IZBI
University Leipzig

UNIVERSITÄT LEIPZIG

Data Integration in Bioinformatics and Life Sciences

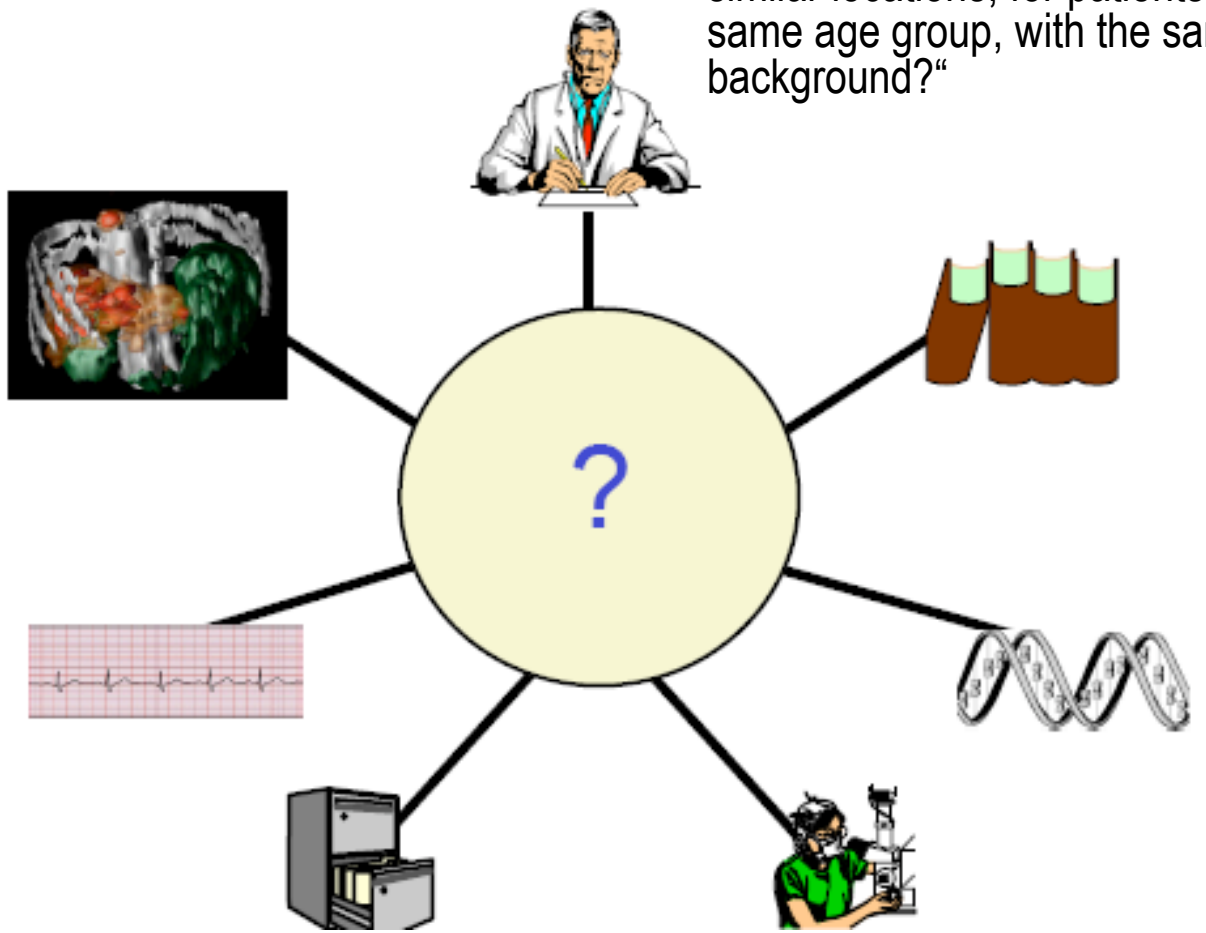
Erhard Rahm, Toralf Kirsten, Michael Hartung

http://dbs.uni-leipzig.de
http://www.izbi.de

EDBT – Summer School, September 2007

What is the Problem?

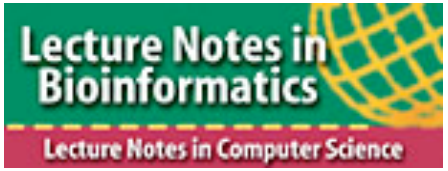
„What protocols were used for tumors in similar locations, for patients in the same age group, with the same genetic background?“



Source: L. Haas, ICDE2006 keynote

DILS workshop series

- International workshop series
  Data Integration in the Life Sciences (DILS)



- DILS2004: Leipzig
  (Interdisciplinary Center for Bioinformatics)
- DILS2005: San Diego, USA
  (UCSD Supercomputing Center)
- DILS2006: Cambridge/Hinxton, UK
  (EBI)
- DILS2007: Philadelphia (UPenn)

- DILS2008: Have you ever been in Paris? ☺

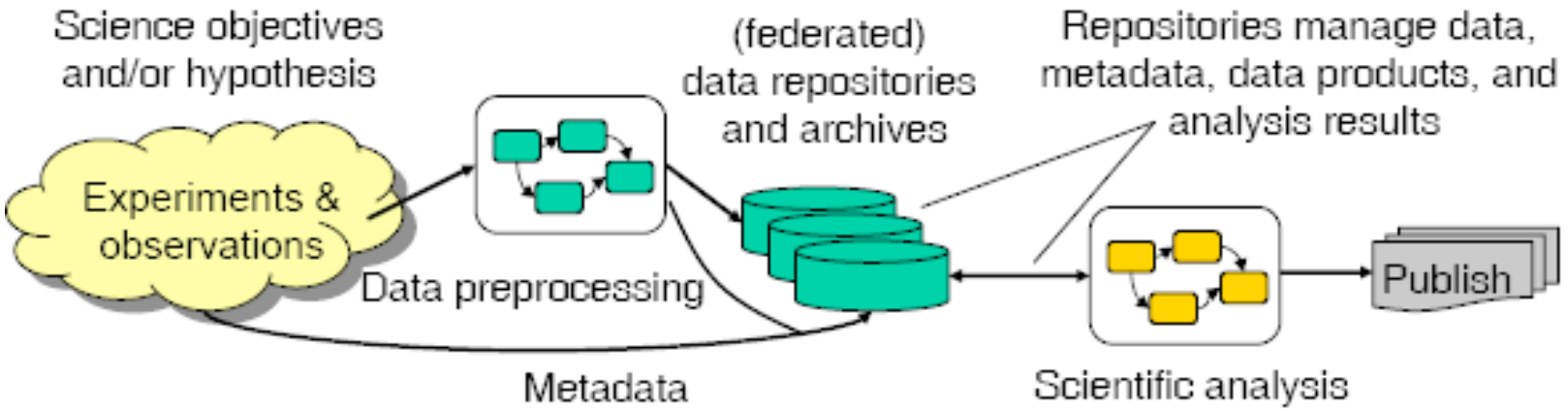
Agenda

- Kinds of data to be integrated
- General data integration alternatives
- Warehouse approaches
- Virtual and mapping-based data integration
- Matching large life science ontologies
- Data quality aspects
- Conclusions and further challenges

Agenda

- Kinds of data to be integrated
  - Experimental data
  - Clinical data
  - Public web data
  - Ontologies
- General data integration alternatives
- Warehouse approaches
- Virtual and mapping-based data integration
- Matching large life science ontologies
- Data quality aspects
- Conclusions and further challenges

Scientific data management process



Source: Gertz/Ludaescher: SDM Tutorial, EDBT2006

"Old style SDM"

1. formulate hypothesis
2. design experiment
3. run experiment
4. analyze result
5. evaluate hypothesis

Trend

1. formulate hypothesis
2. **lookup and explore data**
3. evaluate hypothesis

- Sharing/reuse of data products
- community-oriented research

Data integration in life sciences

- Many heterogeneous data sources
  - Experimental data produced by chip-based techniques
    - Genome-wide measurement of gene activity under different conditions (e.g., normal vs. different disease states)
  - Experimental annotations (metadata about experiments)
  - Clinical data
  - Lots of inter-connected web data sources and ontologies
    - Sequence data, annotation data, vocabularies, …
  - Publications (knowledge in text documents)
  - Private vs. public data

- Different kinds of analysis
  - Gene expression analysis
  - Transcription analysis
  - Functional profiling
  - Pathway analysis and reconstruction
  - Text mining , …

Affymetrix gene expression microarray

Expression experiment and analysis

(1) Cell selection

sample

mRNA

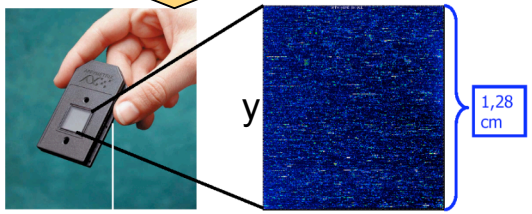
(2) RNA/DNA preparation

labeling

(3) Hybridization

array

(4) Array scan



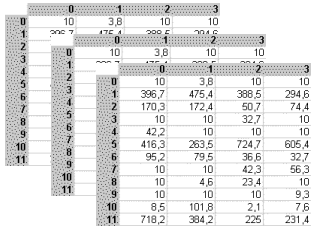
array image

(5) Image analysis

x

y	0	1	2	3
0	10	3,8	10	10
1	396,7	475,4	388,5	294,6
2	170,3	172,4	50,7	74,4
3	10	10	32,7	10
4	42,2	10	10	10
5	416,3	263,5	724,7	605,4
6	95,2	79,5	36,6	32,7
7	10	10	42,3	56,3
8	10	4,6	23,4	10
9	10	10	10	9,3
10	8,5	101,8	2,1	7,6
11	718,2	384,2	225	231,4

array spot intensities



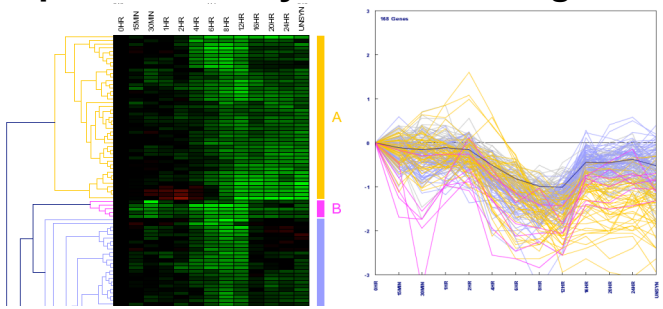
spot intensities for experiment series

(6) Data pre-processing

Genes	Experiments				
	HK-X1	HK-X2	HK-X3	HK-X4	HK-X5
1000_at	24,3	32,6	25,6	35,8	27,2
1001_at	38,5	46,6	35,2	49,8	32,3
1002_at	1002,8	1175,5	1235,7	1193,4	1045,2
1003_at	978,3	1037,8	989,3	1023,8	967,2
1110_at	207,6	239,4	234,1	238,2	214,9
3140_at	757,3	787,6	762,9	764,9	734,2

gene expression matrix

(7) Expression analysis/data mining



(8) Interpretation using annotations

Gene groups (co-regulated, ...)

Experimental data

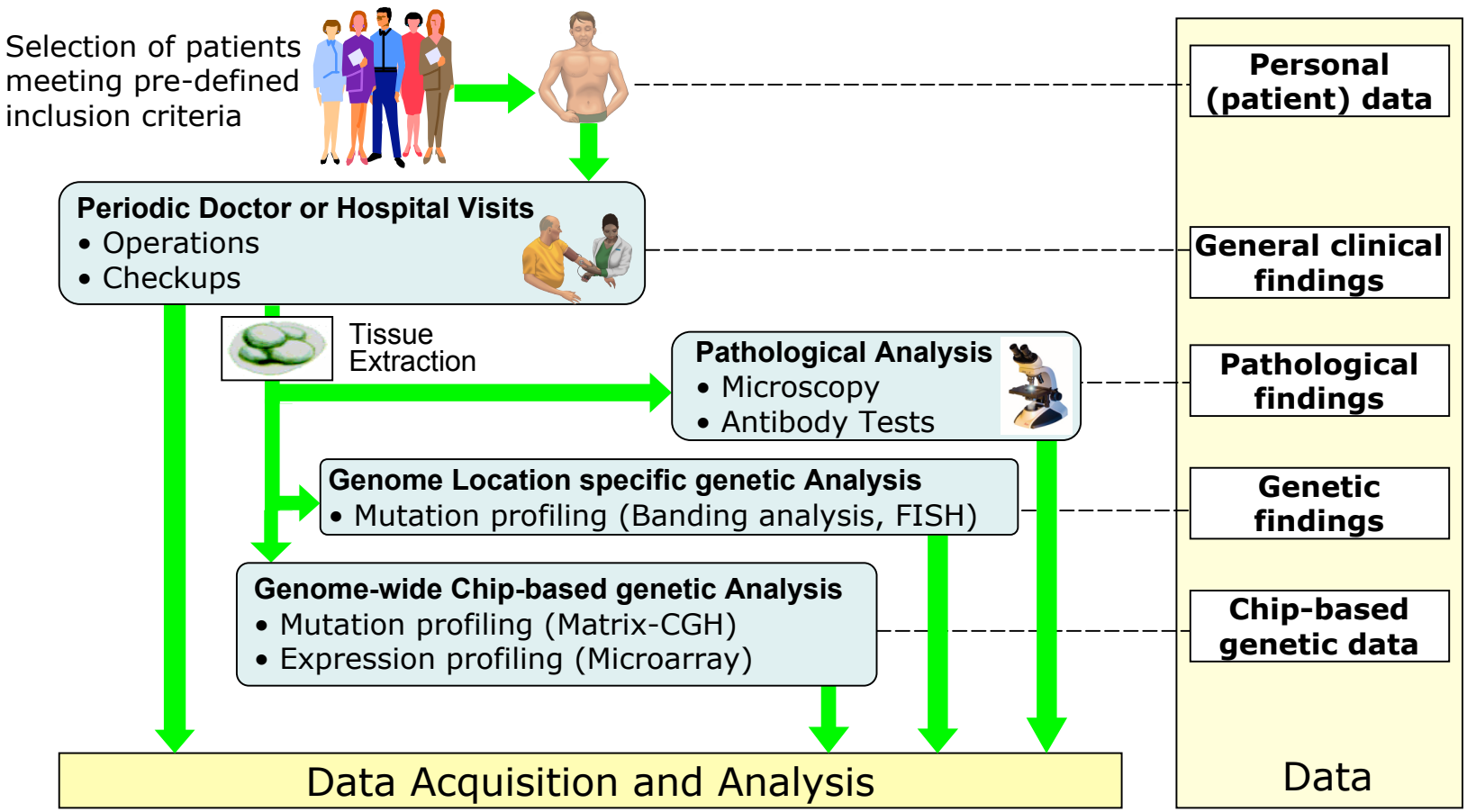
- High volume of experimental data
  - Various existing chip types for gene expression and mutation analysis
  - Fast growing amount of numeric data values
- Need to pre-process chip data (no standard routines)
  - Different data aggregation levels (e.g. Affy probe vs. probeset expression values)
  - Various statistical approaches, e.g. tests and resampling procedures, …
  - Visualizations, e.g. Heatmap, M/A plot, …
- Need for comprehensive, standardized experimental annotations
  - Experimental set up and procedure (hybridization process, utilized devices, …
  - Manual specification by the experimenter
  - Often user-dependent utilization of abbrev. and names / synonyms
  - Recommendation: **M**inimal **I**nformation **a**bout a **M**icroarray **E**xperiment*

* Brazma et al.: *Minimum information about a mircoarray experiment (MIAME) – toward standards for microarray data*. Nature Genetics, 29(4): 365-371, 2001

Clinical data: Requirements

- Patient-oriented data
  - Personal data
  - Different types of findings, e.g. general clinical findings (blood pressure, etc.), pathological findings (tissue samples), genetic findings
  - Applied therapies (timing and dosages of drugs, …)
- Clinical studies to evaluate and improve treatment protocols, e.g. against cancer
  - Data acquisition during complex workflows running in different hospitals
  - Special software systems for study management (eResearch Network, Oracle Clinical, ...)
- New research direction: collect and evaluate genetic data (e.g., gene expression data) within clinical studies to investigate molecular-biological causes of diseases and impact of drugs
- Need to integrate experimental and clinical data within distributed study management workflows
- High privacy requirements: protect identity of individual patients

Clinical trials: Inter-organizational workflows

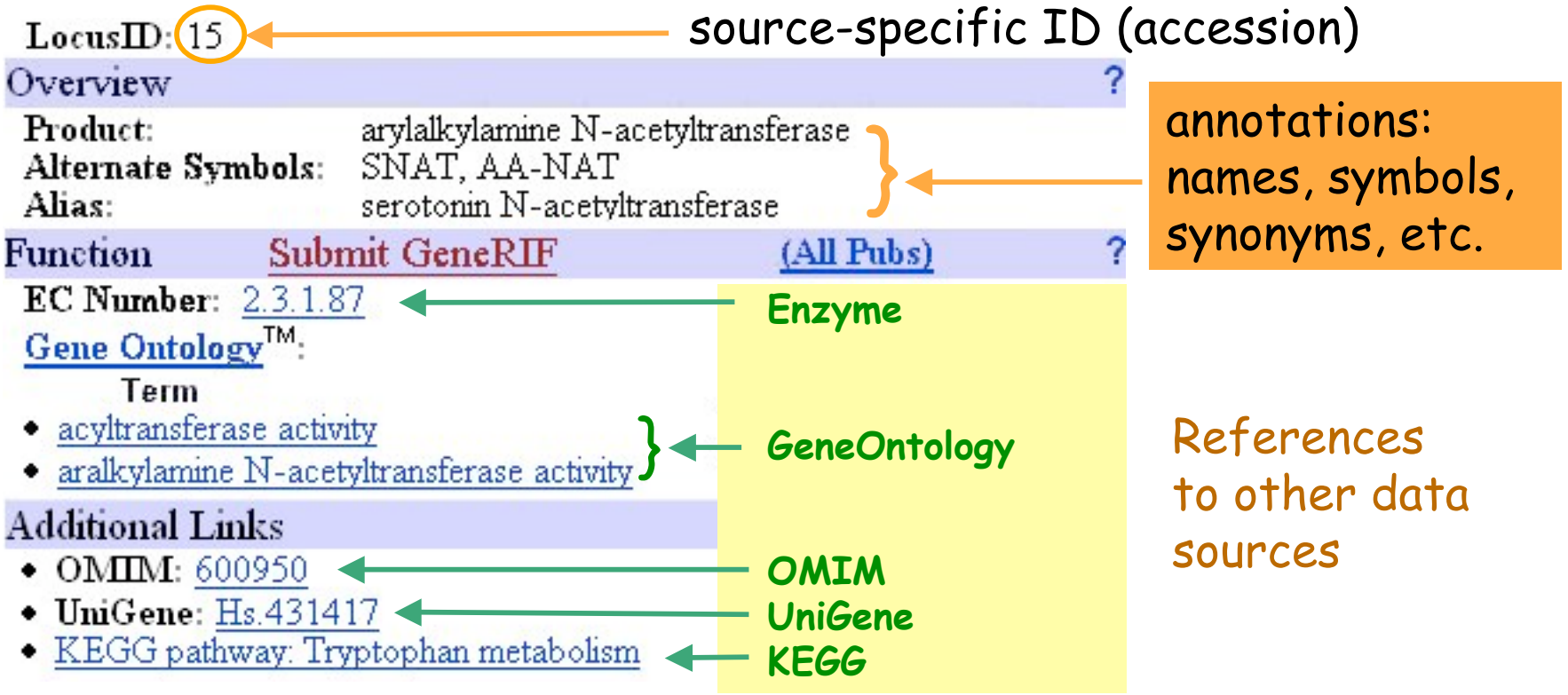


Publicly accessible data in web sources

- Genome sources: Ensembl, NCBI Entrez, UCSC Genome, ...
  - Objects: Genes, transcripts, proteins etc. of different species
- Object specific sources
  - Proteins: UniProt (SwissProt, Trembl), Protein Data Bank, ...
  - Protein interactions: BIND, MINT, DIP, ...
  - Genes: HUGO (standardized gene symbols for human genome), MGD, ...
  - Pathways: KEGG (metabolic & regulatory pathways), GenMAPP, ...
  - ...
- Publication sources: Medline / Pubmed (>16 Mio entries)
- Ontologies
  - Utilized to describe properties of biological objects
  - Controlled vocabulary of concepts to reduce terminology variations
  - Popular examples: Gene Ontology, Open Biomedical Ontologies (OBO)

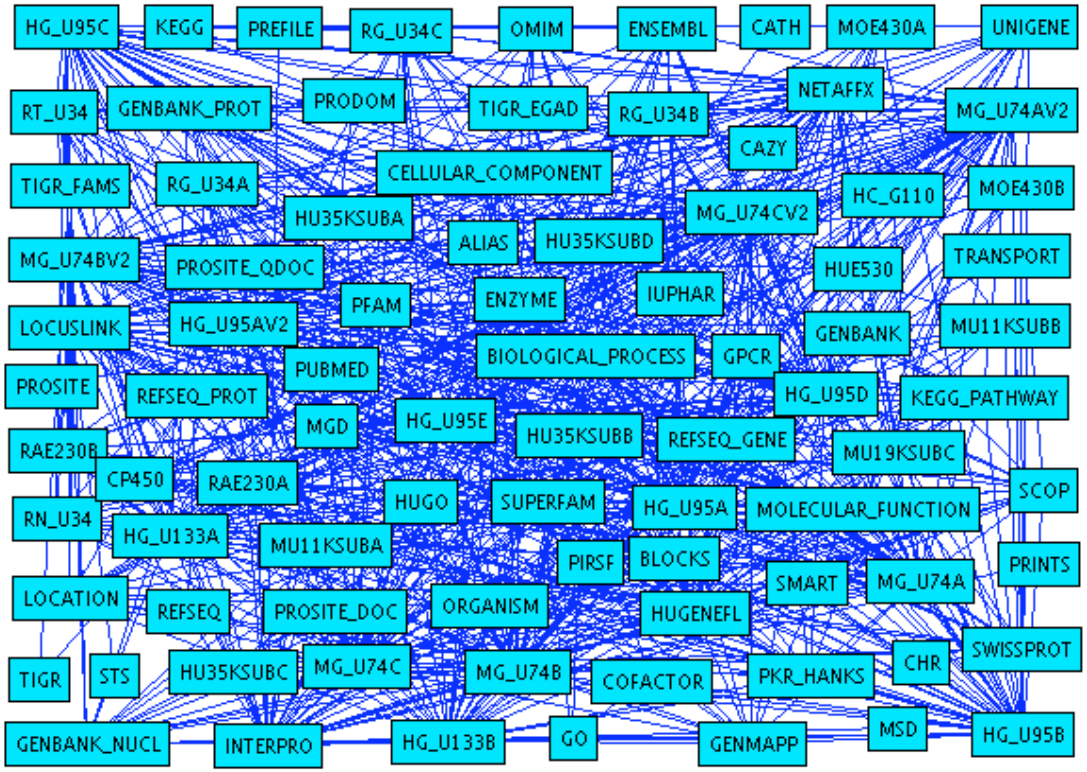
Sample web data with cross-references

- Annotation data vs. mapping data



- Problem: semantics of mappings (missing mapping type)
  - Gene ←→ gene: orthologous vs. paralogous genes

Highly connected data sources



- Many, highly connected data sources and ontologies
- Heterogeneity
  - Files and databases
  - Format and schema differences
  - Semantics
- Incomplete data sources
- Overlapping data sources

  → need to fuse corresponding objects from different sources
- Frequent changes
  - Data, schema, APIs

- common (global) database schema ???

