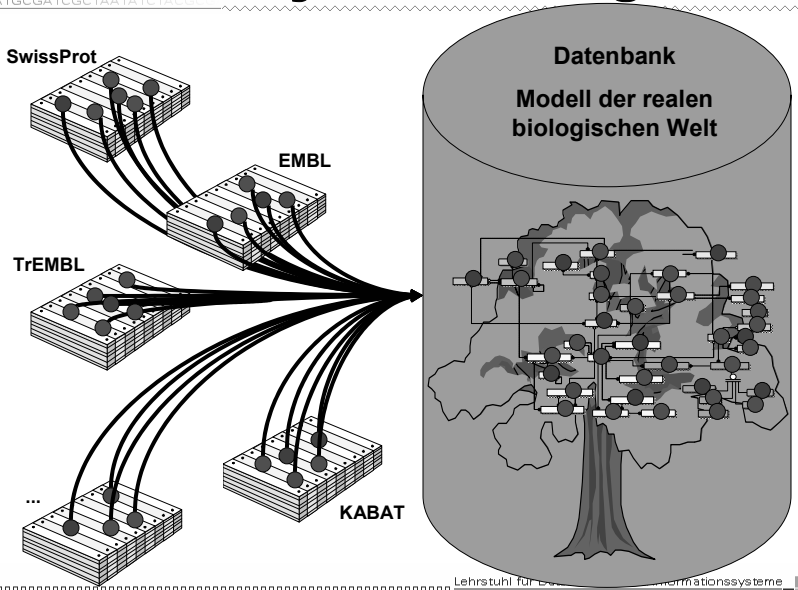
Logische Datenintegration



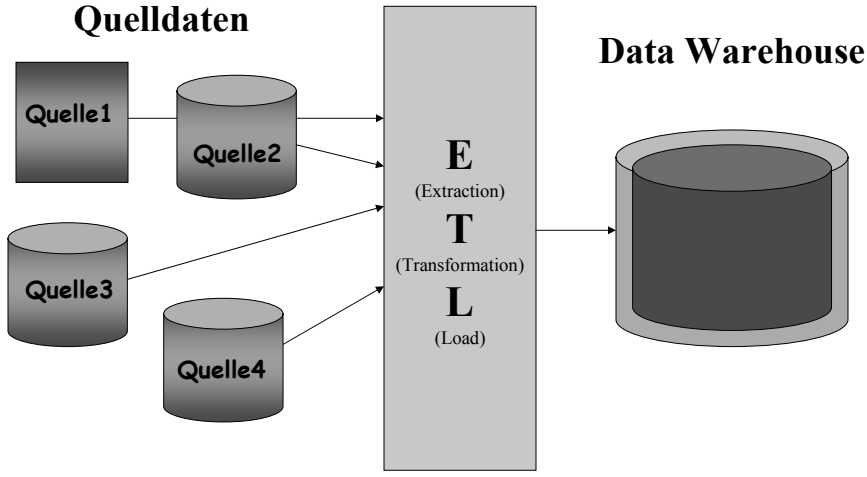
Logische Datenintegration (cont.)