Ontologies

- Increasing use of ontologies in bioinformatics and medicine to organize domains, annotate data and support data integration
  - Develop a shared understanding of concepts in a domain
  - Define the terms used
  - Attach these terms to real data (annotation)
  - Provide ability to query data from different sources using a common vocabulary
- Some popoluar life science ontologies
  - Gene Ontology (http://www.geneontology.org)
    - Species-independent, comprehensive sub-ontologies about Molecular Functions, Biological Processes and Cellular Components
  - UMLS – Unified Medical Language System (http://www.nlm.nih.gov/research/umls/umlsmain.html)
    - Metathesaurus comprising medical subjects and terms of Medical Subject Headings, International Classification of Diseases (ICD), …

OBO – Open Biomedical Ontologies

- An umbrella project for grouping different ontologies in biological/medical field

Why OBO?

- GO only covers three specific domains
- Other aspects could also be annotated: anatomy, …
- No standardization of ontologies: format, syntax, …
- What ontologies do exist in the biomedical domain?
- Creation takes a lot of work → Reuse existing ontol.

Requirements for ontologies in OBO:

- Open, can be used by all without any constraints
- Common shared syntax
- No overlap with other ontologies in OBO
- Share a unique identifier space
- Include text definitions of their terms

Currently covered aspects:

- Anatomies
- Cell Types
- Sequence Attributes
- Temporal Attributes
- Phenotypes
- Diseases
- ….

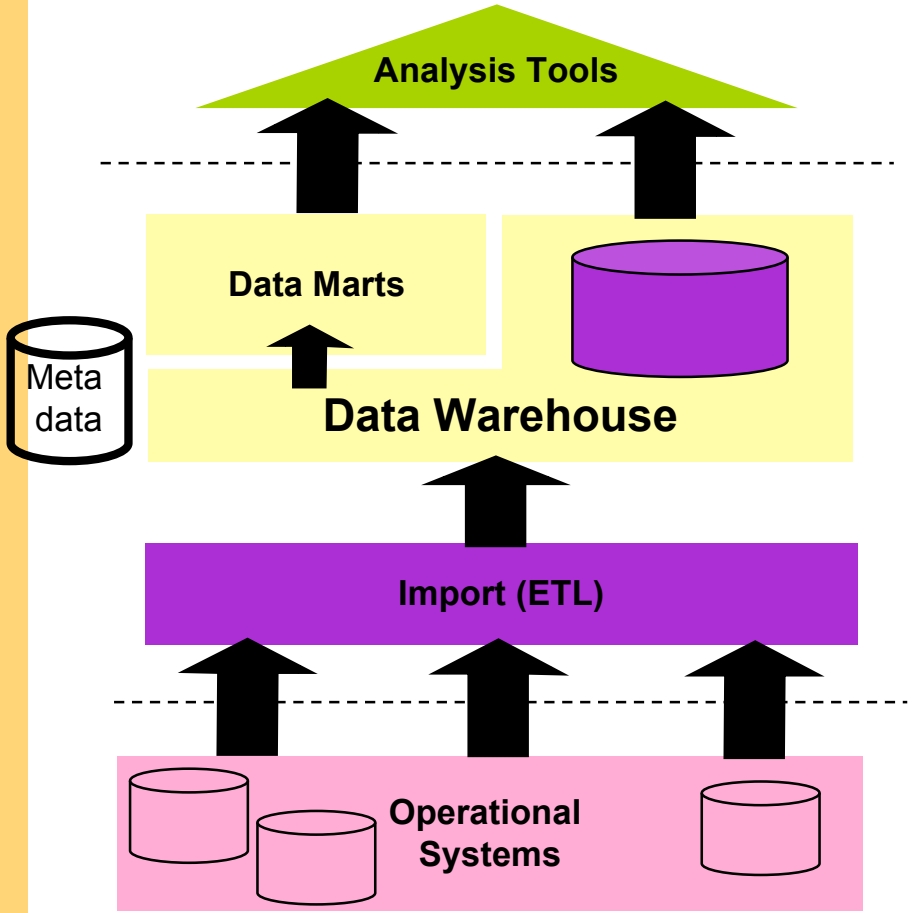
http://obo.sourceforge.net/main.html

Agenda

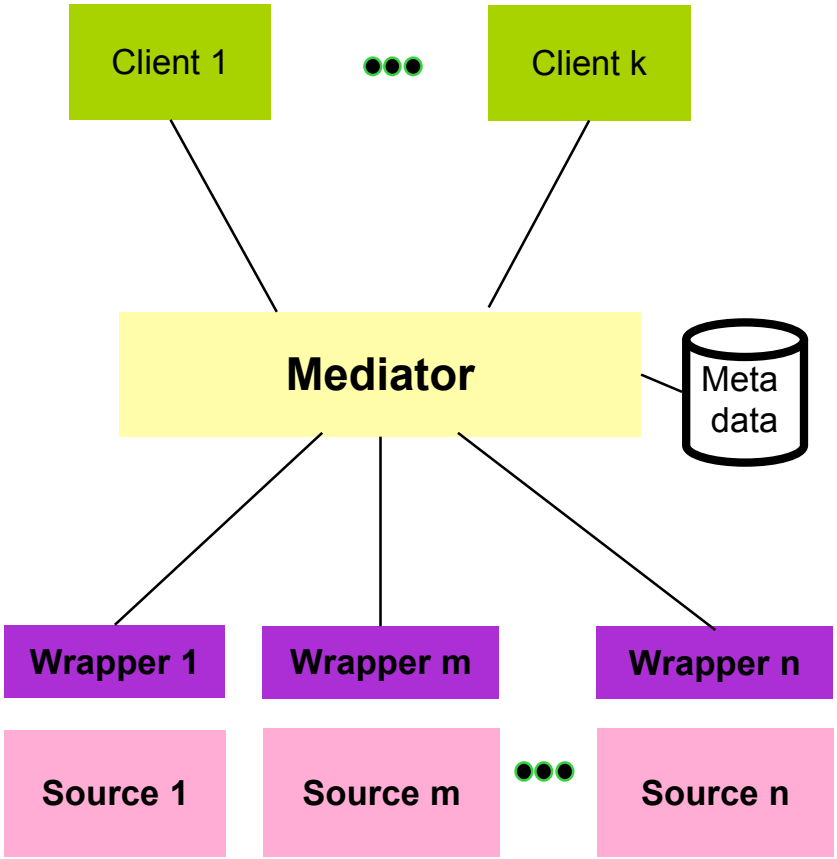
- Kinds of data to be integrated
- General data integration alternatives
  - Physical vs. virtual integration
  - P2P-like / Peer Data Management Systems (PDMS)
  - Scientific workflows
- Warehouse approaches
- Virtual and mapping-based data integration
- Matching large life science ontologies
- Data quality aspects
- Conclusions and further challenges

Instance integration: Physical vs. virtual

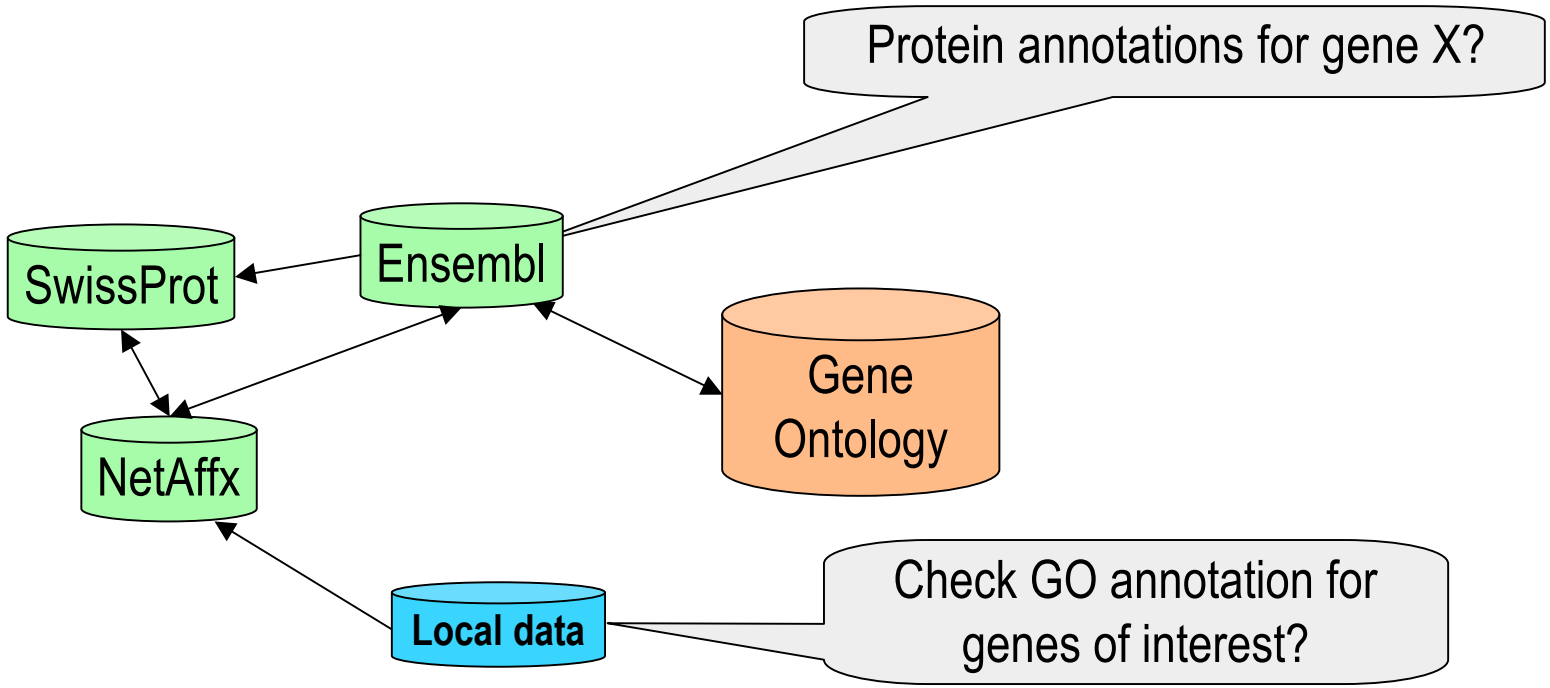
Physical Integration
(Data Warehousing)



Virtual Integration
(query mediators)



Peer Data Integration: Typical Scenario

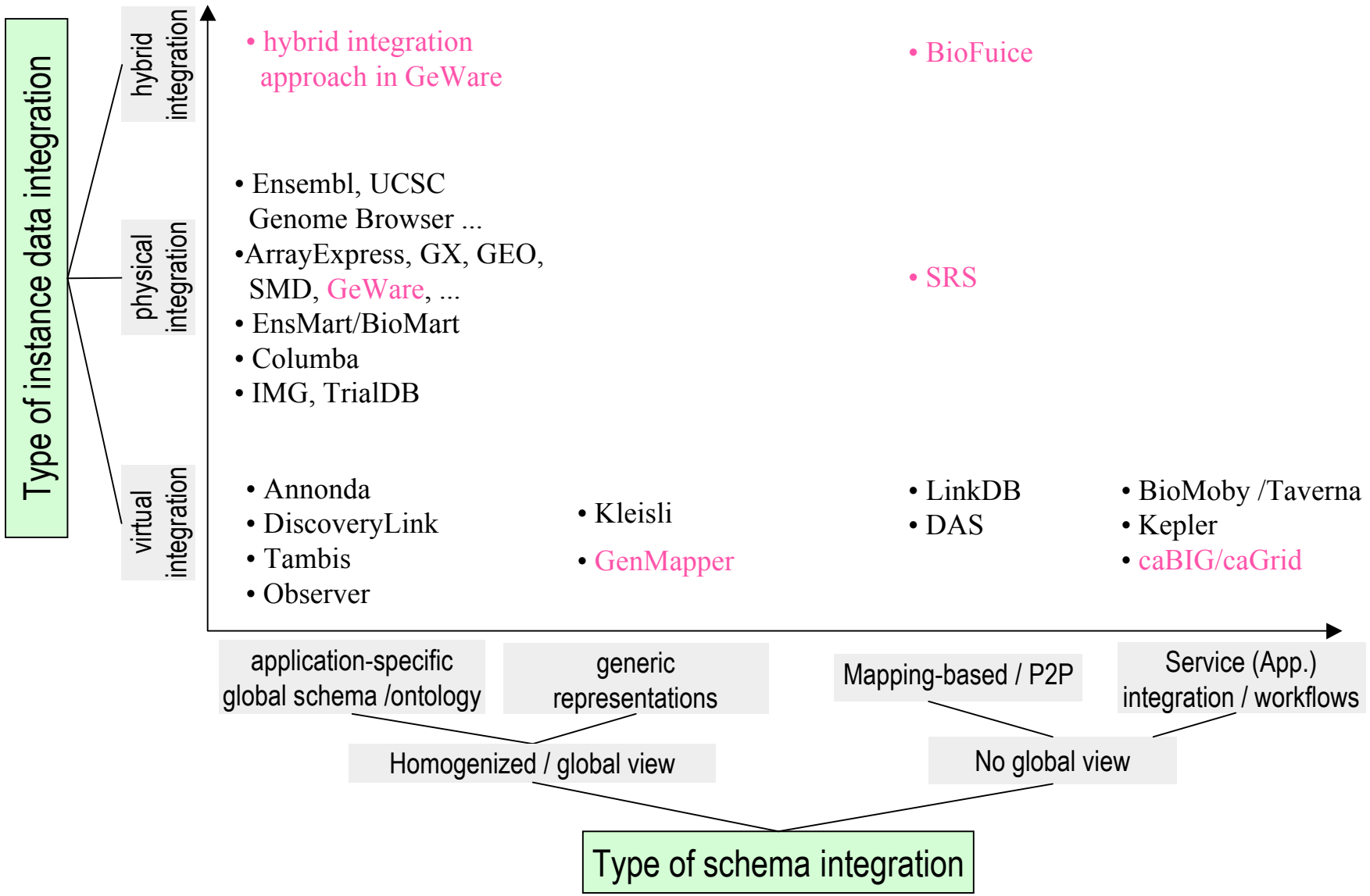


- Bidirectional mappings between data sources instead of global schema
- Queries refer to single source and are propagated to relevant peers
- Adding new sources becomes simpler
  - Support for local data sources (e.g. private gene list)

Data integration: Physical vs. virtual

	Physical (Warehouse)	Virtual	
		Query mediators	Peer Data Mgmt
Schema integration	A priori	A priori	No schema integration
Instance data integration	A priori	At query runtime	At query runtime
Achievable data quality	+	o	o
Analysis of large data volumes	+	-	-
(HW) ressource requirements	-	o	o
Data freshness	o	+	+
Source autonomy	o	+	+
Scalability to many sources	-	-	o

Classification of data integration approaches



Application-specific vs. generic representation

Application-specific global schema

Generic representation using EAV

Protein

Accession	Name	Organism	...
ENSP00000226317	Cytokine B6 precursor	Homo Sapiens	
ENSP00000306512	Interleukin-8 precursor	Homo Sapiens	
...			

Metadata

Instance data

Entity

Entity_ID	Name
1	Protein
2	ProteinFunctionRel
3	Function

Attribute

Attribute_ID	Entity_ID	Name
1	1	Accession
2	1	Name
3	1	Organism
4	...	...

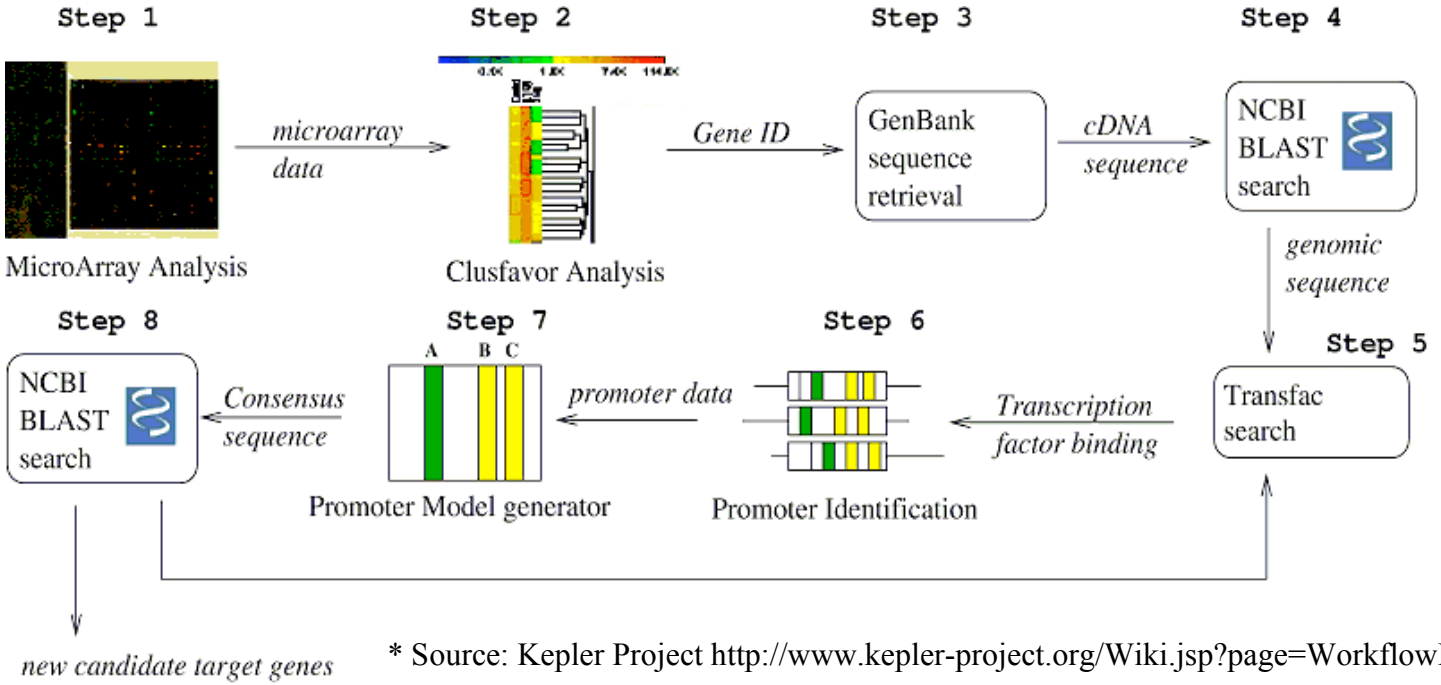
AttributeValue

Tupel_ID	Attribute_ID	Value
1	1	ENSP00000226317
1	2	Cytokine B6 precursor
1	3	Homo Sapiens
2	1	ENSP00000306512
2	2	...

Generic representation
Flexible and extensible, but hard to query

Scientific Workflows

- Integrate data sources at the application (analysis) level
  - Complementary to data-focussed integration approaches
  - Reuse of existing applications, services, and (sub-) workflows
  - Issues: semantically rich service registration, service composition (matching), manipulation of result data, monitoring and debugging workflow execution, …
- Example: Promoter Identification Workflow*



* Source: Kepler Project http://www.kepler-project.org/Wiki.jsp?page=WorkflowExamples

Agenda

- Kinds of data to be integrated
- General data integration alternatives
- Warehouse approaches
  - The GeWare platform for microarray data management
    - Architecture; preprocessing and analysis workflows
    - Integrating data from clinical studies
    - Generic annotation management
  - Hybrid integration for expression + annotation analysis
- Virtual and mapping-based data integration
- Matching large life science ontologies
- Data quality aspects
- Conclusions and further challenges

The GeWare system*

- Many platforms for microarray data management: ArrayExpress (EBI), Gene Expression Omnibus (NCBI), Stanford Microarray Database, ...
- GeWare – Genetic Data Warehouse (U Leipzig)
  - Under development since 2003
- Central data management and analysis platform
  - Data of chip-based experiments (i.e. expression microarrays & Matrix-CGH arrays)
  - Uniform and autonomous specification of experiment annotations
  - Import of clinical data
  - Integration of gene annotations from public sources
  - Various methods for pre-processing, analysis and visualization
  - Coupling with existing tools for powerful and flexible analysis, e.g. R packages, BioConductor

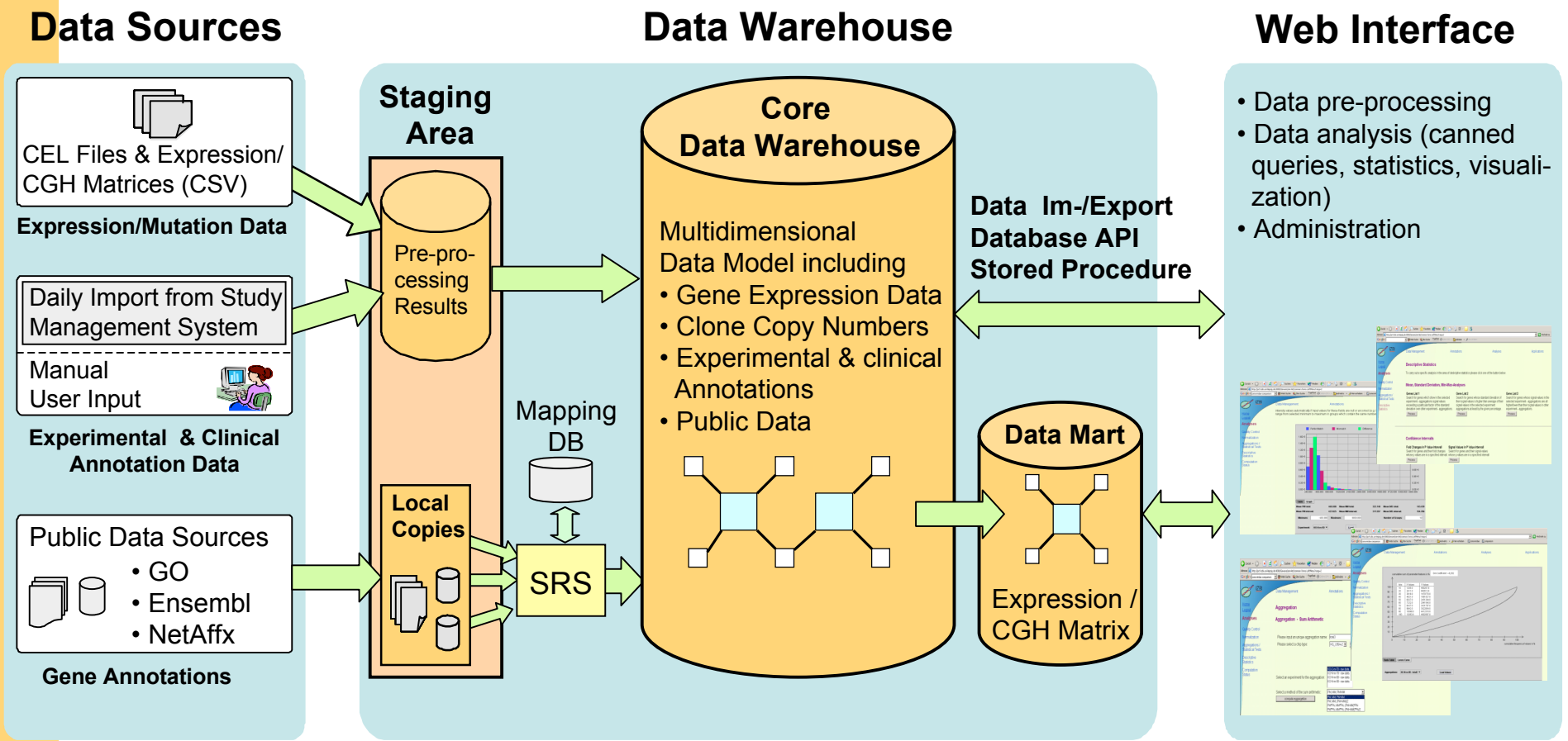
*Rahm, E; Kirsten, T; Lange, J: *The GeWare data warehouse platform for the analysis molecular-biological and clinical data.* Journal of Integrative Bioinformatics, 4(1):47, 2007

GeWare Applications

- Two collaborative cancer research studies
  - Molecular Mechanism in Malignant Lymphoma (MMML) http://www.lymphome.de/Projekte/MMML
  - German Glioma Network: http://www.gliomnetzwerk.de/
  - Data from several national clinical, pathological and molecular-genetics centers
  - Experimental and clinical data for hundreds of patients

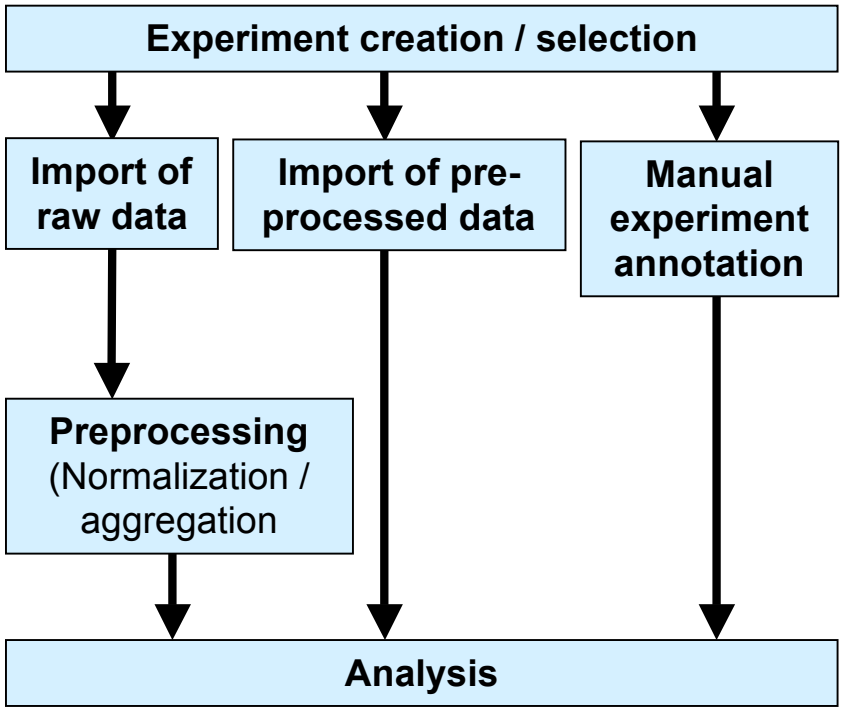
- Local research groups at the Univ. Leipzig, e.g.
  - Expression analysis of different types of human thyroid nodules
  - Expression analysis of physiological properties of mice
  - Analysis of factors influencing the specific binding of sequences on microarrays

System architecture

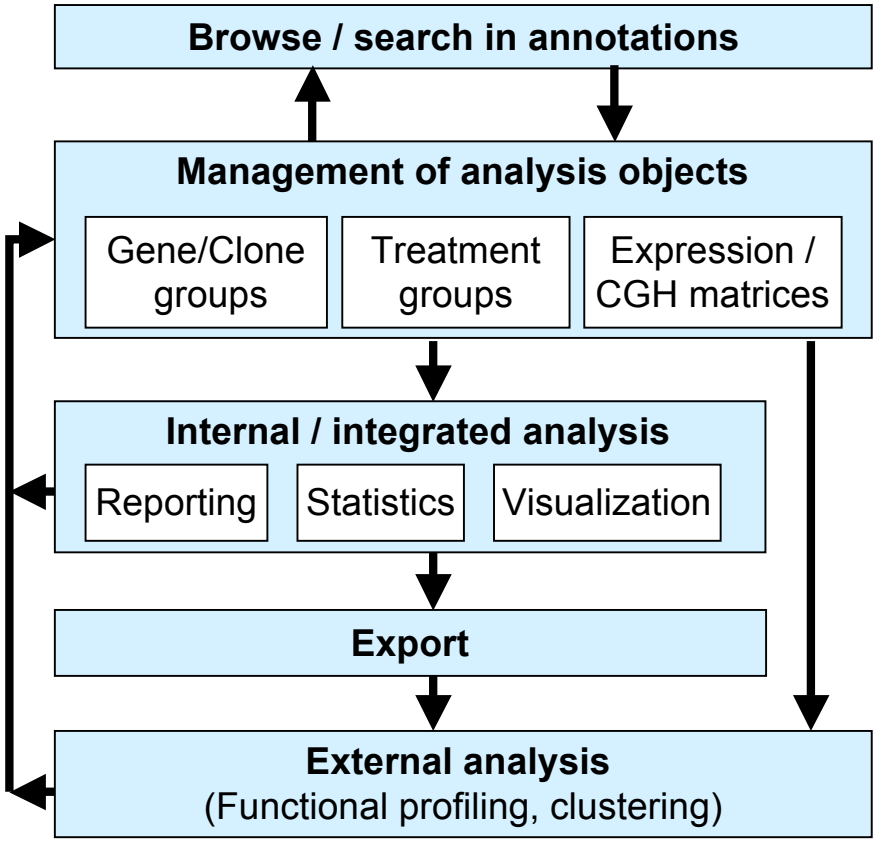


GeWare – System workflows

Import Workflow

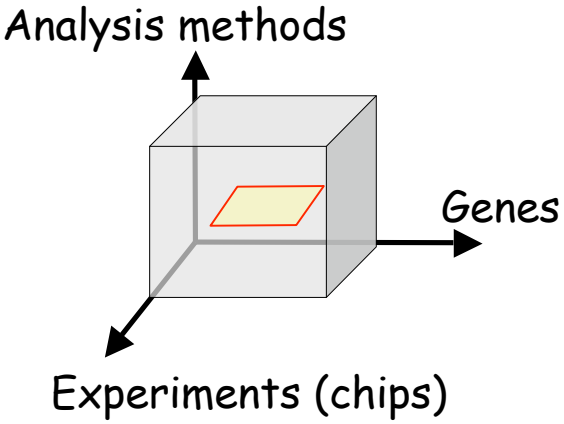


Analysis Workflow

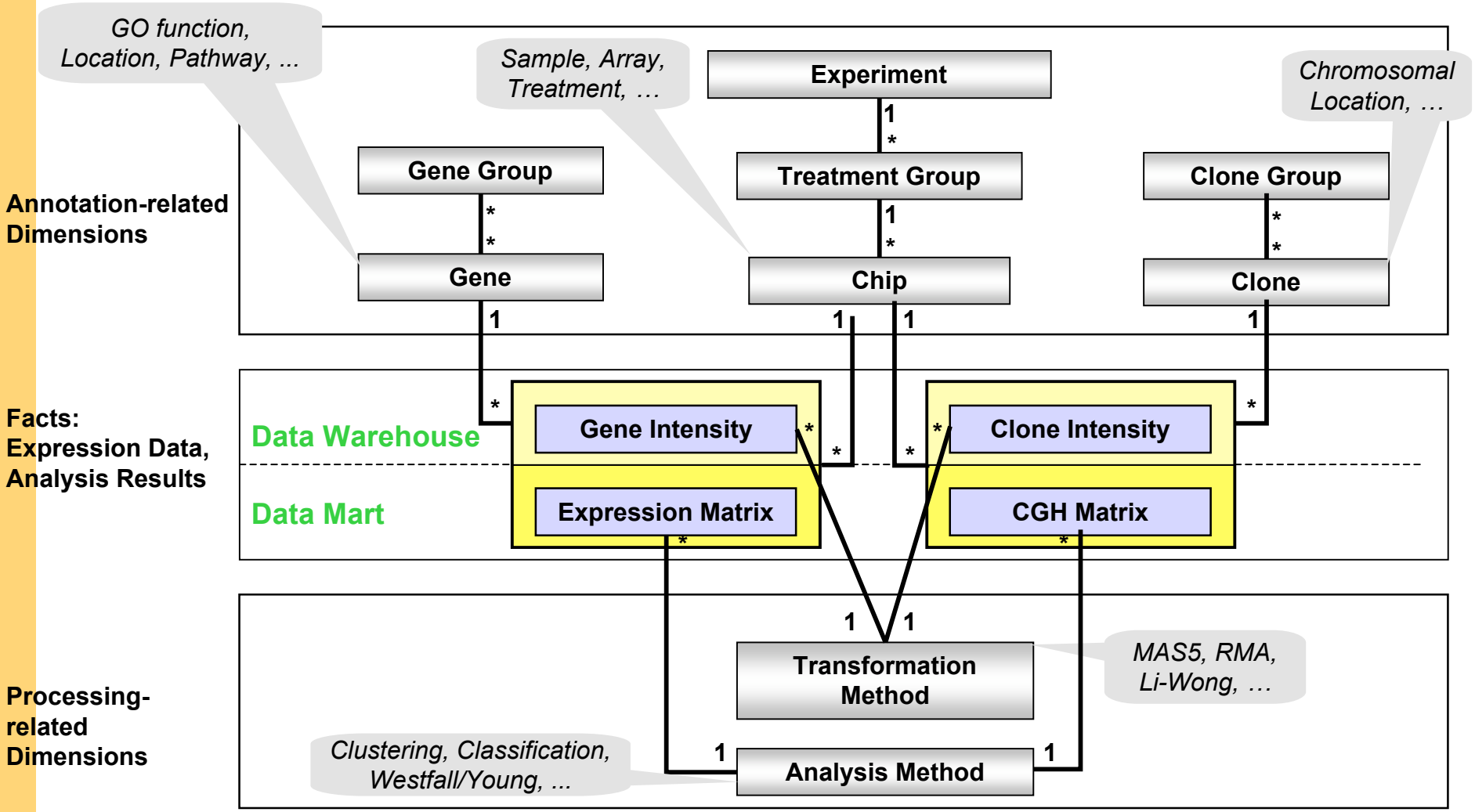


Multidimensional Data Management

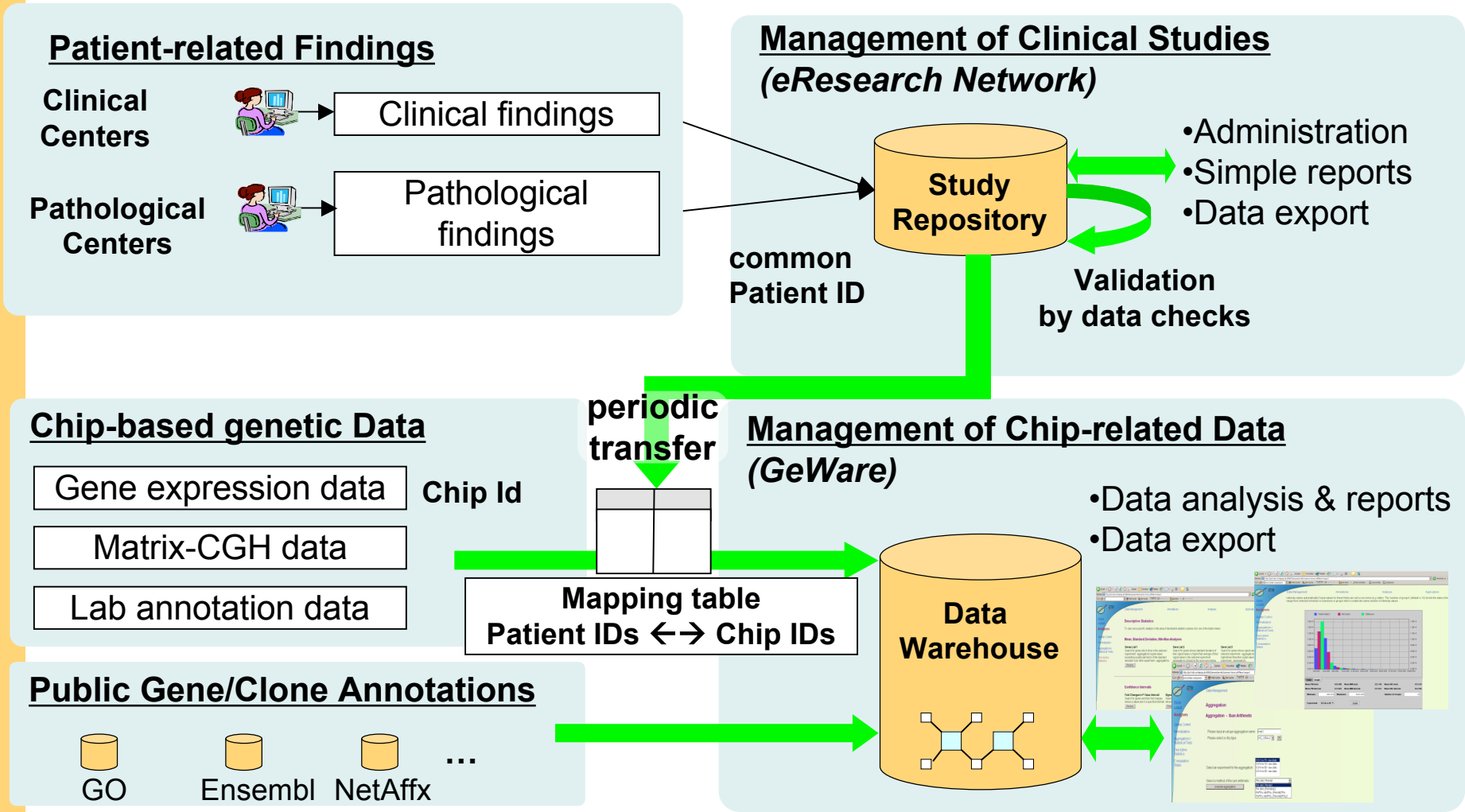
- Fact tables: expression values for different chip types and many chips
  - Scalability and extensibility
- Dimensions (chips/patients, genes, analysis methods)
- Multidimensional analysis
  - Easy selection, aggregation and comparison of values
- Basis to support more advanced analysis methods
  - Focused selection and creation of matrices



GeWare – Data Warehouse Model



Clinical data: integration architecture*

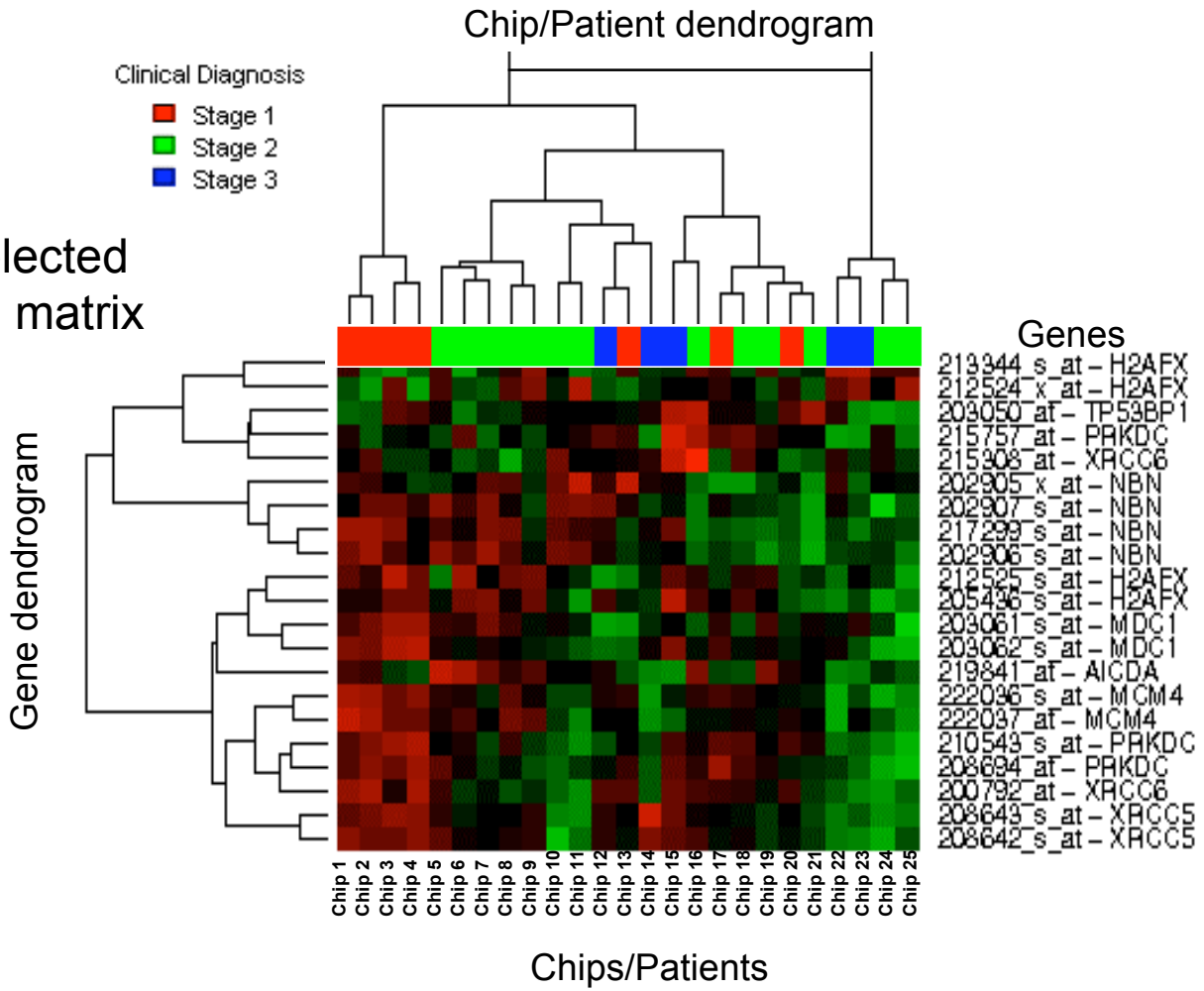


*Kirsten, T; Lange, J; Rahm, E : *An integrated platform for analyzing molecular-biological data within clinical studies.* Information Integration in Healthcare Application, LNCS 4254, 2006

Analysis example

- Visualizations of expression values using clinical data

Heatmap of a selected gene expression matrix

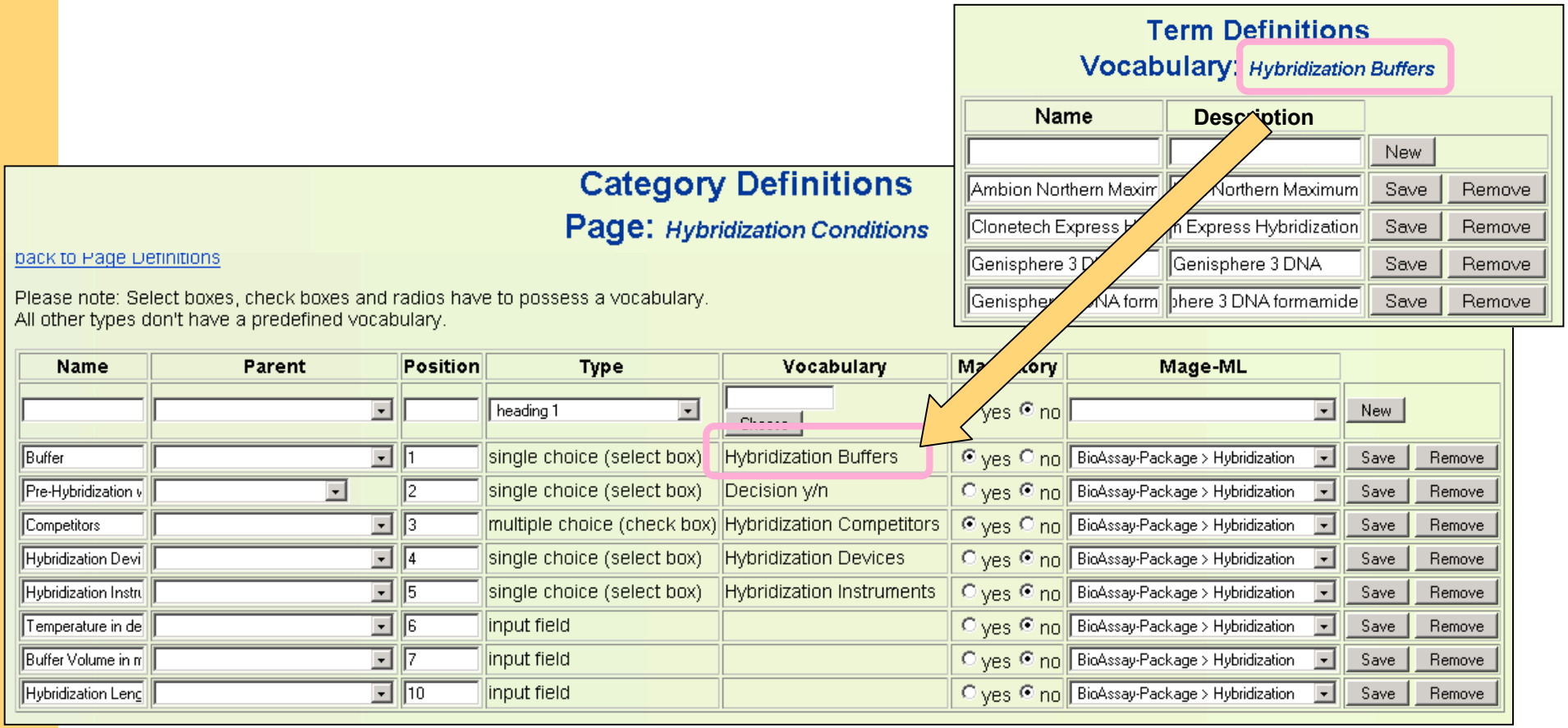


Annotation management

- Generic approach to specify structure and vocabulary for experimental, clinical and genetic annotations
- Consistent metadata instead of freetext or undocumented abbreviations and naming
- Manual specification of experimental annotations
  - describing the experimental set-up and procedure: sample modifications, hybridization process, utilized devices, …
- Automatic import of clinical annotations and genetic annotations
- Annotation templates:
  - collections of hierarchically structured annotation categories
  - permissible annotation values can be restricted to controlled vocabularies
  - MIAME compliant templates
- Controlled vocabularies: locally developed or external (e.g. NCBI Taxonomy)