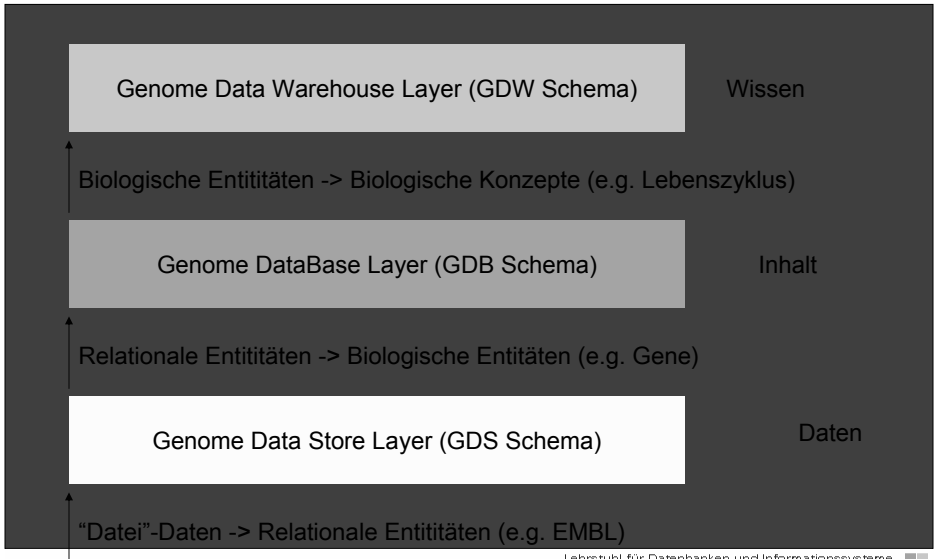
Physische Integration



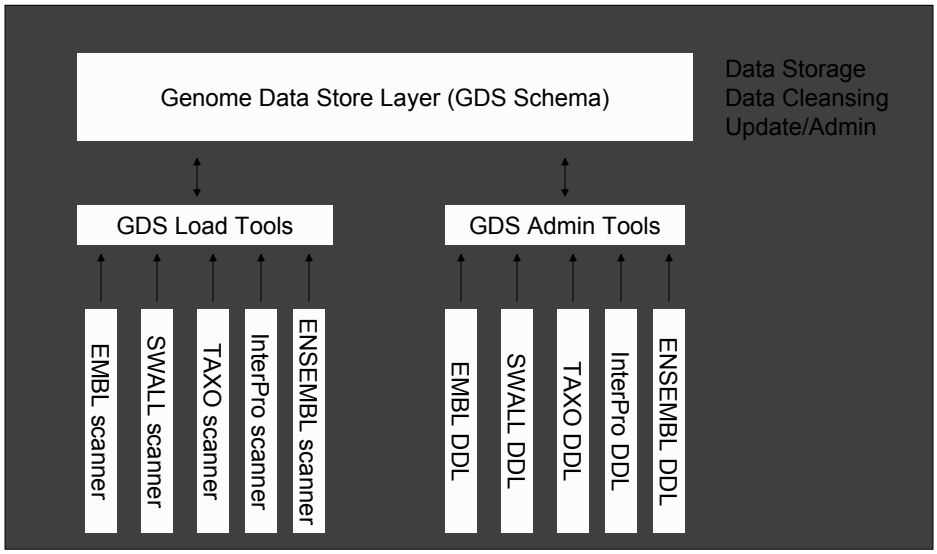
Data Warehousing



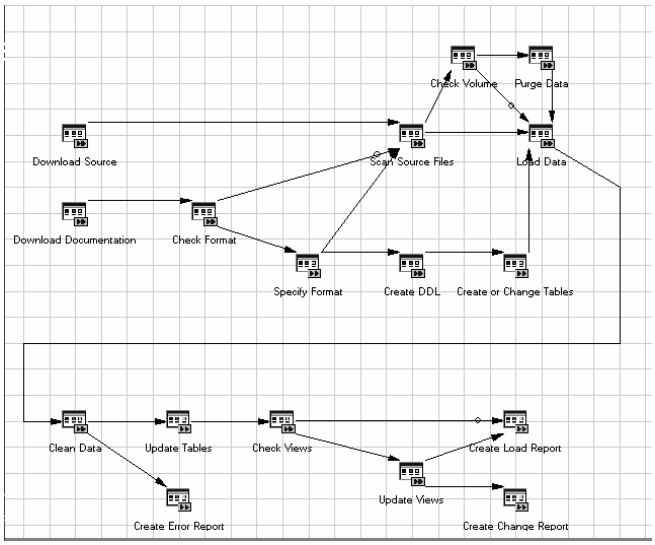
Gene-EYe Integrationsplattform



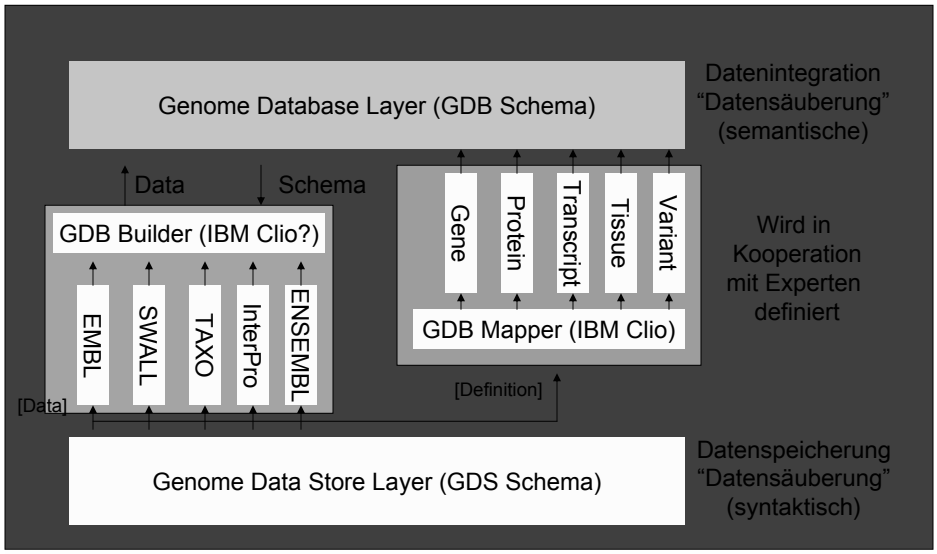
GDS: Von der Datei zur Datenbank



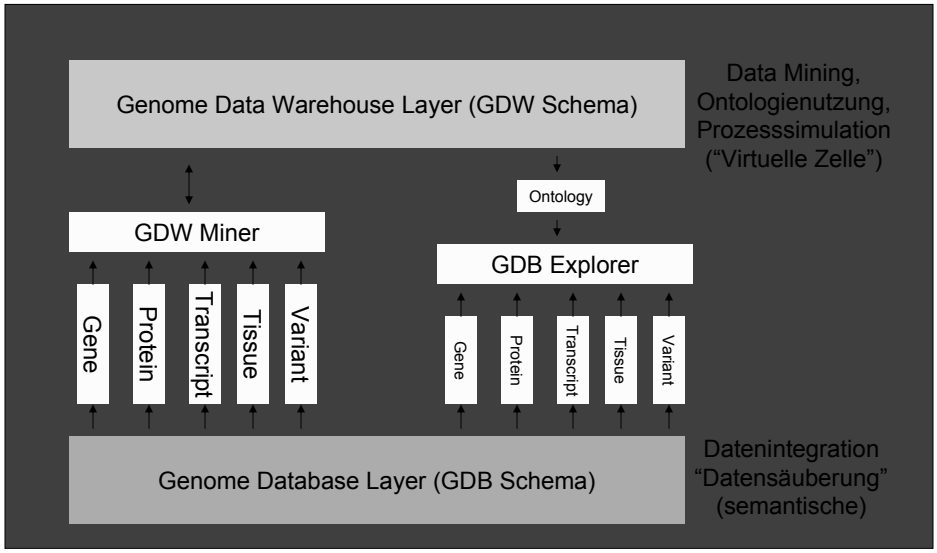
Modellierung des "Wartungsprozesses"



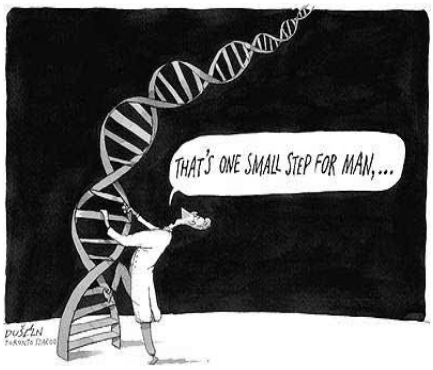
GDB: von den Daten zur Biologie



GDW: Plattform für Forschung



What are the goals?



Source: **Dusan Petricic, Toronto, Ontario** -- The Toronto Star
http://cagle.slate.msn.com/news/gene/gene6.asp

Fragen??