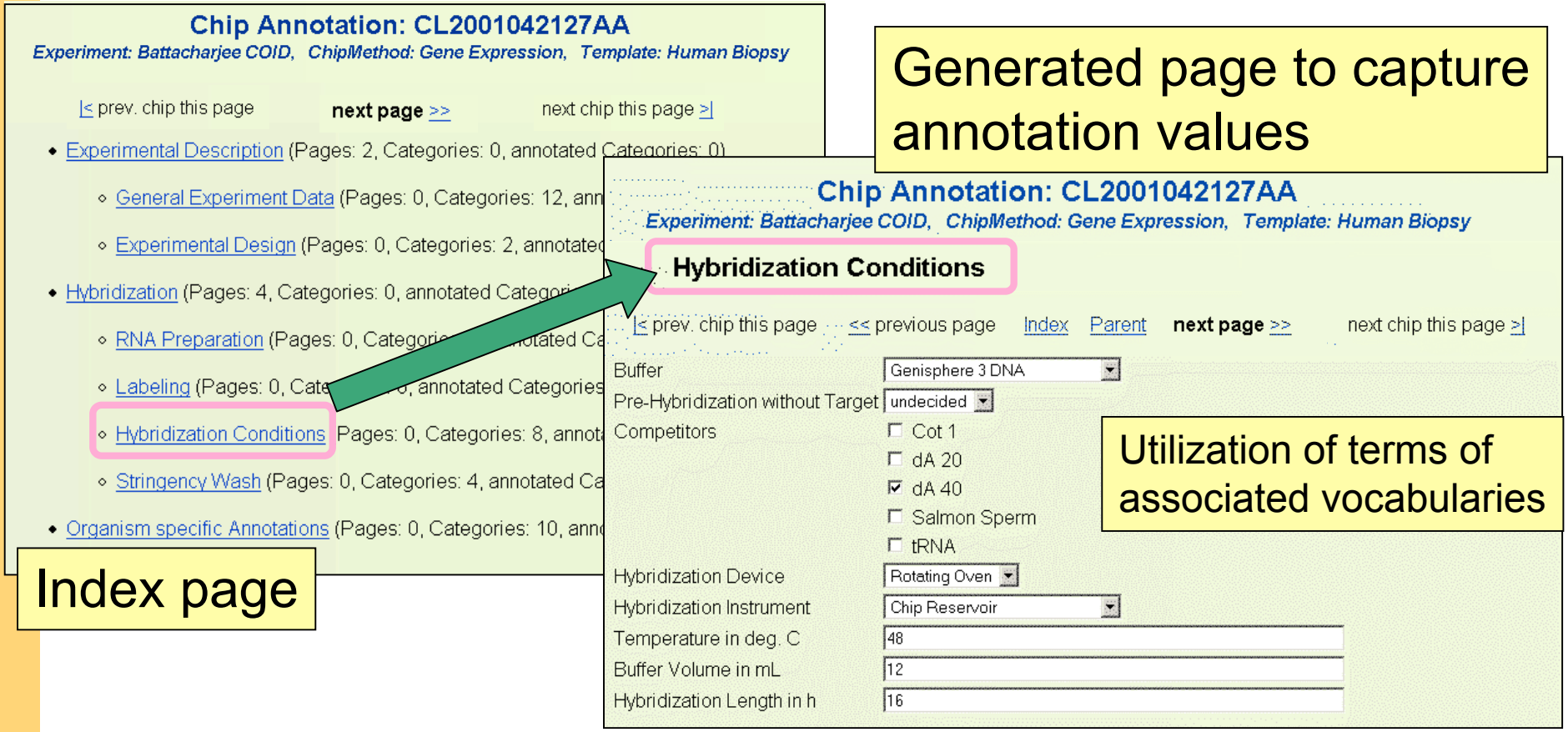
Experiment annotation: implementation (1)

- Template example
  - Easy specification and adaptation
  - Association of available vocabularies



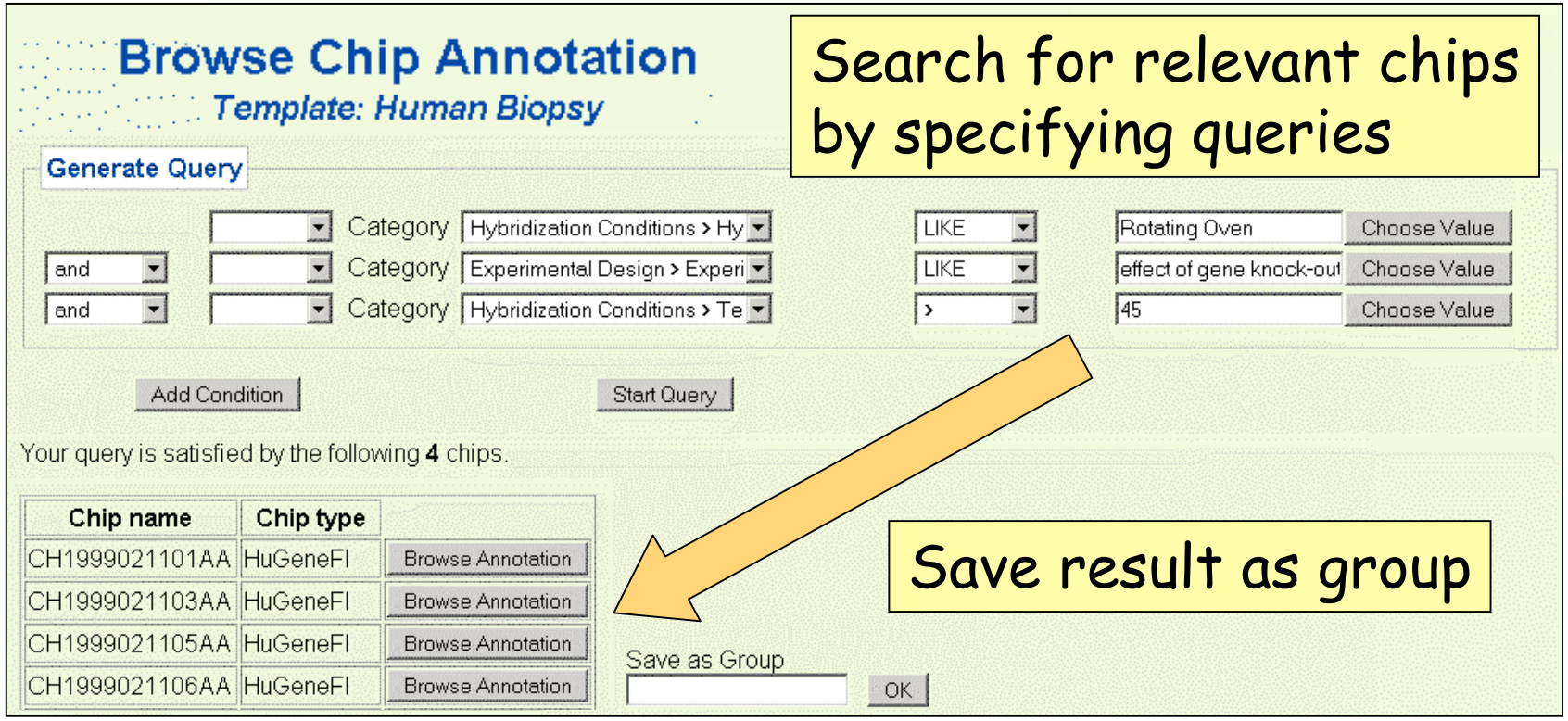
Experiment annotation: implementation (2)

- Template example
  - Automatically generated web GUI
  - Hierarchically ordered categories

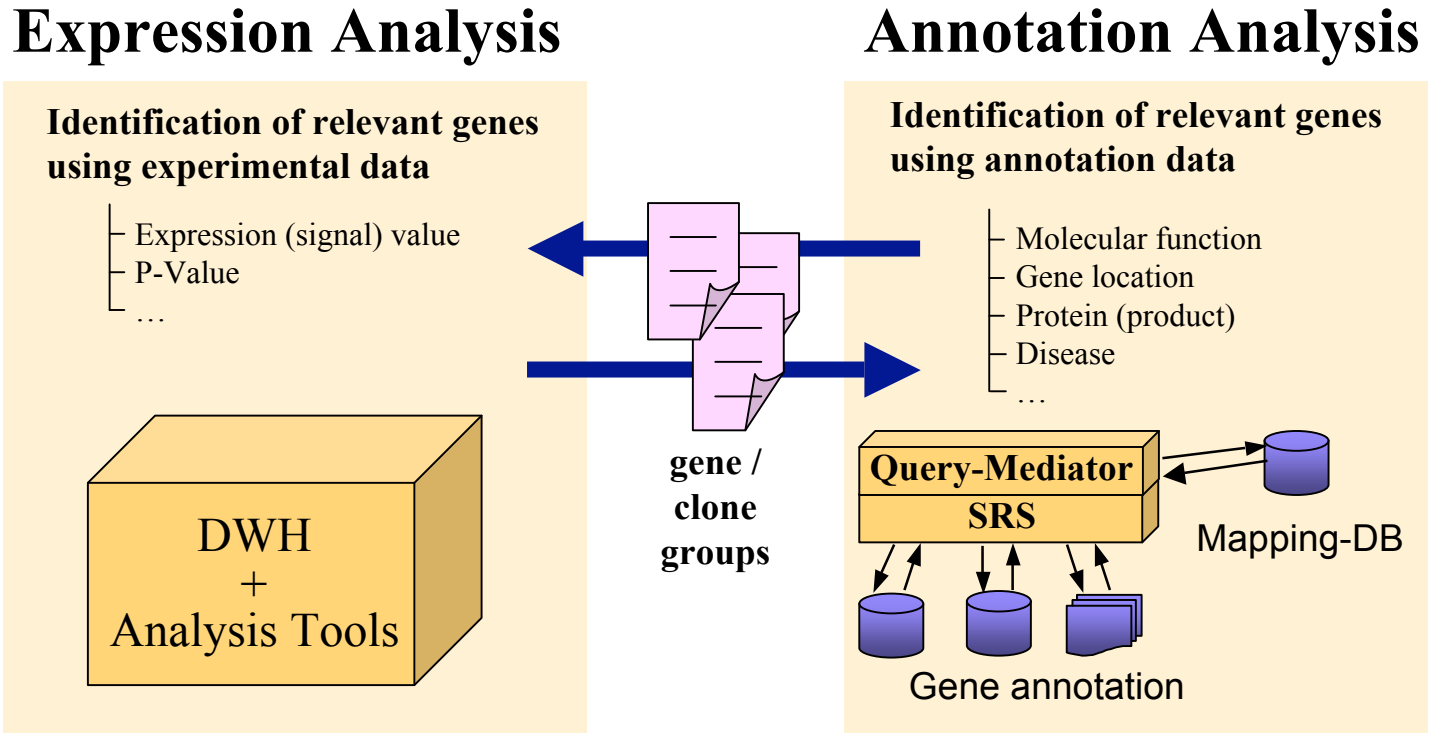


Experiment annotation: application

- Search in experiment annotation: Create treatment groups (later reuse in analysis)



Hybrid integration of data sources*



*Kirsten, T; Rahm, E: *Hybrid integration of molecular-biological annotation data*. Proc. 2nd Intl. Workshop DILS, July 2005

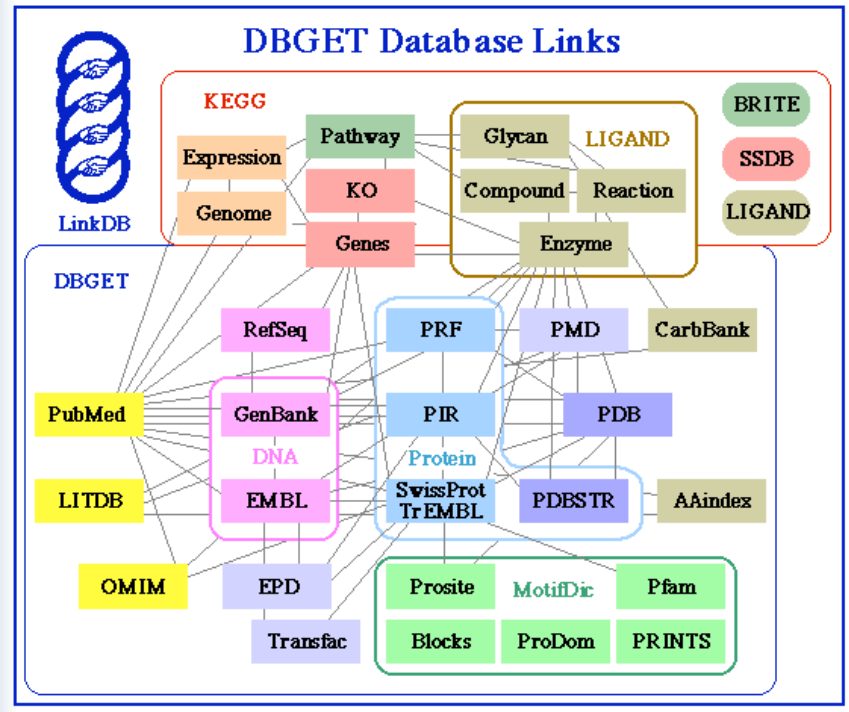
Agenda

- Kinds of data to be integrated
- General data integration alternatives
- Warehouse approaches
- Virtual and mapping-based data integration
  - Web-link integration: DBGet/LinkDB
  - GenMapper
  - Distributed Annotation System (DAS)
  - Sequence Retrieval System (SRS)
  - BioFuice
- Matching large life science ontologies
- Data quality aspects
- Conclusions and further challenges

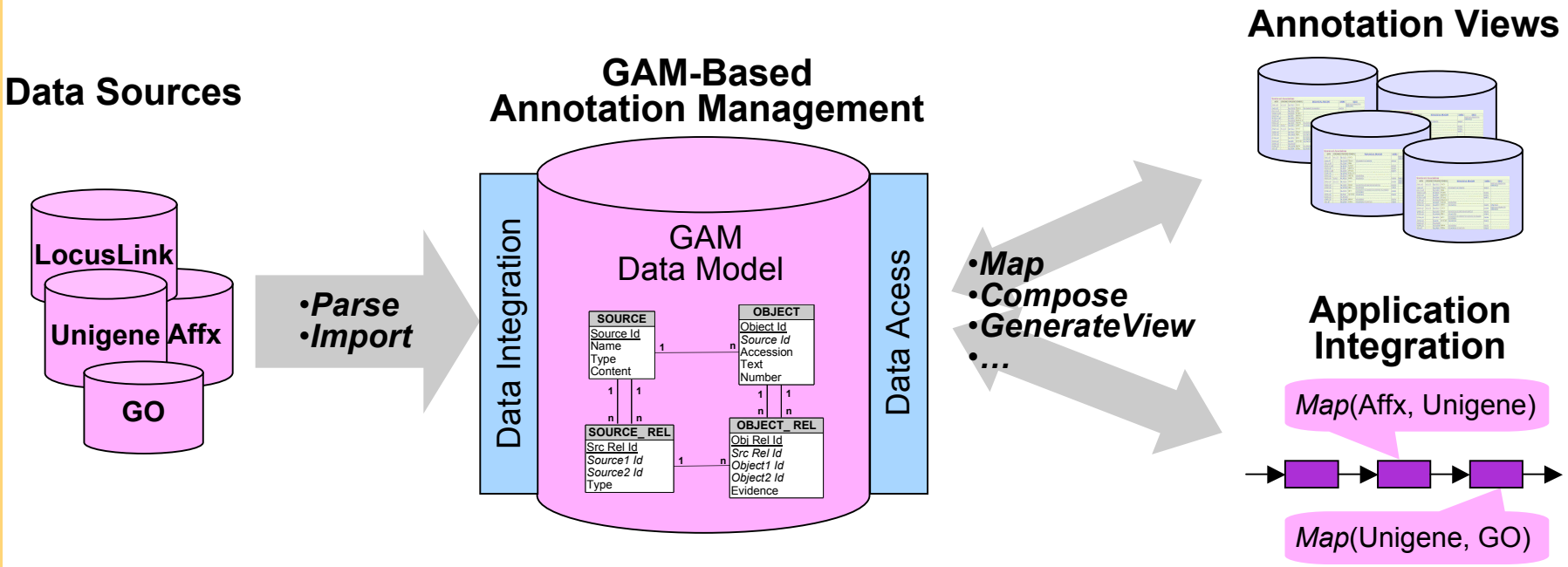
Integration based on available web-links

- Web-Link = URL of a source + ID of the object of interest
- Simple integration approach
  - Little integration effort
  - Scaleable
  - Navigational analysis: only one object at a time)
- DBGET + LinkDB:
  - Collection of web-links between many sources
  - Management of source specific sets of object ID and their connecting mappings
  - No explicit mapping types

www. genome.jp/dbget/



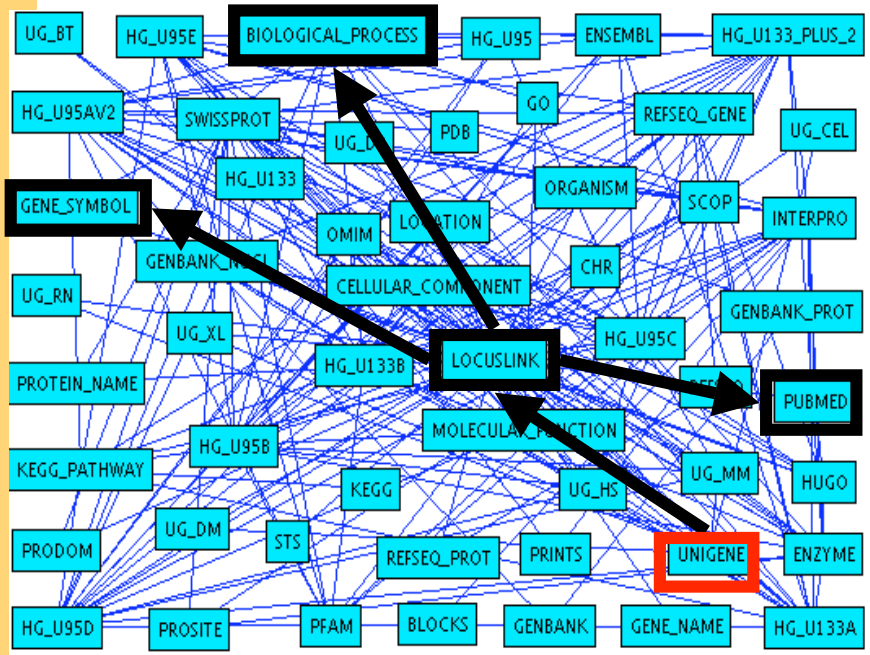
GenMapper*



- *Generic data model*, GAM, to uniformly represent annotation data
  - Flexible w.r.t. heterogeneity, evolution and integration
- Exploits *existing mappings* between objects/sources
  - Valuable knowledge, available in almost every source, scalable
- *High-level operations* to support data integration and data access
- *Tailored annotation views* for specific analysis needs

*Do, H.H.; Rahm, E.: *Flexible integration of molecular-biological annotation data: The GenMapper approach*. Proc. 9th EDBT Conf., 2004

GenMapper: Usage scenario

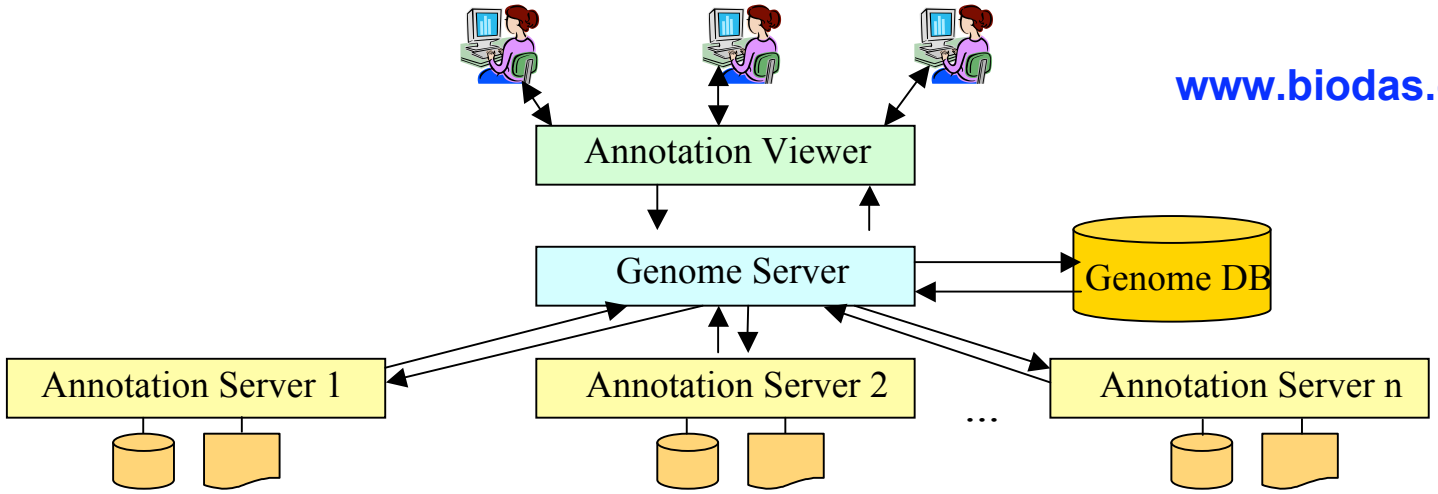


Annotation view

UG_HS	GENE_SYMBOL	BIOLOGICAL_PROCESS	PUBMED
Hs.100002	blp, Dncl2a, DNLC2A, BITH, HSPC162, MGC15113	GO:0007018, GO:0007632	
Hs.100007	RFX2, FLJ14226	GO:0006355	
Hs.100009	CDK3	GO:0000074, GO:0006468, GO:0007067	1639063
Hs.100057	STK35, CLIK1	GO:0006468	
Hs.100058	DPYSL4, CRMP3, DRP-4, ULIP4	GO:0007399	8973361, 9652388
Hs.100071	PGLS, 6PGL	GO:0005975, GO:0006098	10518023
Hs.100072	GJA12, Cx47, CX46.6, PMLDAR	GO:0007154	
Hs.100194	ALOX5AP, FLAP	GO:0006954, GO:0019370	1673682, 10036194

Distributed Annotation System (DAS)

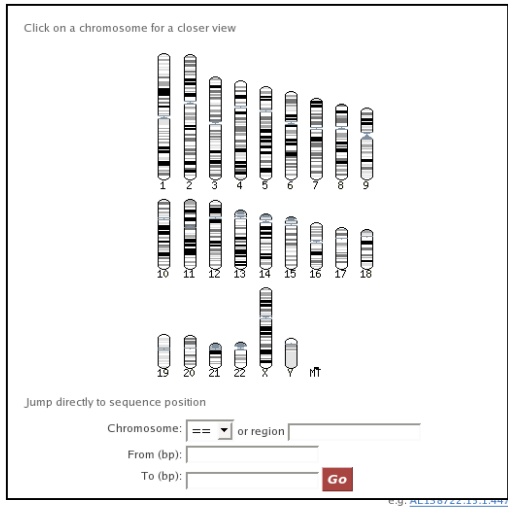
- Integration of distributed data sources with central genome server
  - Genome server: Primary source containing reference genome sequence
  - Annotation server: Wrapped source of a research group / organization
- Annotations are mapped to a reference genome sequence
  - Only sequence coordinates for each object are necessary (i.e., chr, start, stop, strand)
  - Simple and scaleable approach
  - Recalculation of all annotations when the reference sequence has changed

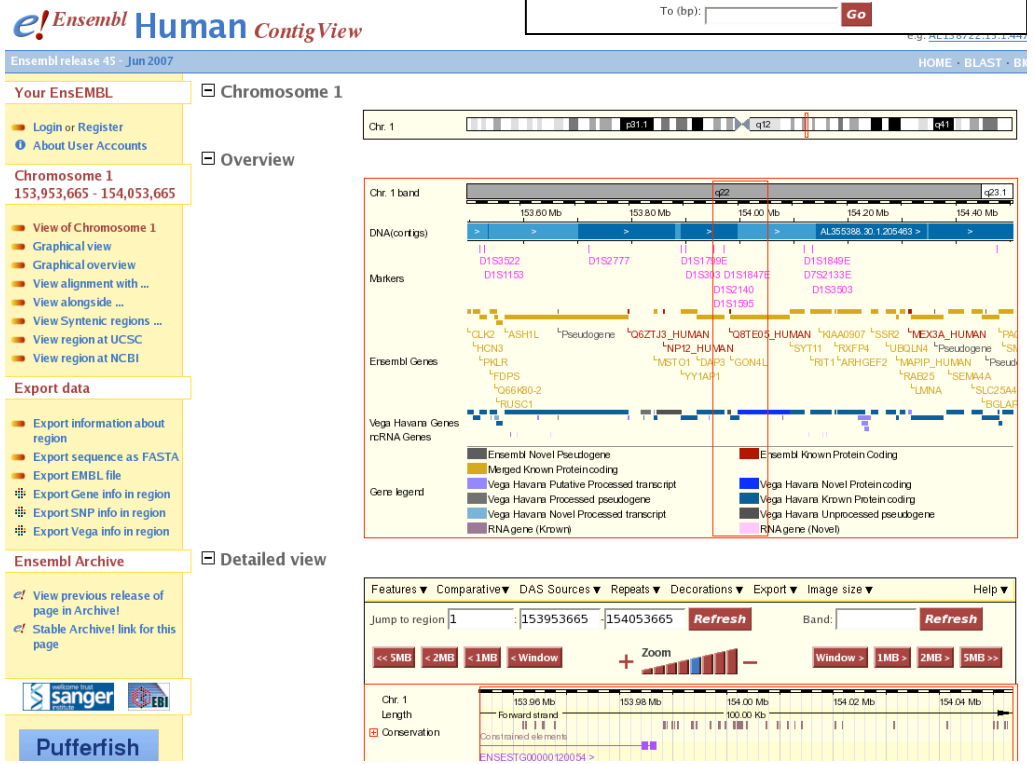


DAS: Query processing

- Query formulation
  - Select organism and chromosome from reference genome
  - Position-based (range) queries for associated objects
- Query processing
  - Send range query to genome DB and relevant annotation servers
  - Merge retrieved results
- Query result can be viewed

  on the genome at different detail levels with associated annotations, i.e., objects of different types

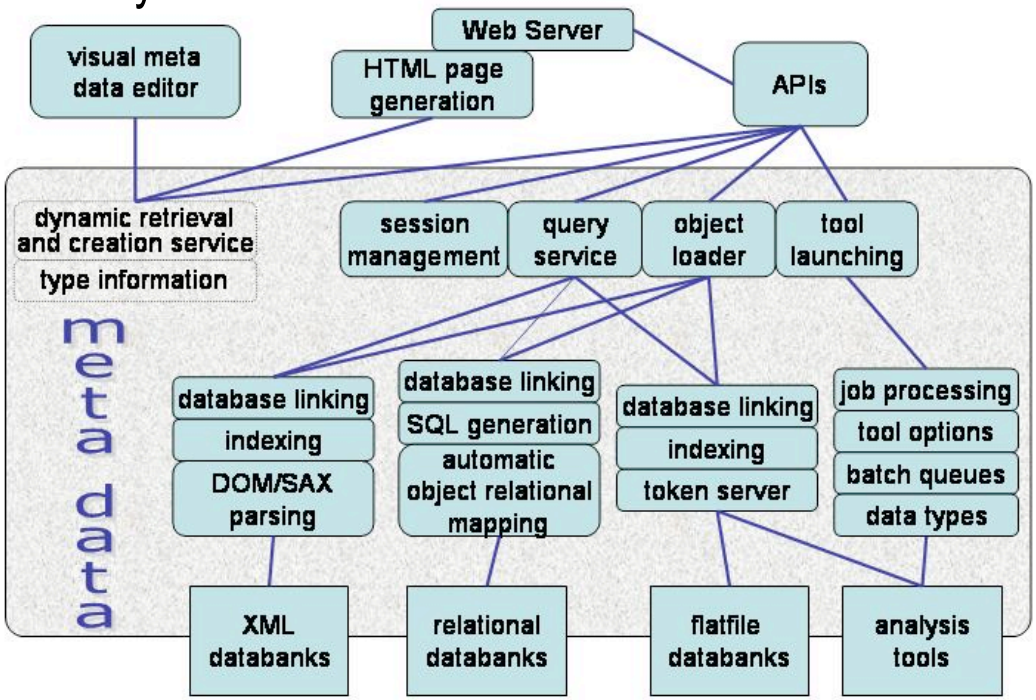




Sequence Retrieval System (SRS)

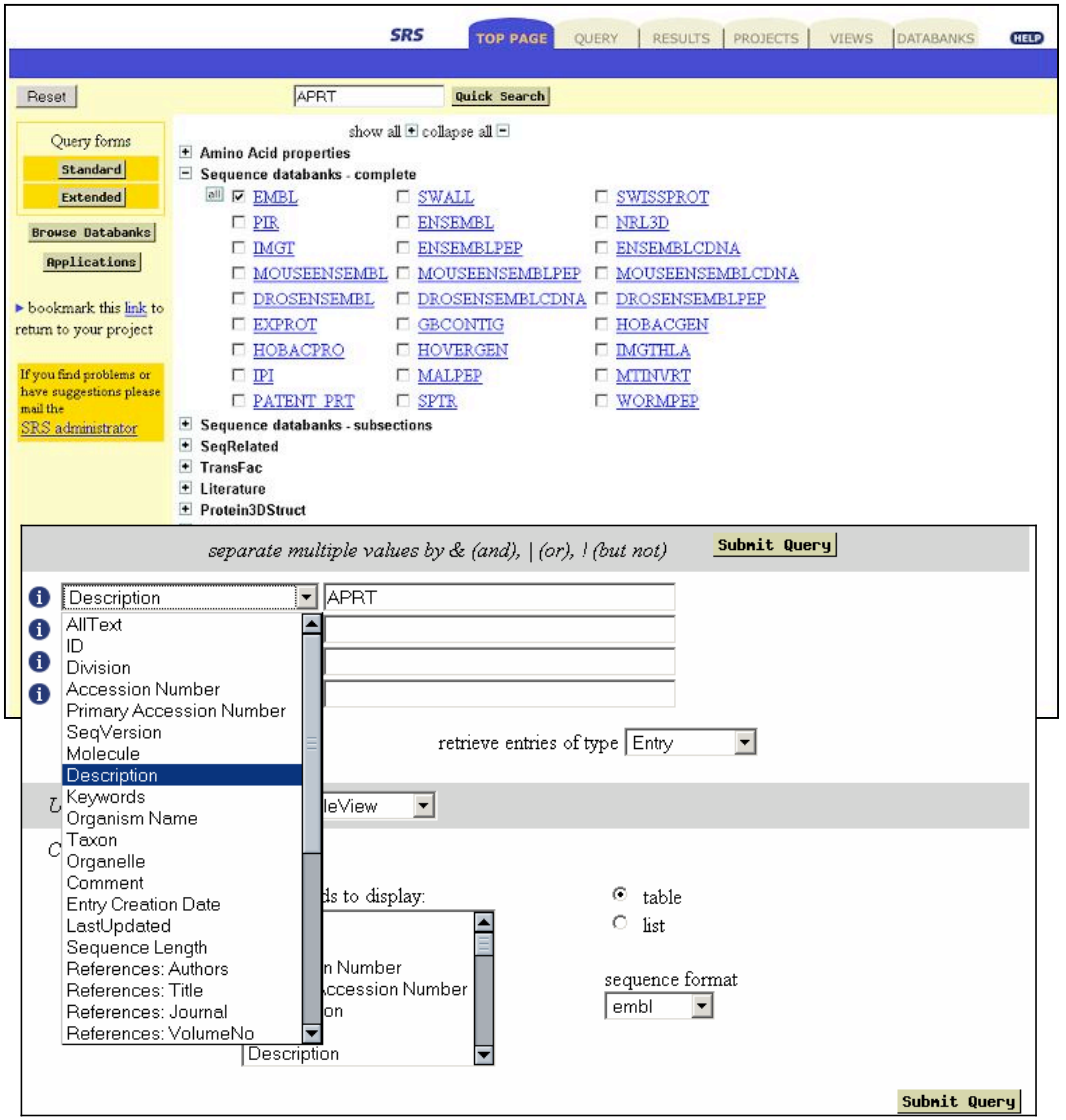
- Originally developed for accessing sequence data at EMBL
  - Commercial version by BioWisdom (before: Lion Bioscience)
- Data integration primarily for file data sources, but extended for database access and analysis tools
  - Mapping-based integration, no global schema
  - Local installation of sources necessary
  - Indexing (queryable attributes) of file-based sources by a proprietary script language
  - Definition of hub-tables (and queryable attributes) in relational sources
- Large wrapper library available for public sources

Source: Lion BioScience



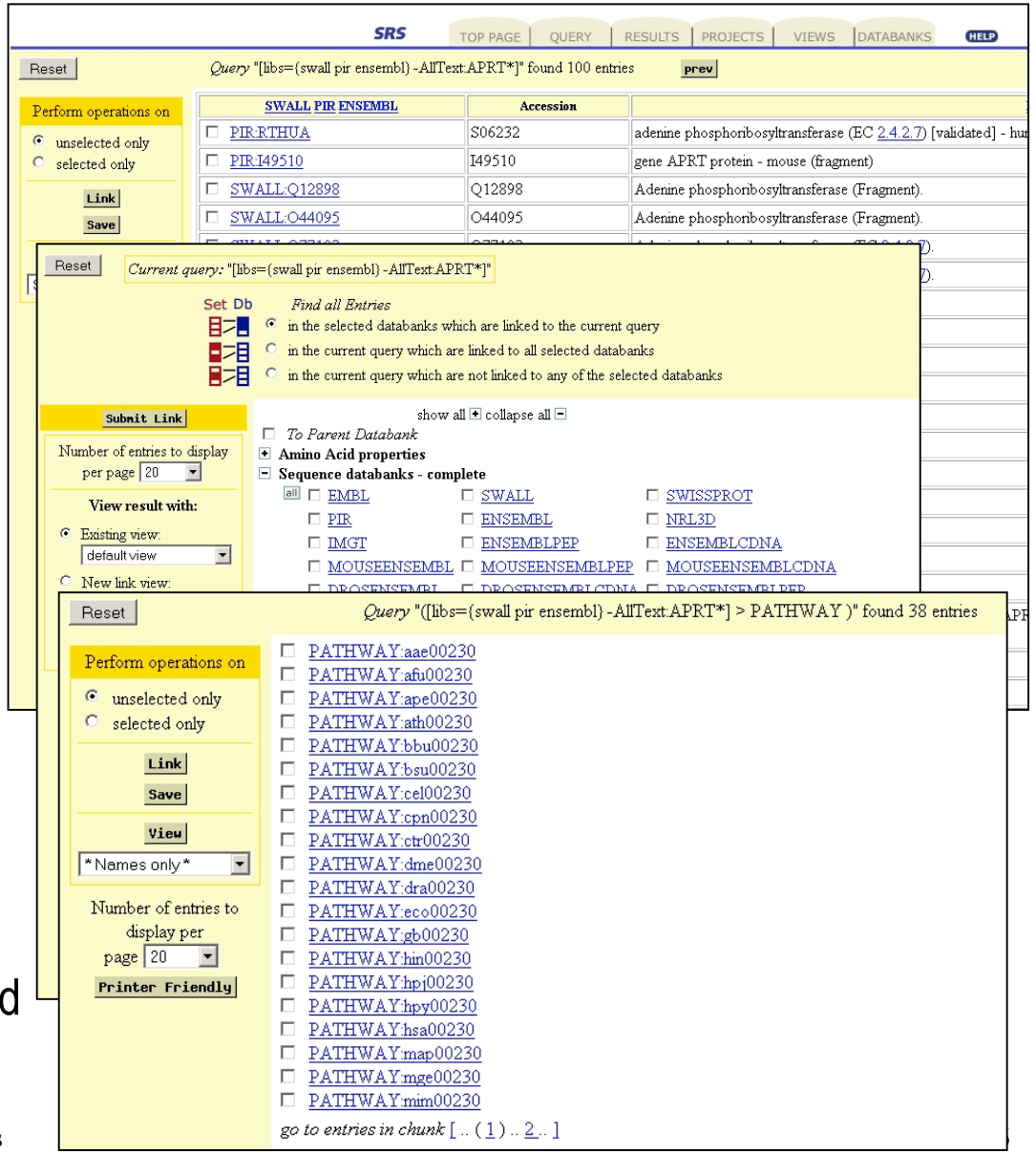
SRS: Query formulation and processing

- Query formulation
  - Source selection
  - Filter specification for queryable attributes
- Query types
  - Keyword search
  - Range search for numeric and date attributes
  - Regular expressions
- Automatic translation to SQL queries for relational sources
- Merge of result sets
  - Intersection
  - Union



SRS: Query formulation and processing cont.

- Explorative analysis
  - Traverse selected objects to objects of another data source
- Automatically generated paths between sources
  - Shortest paths (Dijkstra)
  - No consideration of path / mapping semantics
  - No join, only source graph traversal
- Result
  - Set of associated objects
  - No explicit mapping data (object correspondences) retrieved



Biofuice*: Design goals

- Utilization of instance-level cross-references (often manually curated, high quality data): instance-level mappings between sources
- Navigational access to many sources
- Support for queries and ad-hoc analysis workflows
- Often no full transparency necessary: users want to know from which sources data comes (data lineage / provenance)
- Support for integrating local (non-public) data
- Support for object matching and fusion (data quality)
- Creation of new instance mappings

-> Mapping-based data integration

*Kirsten, T; Rahm, E: *BioFuice: Mapping-based data integration in bioinformatics*. Proc. 3rd DILS, 2006

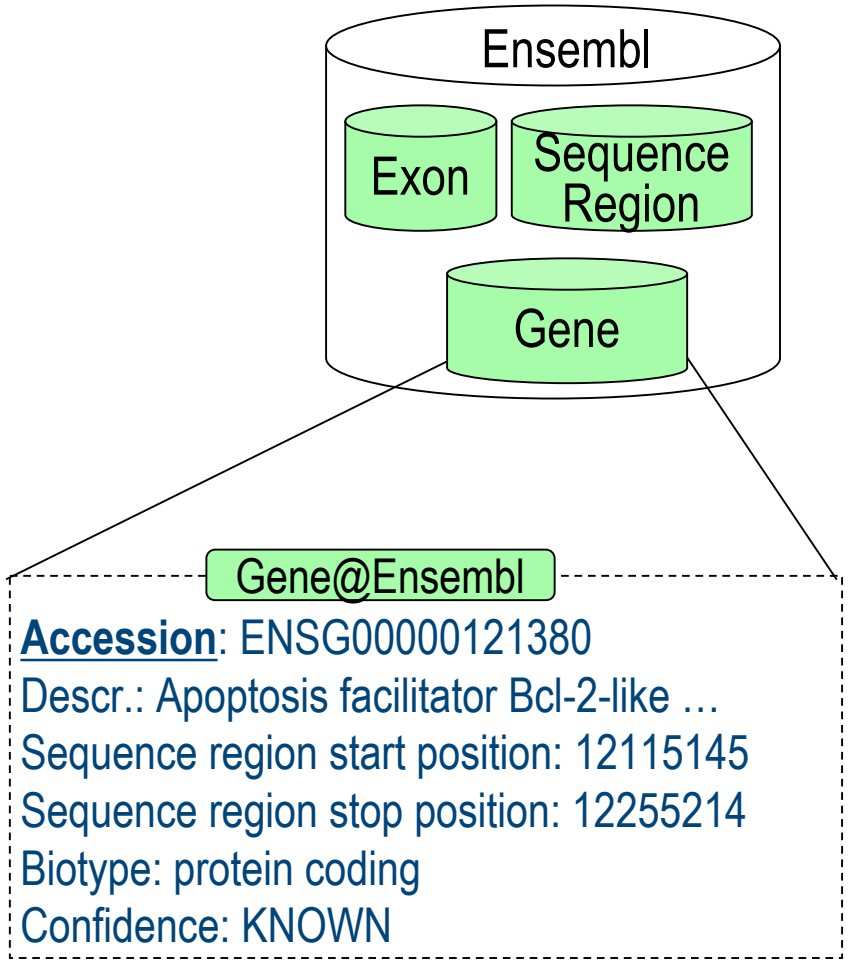
BioFuice (2)

- **BioFuice: Bioinformatics information fusion utilizing instance correspondences and peer mappings**
- Basis: iFuice approach*
  - Generic way to information fusion
  - **High-level operators**
- P2P-like infrastructure
  - Mappings between autonomous data sources (peers), e.g. sets of instance correspondences
  - Simple addition of new sources where they fit best
- Mapping mediator
  - Mapping management and operator execution
  - Downloadable sources are materialized for better performance (hybrid integration)
  - Utilization of application specific semantic domain model

* Rahm, E., et al.: *iFuice - Information Fusion utilizing Instance Correspondences and Peer Mappings.* Proc. 8th WebDB, Baltimore, June 2005

BioFuice: Data sources

- **Physical data source (PDS)**
  - Public, private and local data (gene list, …), ontologies
  - Splitted into logical data sources
- **Logical data source (LDS)**
  - Refers to one object type and a physical data source, e.g. Gene@Ensembl
  - Contains object instances
- Object instances
  - Set of relevant attributes
  - One **id** attribute

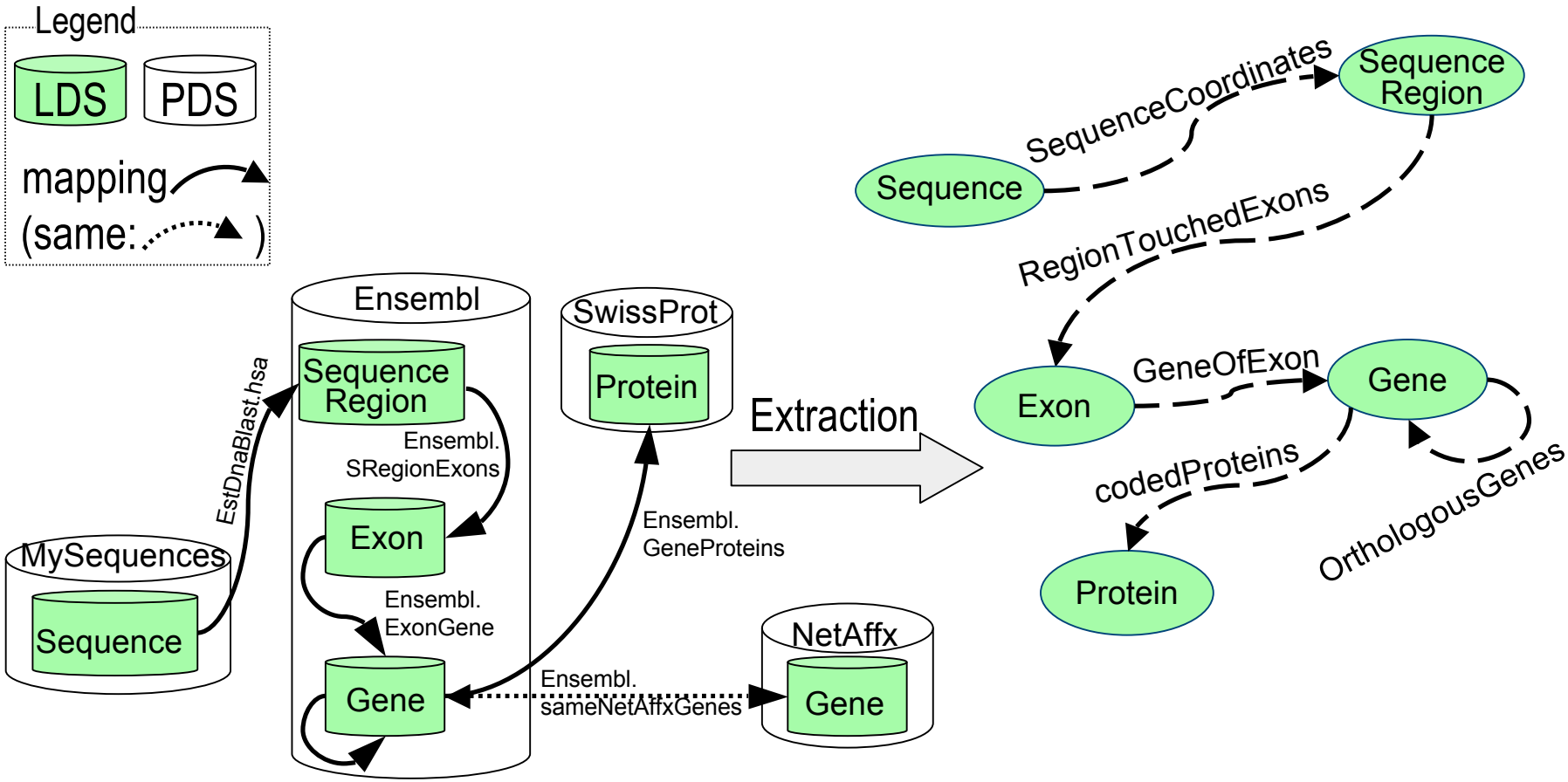


BioFuice Mappings

- Directed relationships between LDS
- Mappings have a semantic mapping type
  - E.g. OrthologousGenes
- Different kinds of mappings
  - Same mappings vs. Association mappings
    - Same: equality relationship
  - ID mappings vs. computed mappings (e.g. query mappings)
  - Materialized mappings (mapping tables) vs. dynamic generation (on the fly)

BioFuice: metadata models

- Used by mediator for mapping/operator execution
- **Domain model** indicates available object types and relationships



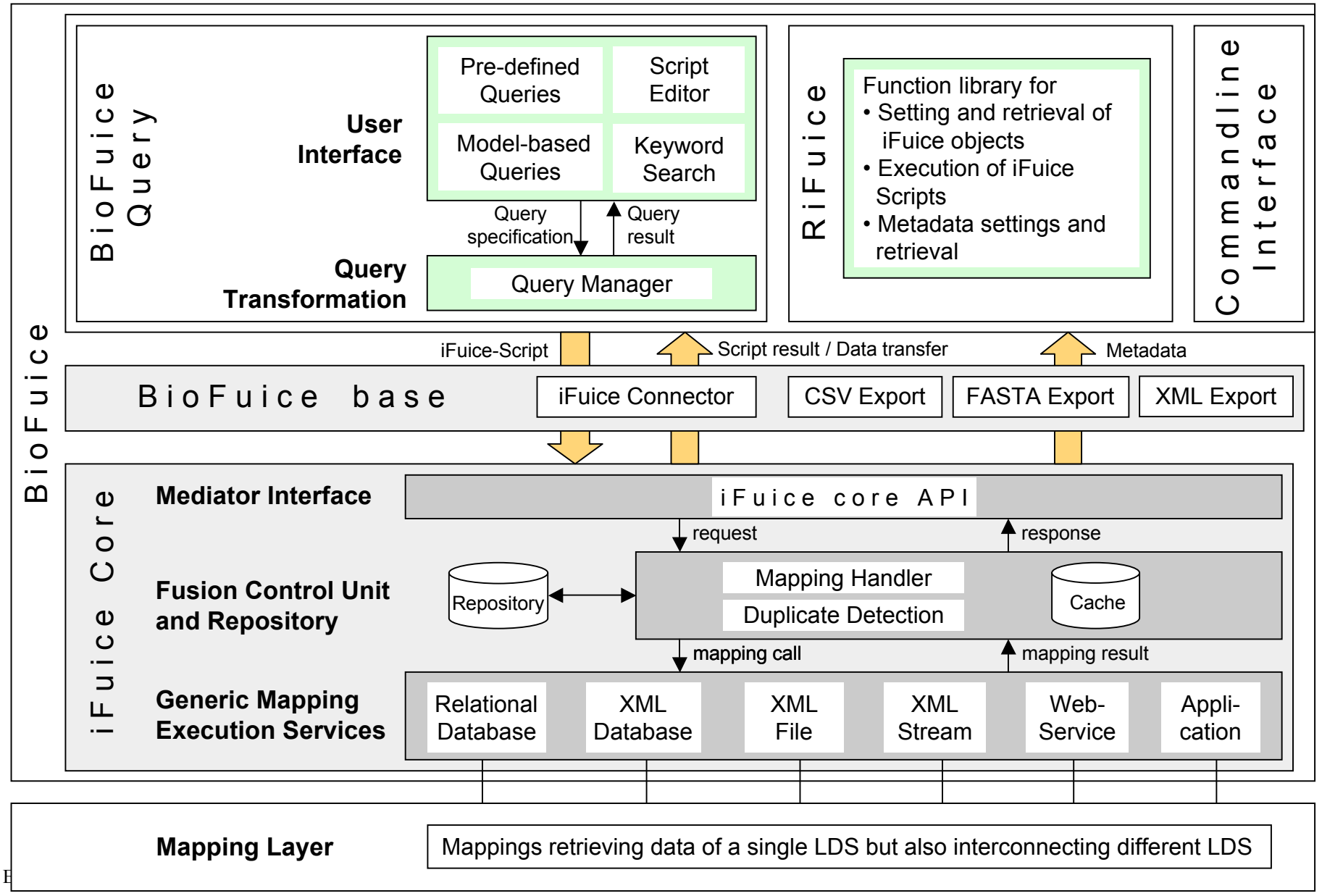
BioFuice Operators

- Query capabilities + scripting support
- **Set oriented operators**
  - Input: Set of objects/mappings
    + parameters / query conditions
  - Output: Set of resulting objects

⟹ Combination of operators within scripts for workflow-like execution

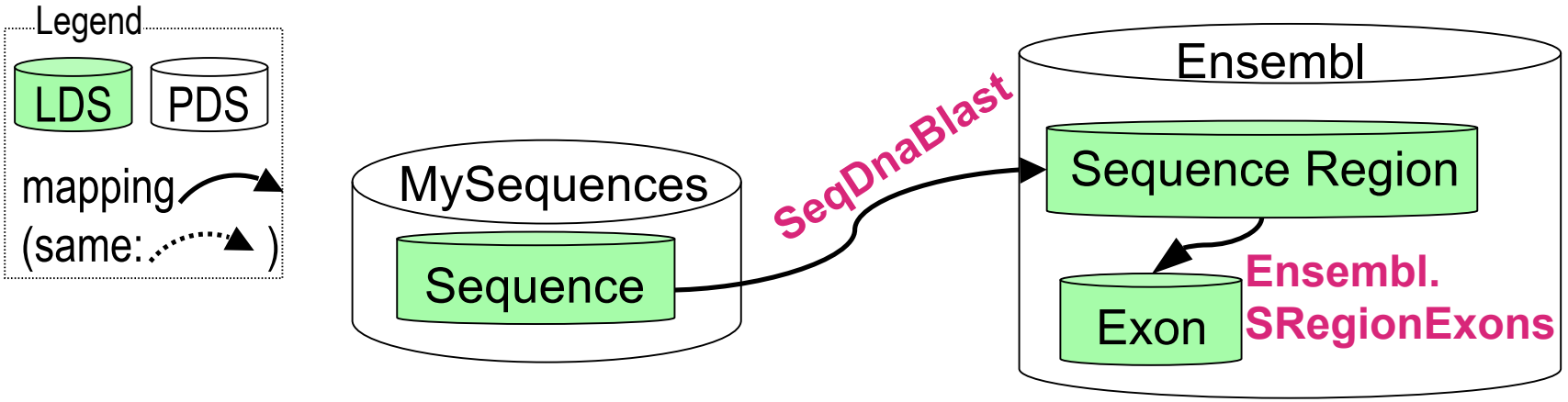
- Selected operators:
  - Single source: **queryInstances, searchInstances, …**
  - Navigation: **traverse, map, compose, …**
  - Navigation + aggregation: **aggregate, aggregateTraverse, …**
  - Generic: **diff, union, intersect, …**

BioFuice architecture



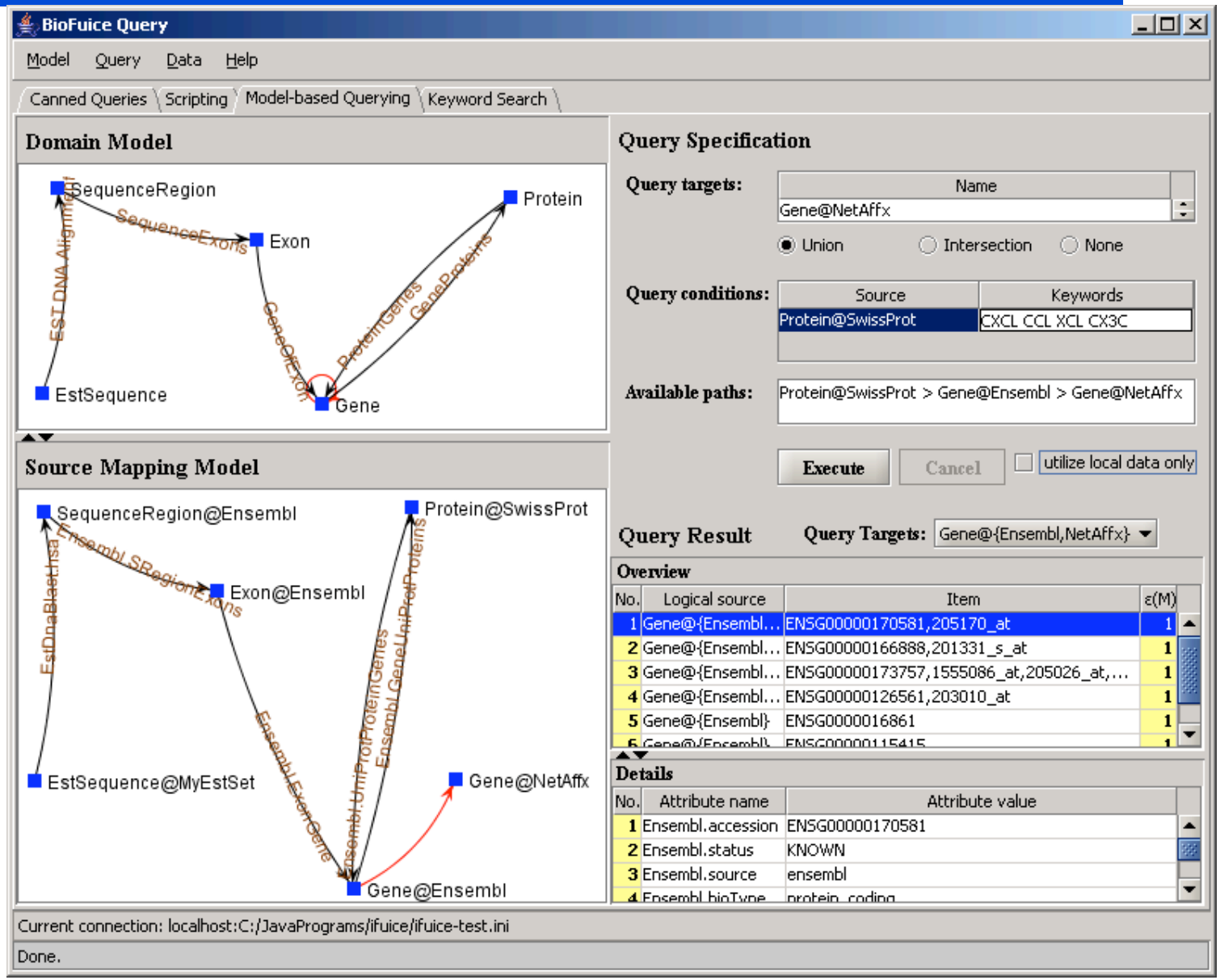
BioFuice: Script example

- Scenario
  - Given: Set of sequences in local source MySequences
  - Wanted: Three classes: unaligned s., non-coding s., protein coding sequences




```
$alignedSeqMR := map( MySequences, { SeqDnaBlast } );
$unalignedSeqOI    := diff ( MySequences, domain ( $alignedSeqMR ));
$codingSeqMR  := compose( $alignedSeqMR, { Ensembl.SRegionExons } );
$protCodingSeqOI  := domain ( $codingSeqMR );
$nonCodingSeqOI  := diff ( domain ( $alignedSeqMR ) , $protCodingSeqOI );
```

BioFuice Query Processing



iFuice application: citation analysis*

- Citation analysis important for evaluating scientific impact of publications venues, researchers, universities etc.
  - What are the most cited papers of journal X or conference Y?
  - What is the H-index of author Z ?
  - Frequent changes: new publications & new citations
- Idea: Combine publication lists, e.g. from DBLP or Pubmed, with citation counts, e.g from Google Scholar, Citeseer or Scopus
- Warehousing approach, virtual (on the fly) or hybrid integration
- Fast approximate results by Online Citation Service (OCS)**
  - **http:// labs.dbs.uni-leipzig.de/ocs**

* Rahm, E, Thor, A.: *Citation analysis of database publications*. ACM Sigmod Record, 2005

** Thor, A., Aumueller, D., Rahm, E.: *Data integration support for Mashups*. Proc. IIWeb 2007

Sample OCS result

OCS result for venue Bioinformatics 2004

Title	Year	Authors
80%	+/- two years	50%
85%	+/- one year	60%
90%	Equal year	70%
95%		80%
100%		90%
		100%

- Found 358 GS publications for 336 DBLP publications.
- No GS publications found for 225 DBLP publications.
- Overall: 561 DBLP publications having 7448 citations.
- Average: 13,3 citations per publication.
- H-Index: 41
- Match configuration: 80% title similarity, max. 0 year(s) difference, 50% author similarity.

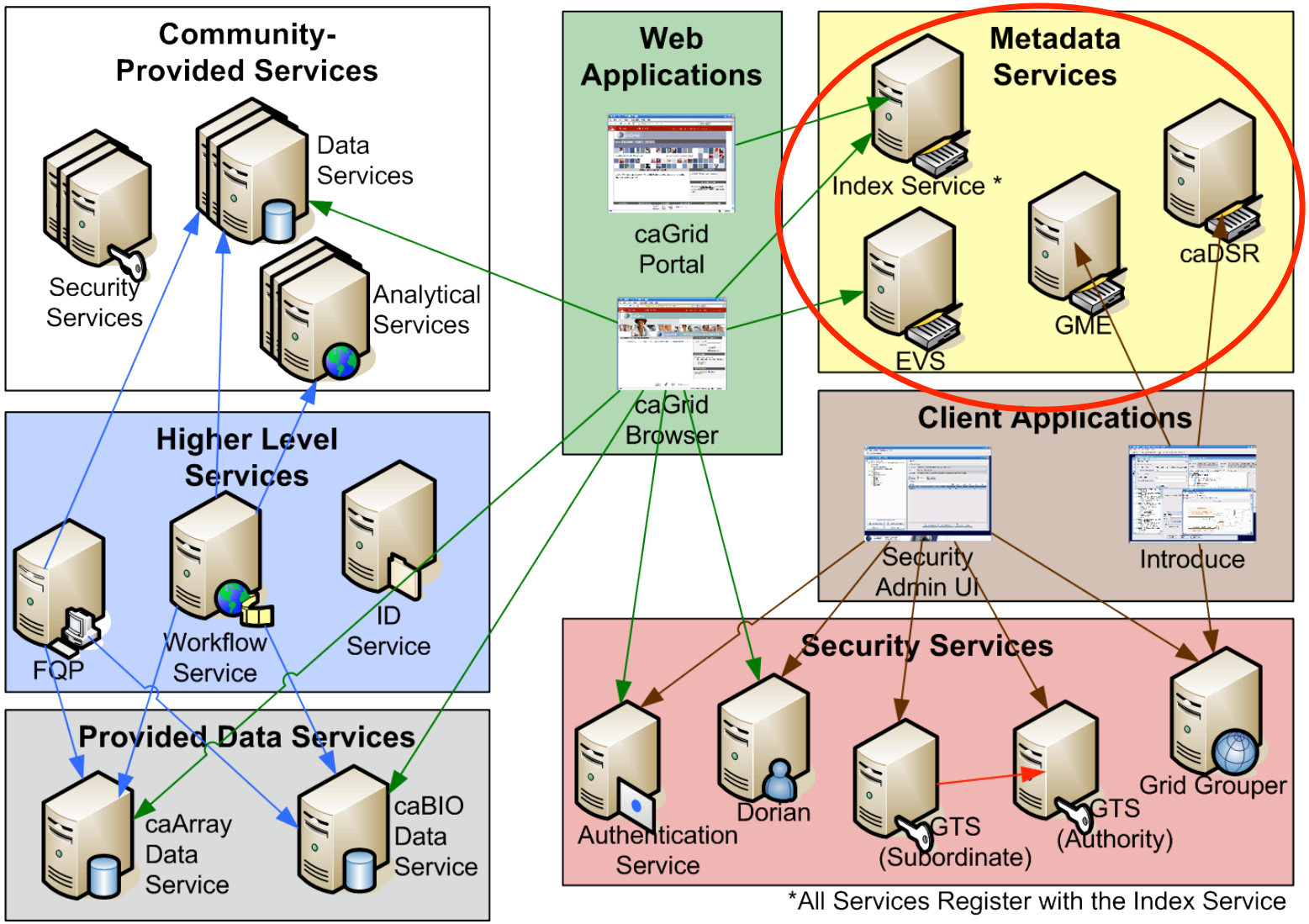
	Title	Authors	Venue	Year	Citation ▼
✚	FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes.	Fátima Al-Shahrour, Ramón Díaz-Uriarte, Joaquín Dopazo	Bioinformatics	2004	289
▬	Taverna: a tool for the composition and enactment of bioinformatics workflows. 215 T Oinn, M Addis, J Ferris, D Marvin, M Senger, M : *Taverna: a tool for the composition and enactment of bioinformatics workflows* (2004) 2 T Oinn, M Addis, J Ferris: *other authors (2004). Taverna: a tool for the composition and enactment of bioinformatics workflows* 2 T Oinn, M Addis, J Ferris, D Marvin, M Senger, M : , *A. Wipat, and P. Li. Taverna: a tool for the composition and enactment of bioinformatics workflows*	Thomas M. Oinn, Matthew Addis, Justin Ferris, Darren Marvin, Martin Senger, R. Mark Greenwood, Tim Carver, Kevin Glover, Matthew R. Pocock, Anil Wipat, Peter Li	Bioinformatics	2004	219
✚	The Jalview Java alignment editor.	Michele E. Clamp, James A. Cuff, Stephen M. J. Searle, Geoffrey J.	Bioinformatics	2004	212

caBIG™/caGRID*

- cancer Biomedical Informatics Grid™ (caBIG™)
  - Virtual network connecting individuals and organizations to enable the sharing of data and tools, creating a World Wide Web of cancer research
  - Overall goal: Speed the delivery of innovative approaches for the prevention and treatment of cancer
- Objectives
  - Common, widely distributed infrastructure that permits the cancer research community to focus on innovation
  - Service-based integration of applications and data
  - Shared, harmonized set of terminology, data elements, and data models that facilitate information exchange to overcome syntactic and semantic interoperability
  - Collection of interoperable applications developed to common standards
  - Raw published cancer research data is available for mining and integration

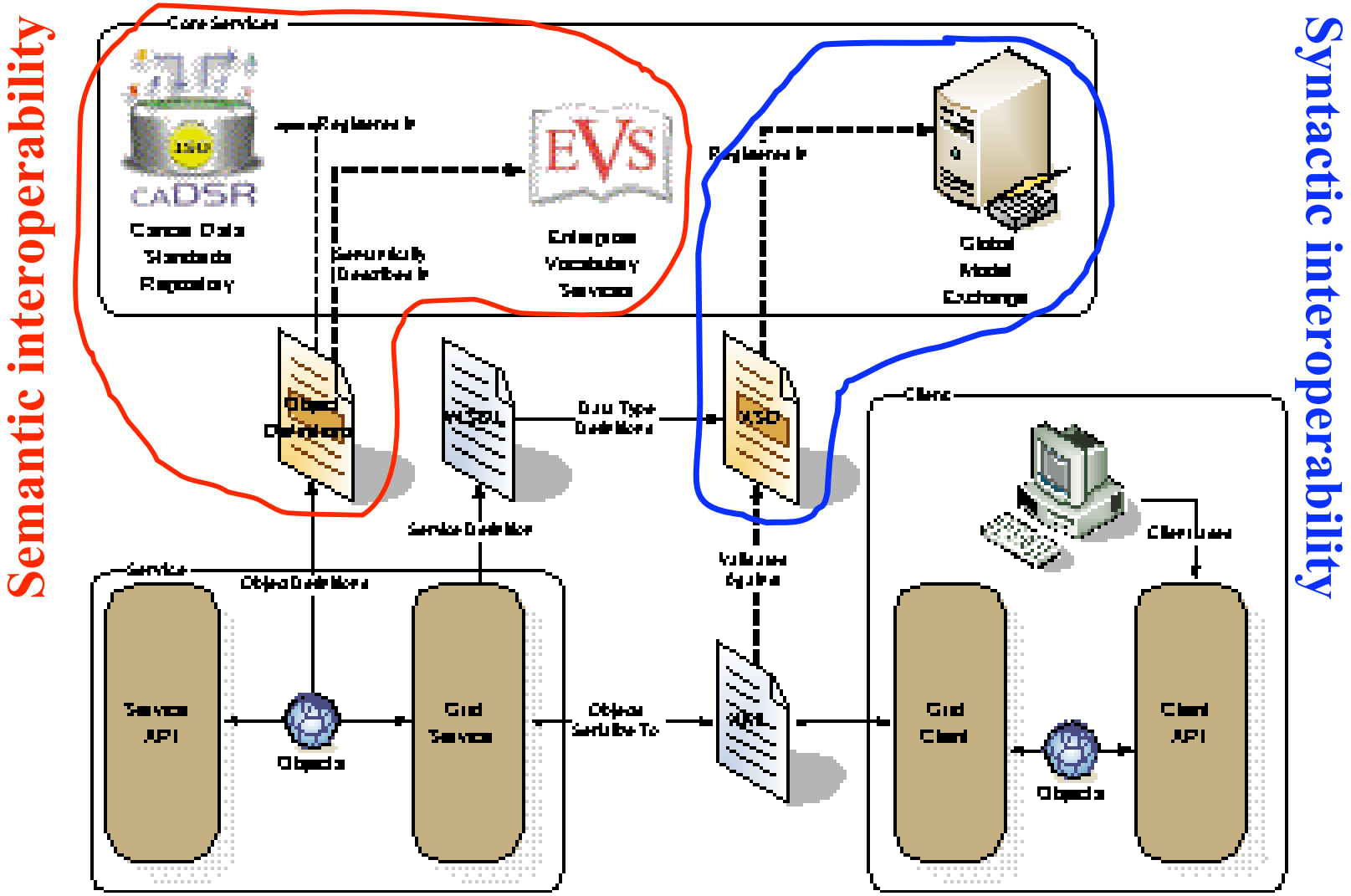
*Joel H. Saltz, et al.: *caGrid: design and implementation of the core architecture of the cancer biomedical informatics grid.* Bioinformatics, Vol. 22, No. 15, 2006, pp. 1910-1916

Service-based data integration in caGrid



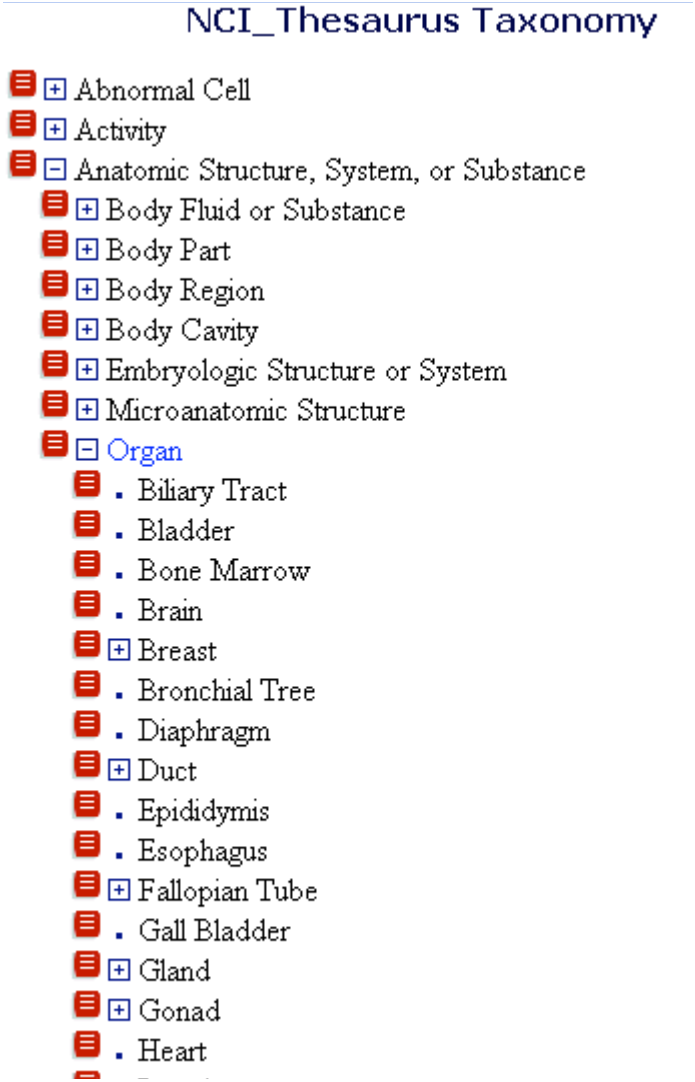
Source: T. Kurc et al.: Panel Discussion, caBIG Annual Meeting 2007

caBIG/caGRID: Data description infrastructure

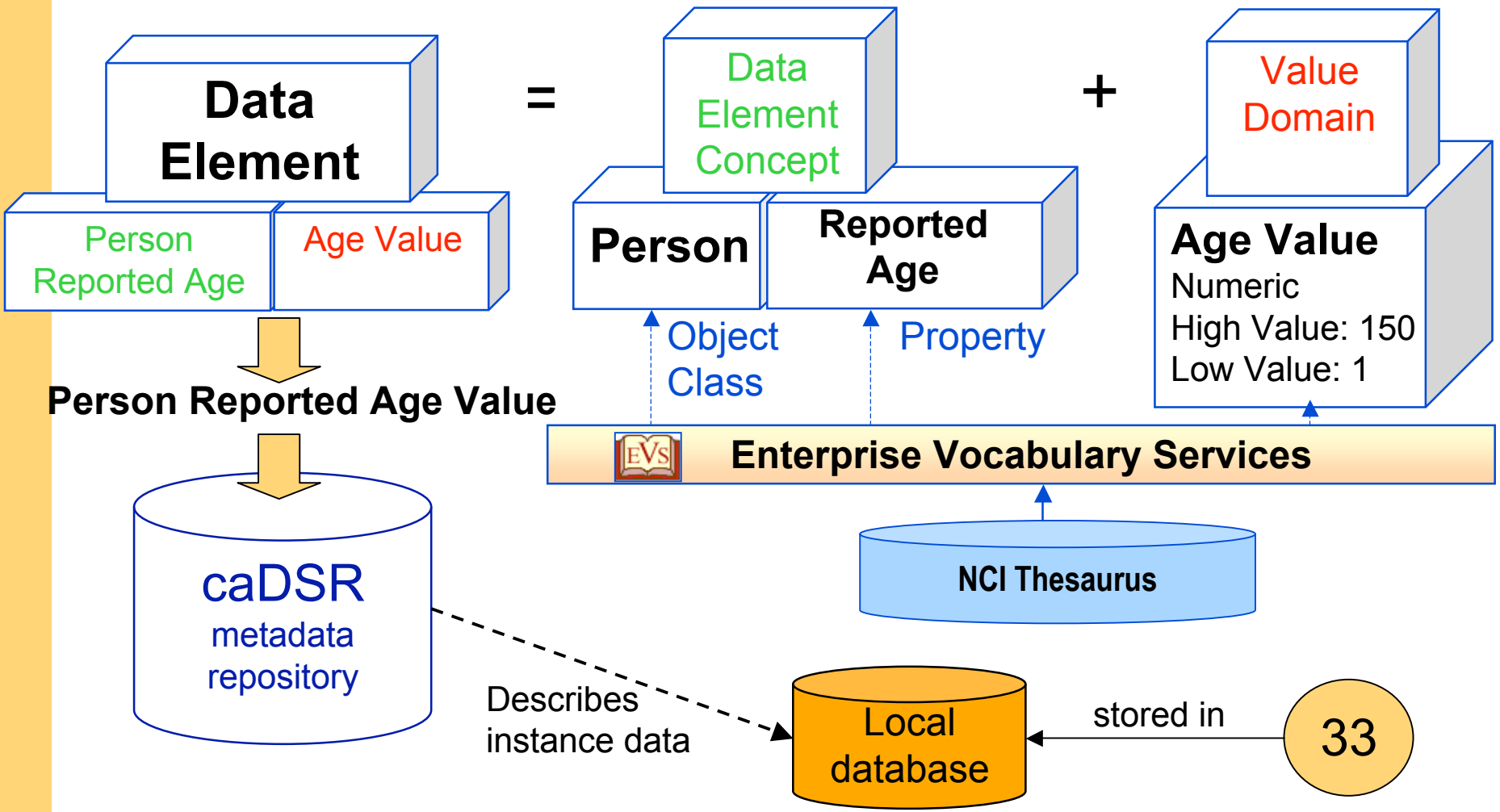


caBIG/caGRID: Basis Vocabulary -NCI Thesaurus

- About NCI Thesaurus
  - Reference terminology for NCI
  - About 54000 concepts in 20 hierarchies
  - Broad coverage of cancer domain
    - Findings and Disorders
    - Anatomy
    - Drugs, Chemicals
    - Administrative Concepts
    - Conceptual Entities/Data Types
- Advantages
  - Uniform conceptualization in a domain
  - Standardization, interoperability, classification
  - Enable reuse of data and information
- Usage in caBIG/caGrid
  - Annotation of medical data (images, …)
  - Service Discovery in grids
  - Building of Common Data Elements (CDE) for exchange of medical data



caBIG/caGRID: Building common data elements



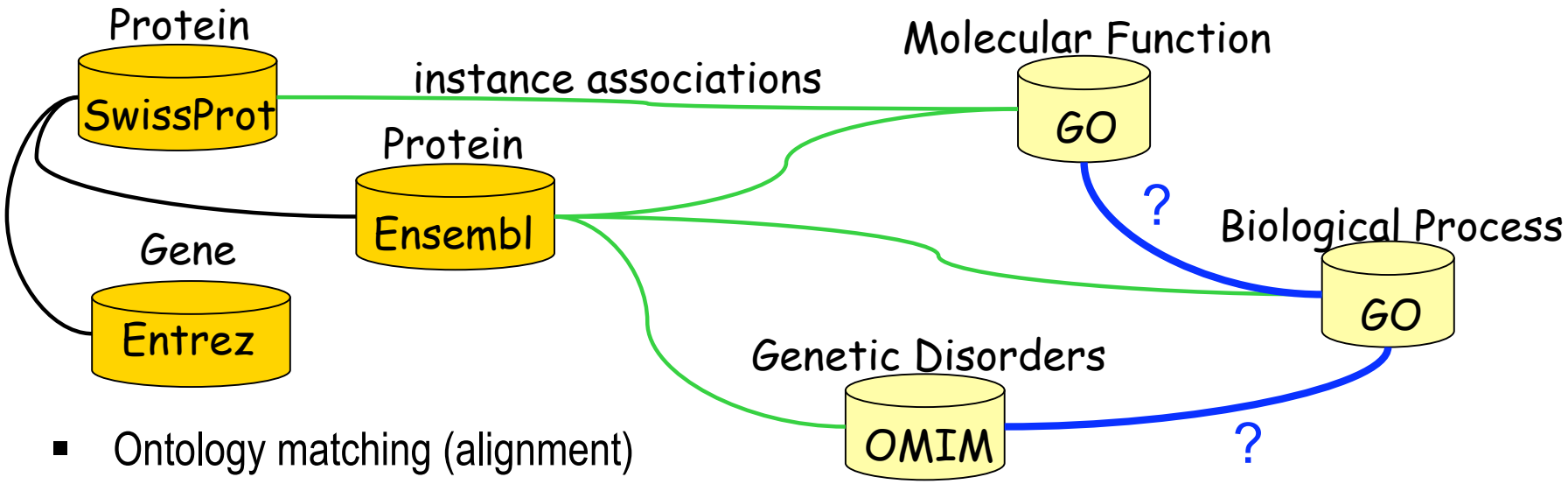
Source: caDSR & ISO 11179 Training - Jennifer Brush, Dianne Reeves

Agenda

- Kinds of data to be integrated
- General data integration alternatives
- Warehouse approaches
- Virtual and mapping-based data integration
- Matching large life science ontologies
  - Motivation
  - Match approaches and frameworks (Coma++, Prompt, Sambo)
  - Instance-based match approach (DILS07), evaluation results
- Data quality aspects
- Conclusions and further challenges

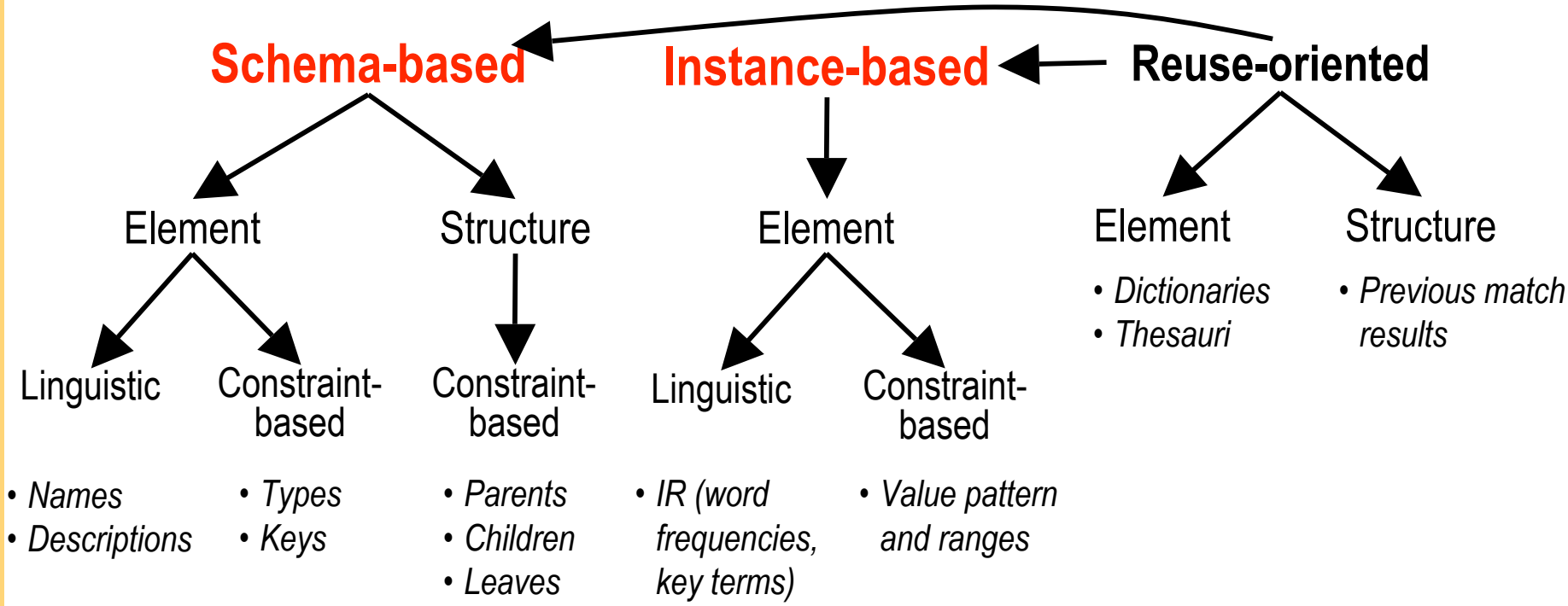
Motivation

- Increasing number of connected sources and ontologies



- Ontology matching (alignment)
  - Goal: Find semantically related concepts
  - Output: Set of correspondences (ontology mapping)
    - Ideally: + semantic mapping type (equivalence, is-a, part-of, …)
  - Use:
    - Improved analysis
    - Validation (curation) and recommendation of instance associations
    - Ontology merge or curation, e.g. to reduce overlap between ontologies

Automatic Match Techniques*



- Combined Approaches: Hybrid vs. Composite
- Many frameworks / prototypes: COMA++, Prompt, FOAM, Clio, … but mostly not used in bioinformatics

*Rahm, E., P.A. Bernstein: *A Survey of Approaches to Automatic Schema Matching.* VLDB Journal 10(4), 2001

Frameworks: PROMPT*

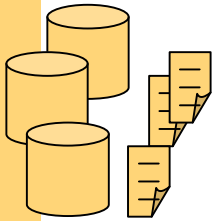
- Framework for ontology alignment and merging
  - Plug-in tool for Protege 2000
- Linguistic matching
- Iterative user feedback and match result manipulation
  - Automatic detection of ontology conflicts
  - Interactive conflict resolution and automaticconflict resolution based on user-preferred ontology
- Merge operation: Create a new ontology or extend one selected ontology
  - Automatic creations of parent- and sub-concept relationships
  - Suggestions of similar concepts based on ontology matches

*Noy, N.; Musen, M.: *PROMPT – Algorithm and tool for automated ontology merging and alignment*. Proc. Conf. on Artificial Intelligence and Innovative Applications of Artificial Intelligence, 2000.

System Architecture*

Graphical User Interface

External Schemas, Ontologies



Execution Engine

Matcher Strategy

Component Identification → Matcher Execution → Similarity Combination

Model Pool

Model Manipulation

Match Customizer

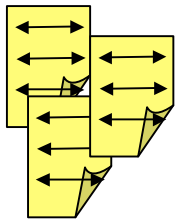
Matcher Configs

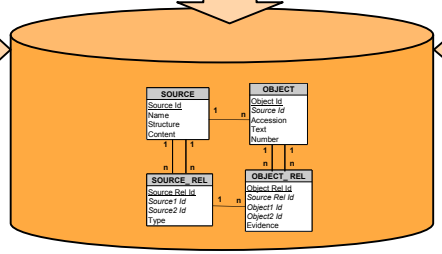
Match Strategies

Mapping Pool

Mapping Manipulation

Exported Mappings





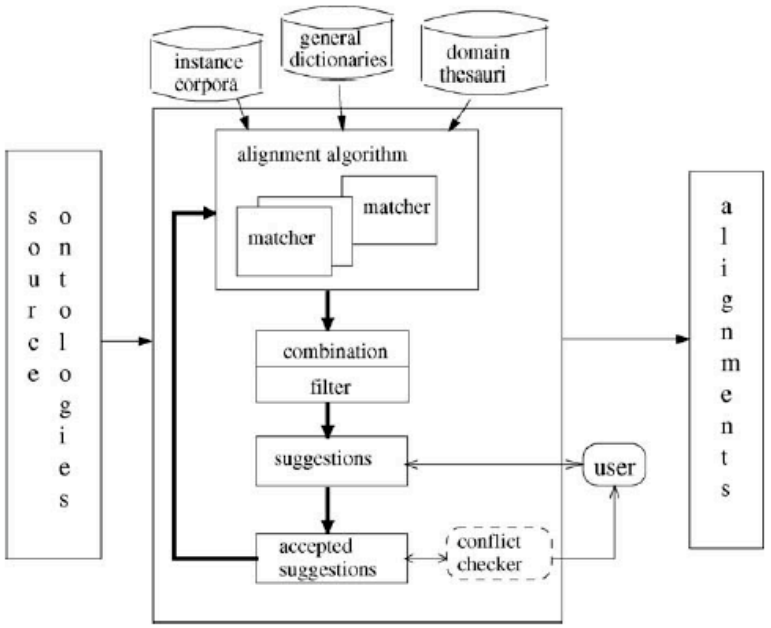
Repository

*Do, H.H., E. Rahm: *COMA - A System for Flexible Combination of Schema Matching Approaches*. VLDB 2002

Aumüller D., H.-H. Do, S. Massmann, E. Rahm: *Schema and Ontology Matching with COMA++*. Sigmod 2005

Frameworks: SAMBO*

- System for aligning and merging biomedical ontologies
- Framework to find similar concepts in overlapping ontologies for alignment and merge tasks
  - Import of OWL ontologies
  - Support of various match strategies by applying / combining different matchers and use of auxiliary information
    - Linguistic, structure-based, constraint-based, instance-based matcher
  - Iterative user feedback for match results
  - Result manipulation by description logic reasoner checking for ontology consistency, cycles, unsatisfiable concepts



*Lambrix, P; Tan, H.: *SAMBO – A system for aligning and merging biomedical ontologies*. Journal of Web Semantics, 4(3):196-206 , 2006.

Metadata-based match approaches

- Metadata: Concept names, descriptions, ontology structure, ...
- Match mainly based on syntax and structure

- Limited use of domain knowledge
- Highly similar names with opposite semantics,
  e.g., ion vs. anion, organic vs. inorganic

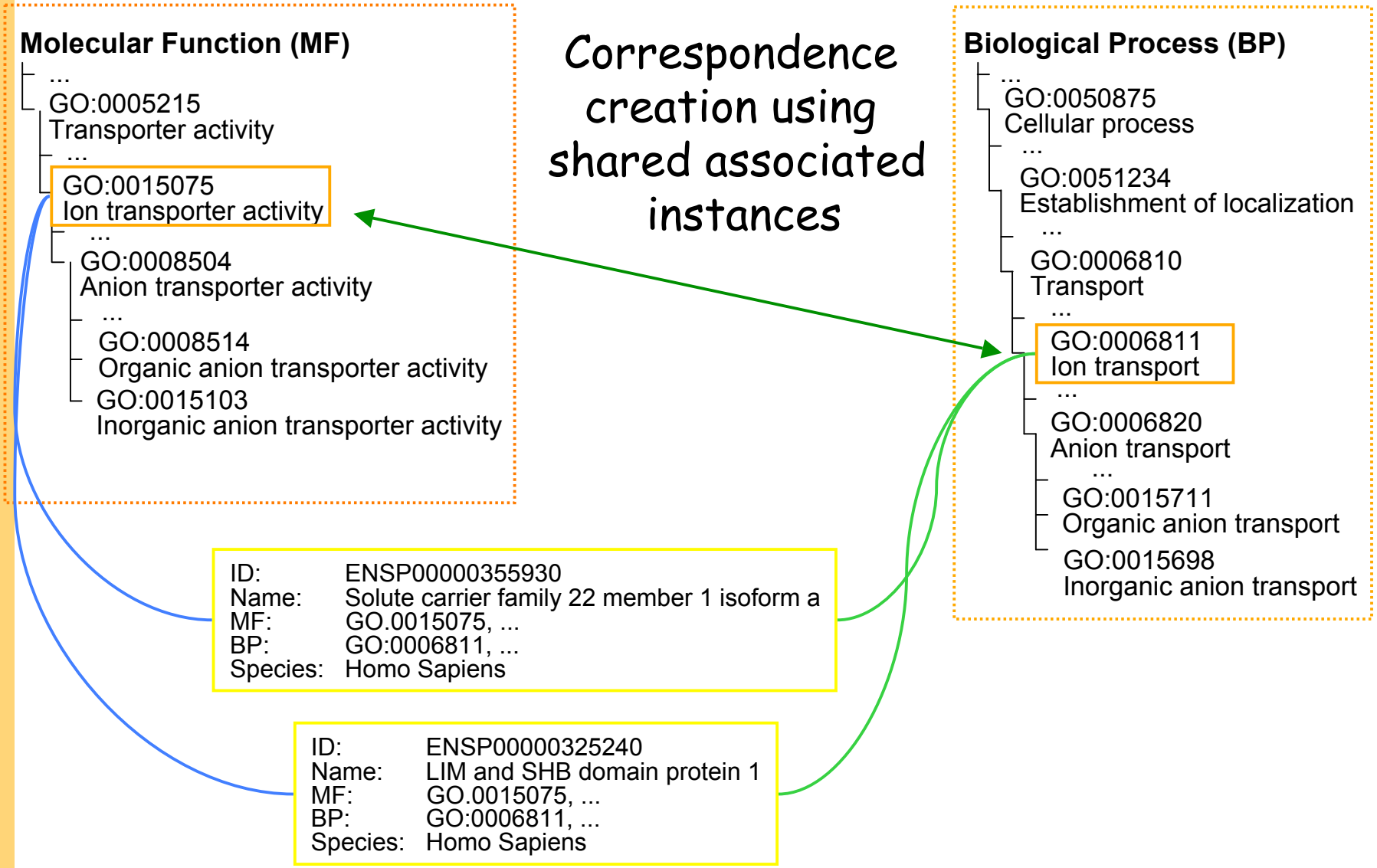
		$Sim_{2\text{-}Gram}$
ion transporter	– anion transport	0.77
ion transporter activity	– ion transport	0.66

Instance-based match approach*

- Approach
  - Use domain-specific knowledge expressed in existing instance associations to create ontology mappings
- **Key idea**: "Two concepts are related if they share a significant number of associated objects"
- Flexible and extensible approach
  - Instance associations of pre-selected sources
  - Different metrics to determine the instance-based similarity
  - Combination of different ontology mappings

* Kirsten, T, Thor, A; Rahm, E.: *Instance-based matching of large life science ontologies.* Proc. 4th Intl. Workshop DILS, July 2007

Instance-based matching



Selected similarity metrics

- Baseline similarity Sim_{Base}

$$Sim_{Base}(c_1, c_2) = \begin{cases} 1 \ , \text{ if } N_{c_1c_2} > 0 \\ 0 \ , \text{ if } N_{c_1c_2} = 0 \end{cases}$$

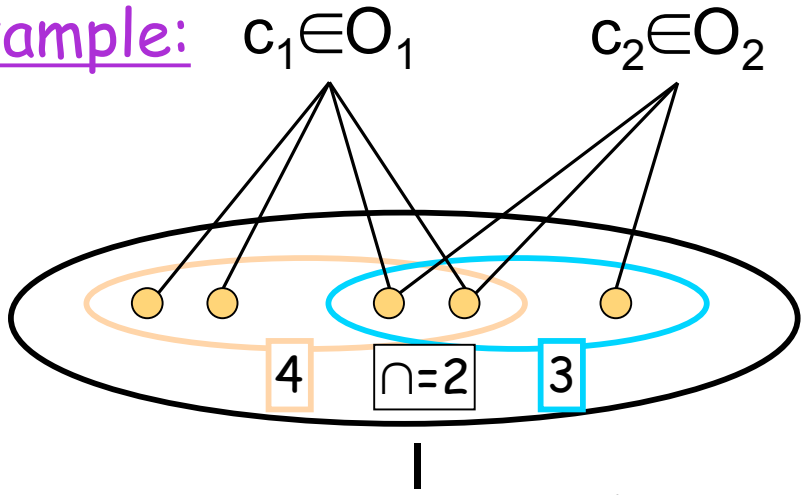
- Dice similarity Sim_{Dice}

$$Sim_{Dice}(c_1, c_2) = \frac{2 \cdot N_{c_1c_2}}{N_{c_1} + N_{c_2}}$$

- Minimum similarity Sim_{Min}

$$Sim_{Min}(c_1, c_2) = \frac{N_{c_1c_2}}{\min(N_{c_1}, N_{c_2})}$$

Example:



Sim_{Base} = 1

Sim_{Dice} = 2*2/(4+3) = 0.57

Sim_{Min} = 2/3 = 0.67

$0 \leq Sim_{Dice} \leq Sim_{Min} \leq Sim_{Base} \leq 1$

Evaluation metrics

- Computation of precision & recall needs a perfect mapping
  - Laborious for large ontologies
  - Might not be well-defined
- Metric *Match Coverage* to approximate "recall"
  - Idea: Measure fraction of matched concepts

$$MatchCoverage_{O_1} = \frac{|C_{O_1-Match}|}{|C_{O_1}|} \in [0...1] \qquad Combined\ InstMatchCoverage = \frac{|C_{O_1-Match}| + |C_{O2-Match}|}{|C_{O_1-Inst}| + |C_{O_2-Inst}|} \in [0...1]$$

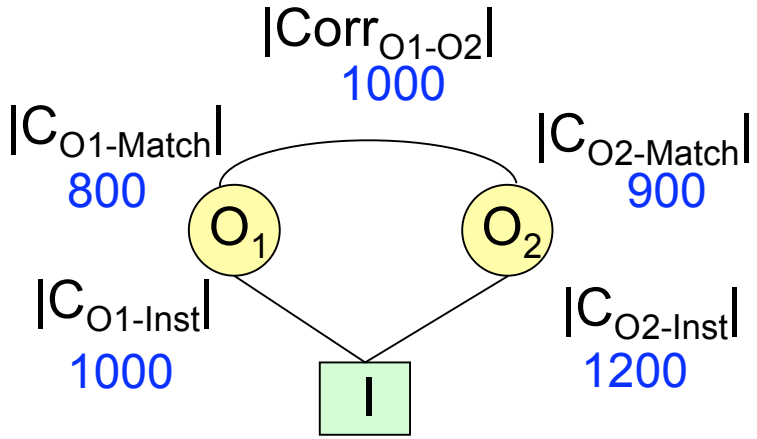
- Metric *Match Ratio* to approximate "precision"
  - Idea: Measure average number of match counter-parts per matched concept

$$MatchRatio_{O1} = \frac{|Corr_{O1-O2}|}{|C_{O1-Match}|} \geq 1 \qquad CombinedMatchRatio = \frac{2 \cdot |Corr_{O1-O2}|}{|C_{O1-Match}| + |C_{O2-Match}|} \geq 1$$

- Goal: high Match Coverage with low Match Ratio

Evaluation metrics cont.

- Example:

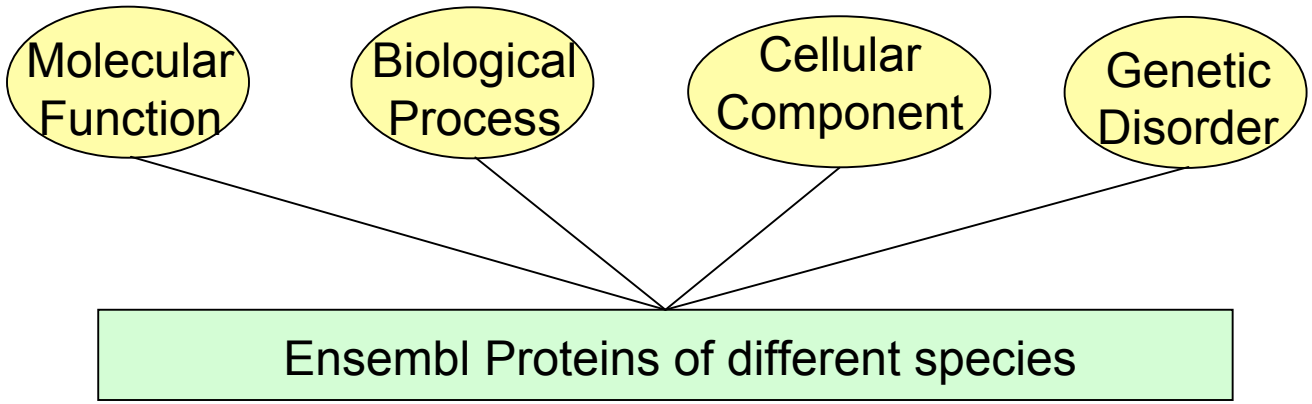


$InstMatchCoverage_{O1} = 800/1000 = 0.80$
$InstMatchCoverage_{O2} = 900/1200 = 0.75$

$MatchRatio_{O1} = 1000/800 = 1.25$
$MatchRatio_{O2} = 1000/900 = 1.11$

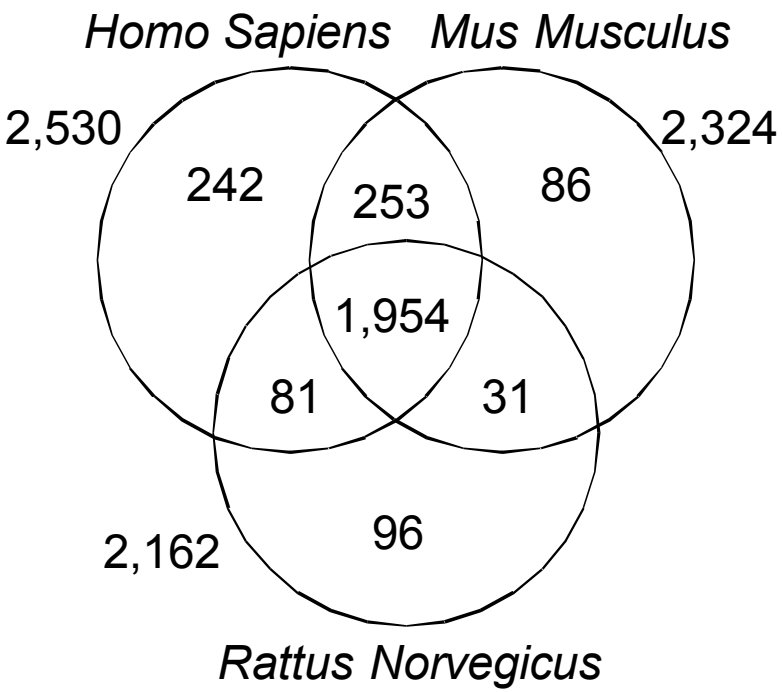
Match scenario

- Ontologies
  - Subontologies of GeneOntology: Mol. function, biol. processes and cell. components
  - Genetic disorders of OMIM
- Instances: Ensembl proteins of different species, i.e., homo sapiens, mus musculus, rattus norvegicus



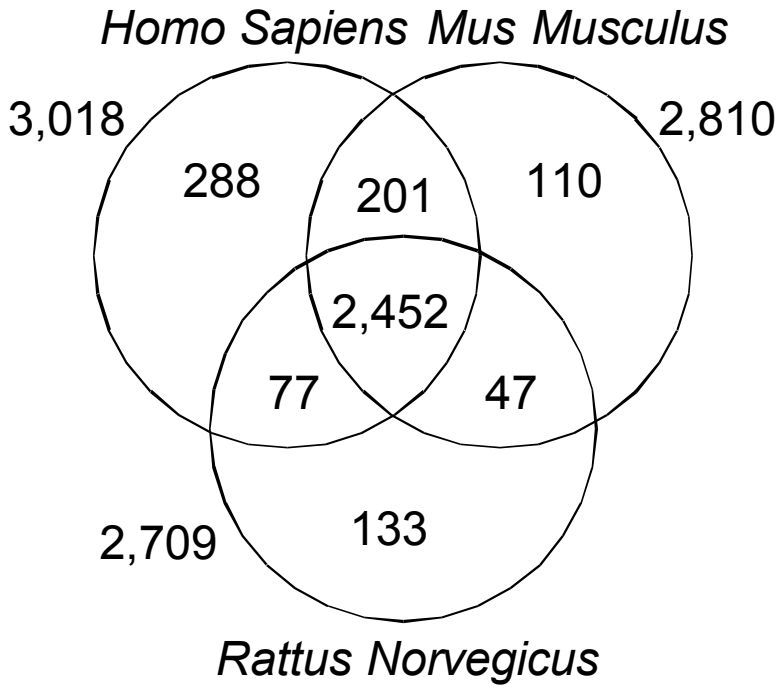
Ontology overlap between species

Total # functions: 7,514



Number of associated Molecular Functions

Total # processes: 12,555



Number of associated Biological Processes

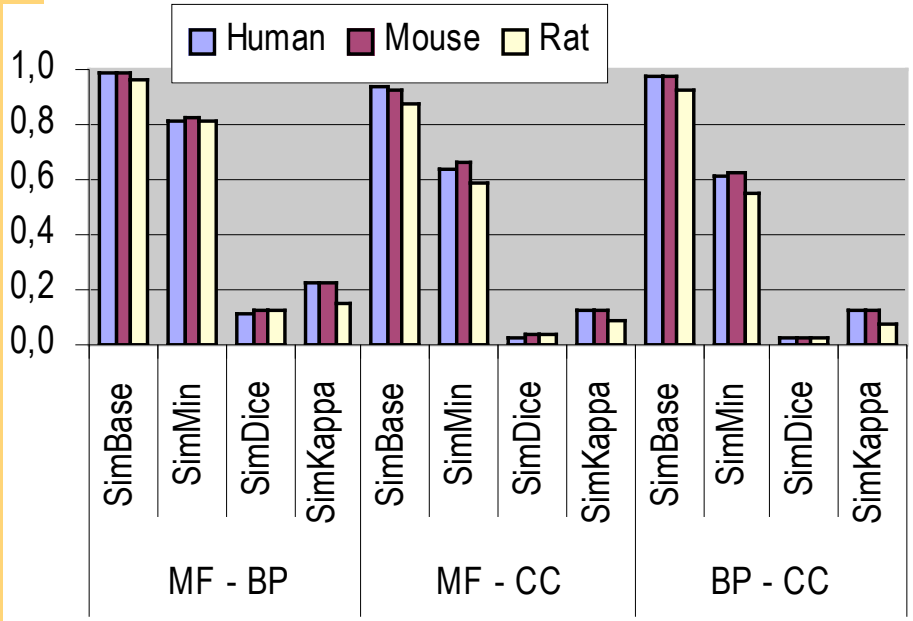
Exhaustive match study

- Instance-based matching
  - Direct protein associations of human, mouse, rat
  - Study of match combinations: Union, intersection
  - Utilization of indirect associations
- (Simple) Metadata-based matching
  - Utilization of concept names
  - Trigram string similarity; different thresholds
- Comparison of instance- and metadata-based match results

Match results: Direct instance associations

- Sim_{Base}: High Coverage (99%), moderate to high Match Ratios
- Sim_{Dice}: Very restrictive (Coverage < 20%) but low Match Ratios
- Sim_{Min}: High Coverage (60%-80%) with high number of covered concepts but significantly lower Match Ratios than Sim_{Base}

Combined Instance Coverage



Match Ratios per ontology

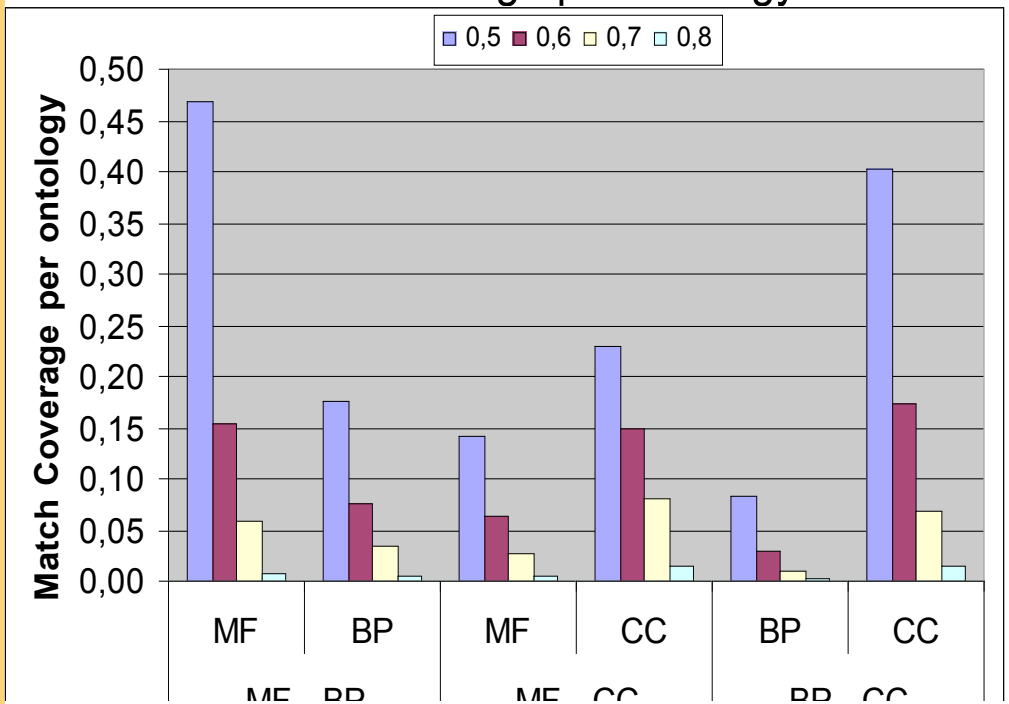
	MF - BP		MF - CC		BP - CC	
	MF	BP	MF	CC	BP	CC
Base	20.4	17.0	7.6	28.6	9.8	46.3
Min	4.4	4.0	2.2	7.8	2.4	8.6
Dice	1.3	1.2	1.0	1.3	1.0	1.3
Kappa	2.0	2.0	1.9	2.7	1.7	2.6

(Match Ratios for Homo Sapiens)

Match results: Metadata-based matching

- Growing Coverage and Match Ratios for lower thresholds
- No correspondences with a similarity ≥ 0.9
- Moderate to low Match Ratios
- Inclusion of false positives for low thresholds, e.g. 0.5

Match Coverage per ontology

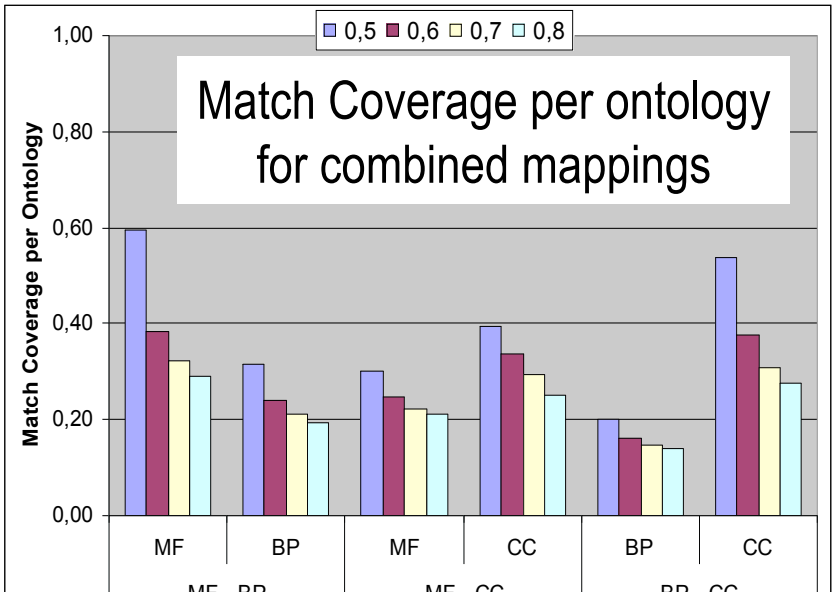


Match Ratios per ontology

	MF - BP		MF - CC		BP - CC	
	MF	BP	MF	CC	BP	CC
0.5	4.4	6.9	2.5	6.3	2.5	3.4
0.6	2.4	2.9	2.7	4.6	1.7	2.0
0.7	1.4	1.4	1.1	1.5	1.4	1.4
0.8	1.1	1.1	1.1	1.2	1.1	1.2

Match results: Match combinations

- Combinations between instance- (Sim_{Min}) and metadata-based match approach
  - Union: Increased coverage, higher influence of Sim_{Min} for increased thresholds of the metadata-based matcher
  - Intersection: Low Match Coverage (<1%) and Match Ratios
- Low overlap between instance- and metadata-based mappings



Match Ratios per ontology (threshold 0.7)

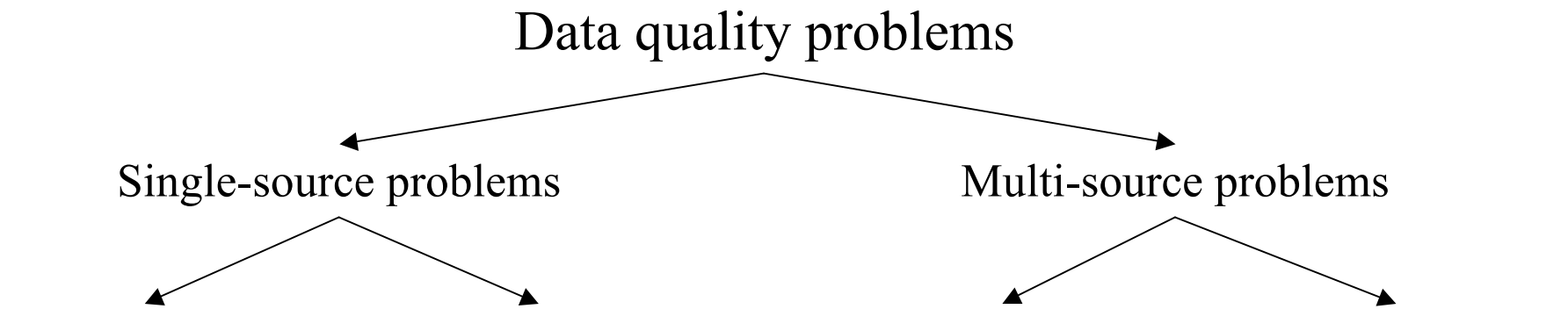
	MF - BP		MF - CC		BP - CC	
	MF	BP	MF	CC	BP	CC
∪	4.1	3.7	2.2	6.7	2.4	7.6
∩	1.0	1.0	1.0	1.0	1.0	1.3

(Sim_{Min} = 1.0, Homo Sapiens)

Agenda

- Kinds of data to be integrated
- General data integration alternatives
- Warehouse approaches
- Virtual and mapping-based data integration
- Data quality aspects
  - Overview and examples of quality problems
  - Object Matching
  - Data cleaning frameworks
- Conclusions and further challenges

Overview*



- Uniqueness
- Referential integrity
- ...

- Mispellings
- Redundancy, duplicates
- Contradictory values
- ...

- Naming conflicts
- Structural conflicts
- ...

- Inconsistent aggregating
- Inconsistent timing
- ...

*Rahm, E; Do, H.-H.: *Data cleaning: Problems and current approaches*. IEEE Techn. Bulletin on Data Engineering, 23(4), 2000

Single-source problems

- Example: Protein data
- Causes
  - Schemaless storage, e.g., file-based data storage
  - Lack of input / acceptance integrity constraints
  - ...

Uniqueness

Synonyms

Case insensitivity

Accession	Entry-Name	Protein-Name	Species	Comment	Sequence
P68511	1433F_RAT	14-3- protein eta	Rat		MGDREQLL...
P11576	1433F_RAT	14-3- protein eta	Rattus norvegicus		mgdreqll...
P0A5B7	14KD_MYCTU	14 kDa antigen, also: 16kDa antigen, HSP16.3	Mycobaterium tuberculosis	[ENSEMBL: ENSP00007463]	

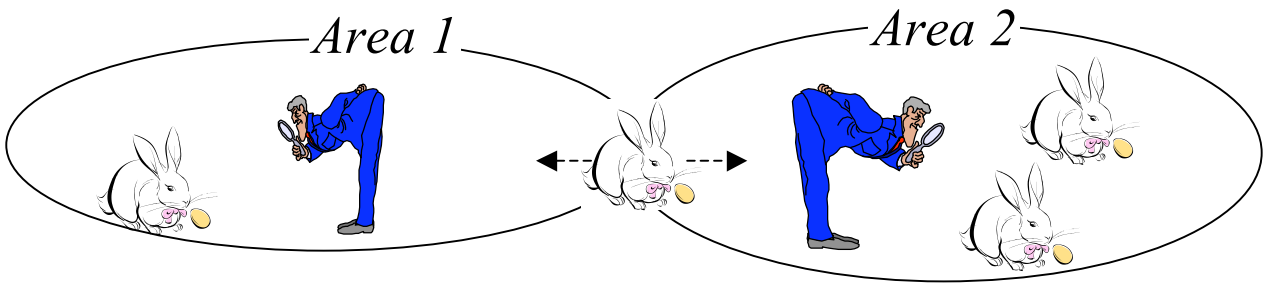
Multiple values

Encoding of further annotations and links

Missing values

Multi-source problems (selection)

- Multiple experiments on same problem with different results
  - Different normalization and analysis methods
  - Human interpretation !
- Observations of mobile things, e.g., animals in bordering areas
  - Human observations
  - Varying annotations (difficult to be objective):
    - white-brown vs. brown-white, full vs. complete
  - Example: Describe and count animal populations



Nr	Colour	Pattern	...
1	snow-white	full	...
2	white-brown	spotted	...

Integration with object fusion

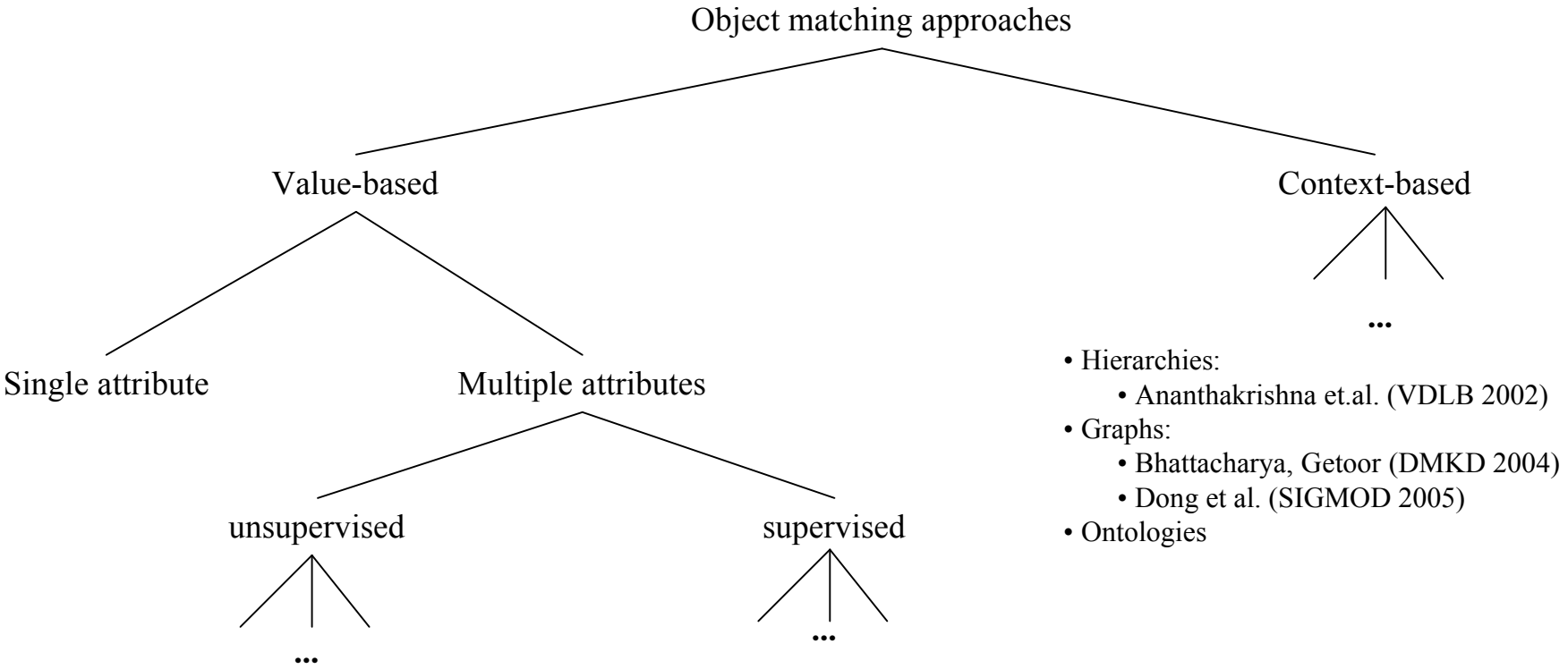
Nr	Colour	Pattern	...
1	white-brown	spotted	...
2	beige	complete	...
3	white	complete	...

Simple solution strategies

- Uniqueness
  - Utilization of global identifiers
  - Use identifier mappings to a second source (of the same type and detail level)
- Multiple values / encodings
  - Extract atomic values by specific parsers, regular expressions
  - Normalization of dependent attributes
- Synonyms: Use of available controlled vocabularies / ontologies as much as possible, e.g., NCBI Taxonomy for species
- Case insensitives: Compare case insensitively or transform all values to upper/lower case before the comparison starts; evtl. delete blanks

Accession	Entry-Name	Protein-Name	Species	Comment	Sequence
P68511	1433F_RAT	14-3- protein eta	Rat		MGDREQLL...
P11576	1433F_RAT	14-3- protein eta	Rattus norvegicus		mgdreqll...
P0A5B7	14KD_MYCTU	14 kDa antigen, also: 16kDa antigen, HSP16.3	Mycobaterium tuberculosis	[ENSEMBL: ENSP00007463]	

Object matching approaches



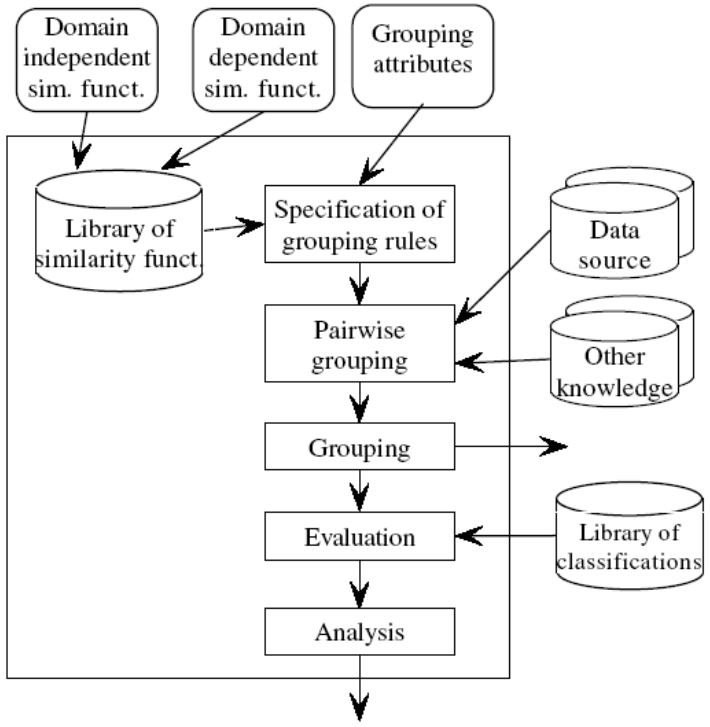
- Hierarchies:
  - Ananthakrishna et.al. (VDLB 2002)
- Graphs:
  - Bhattacharya, Getoor (DMKD 2004)
  - Dong et al. (SIGMOD 2005)
- Ontologies

- Aggregation function with threshold
- User-specified Rules:
  - Hernandez et al. (SIGMOD 1995)
- Clustering
  - Monge, Elkan (DMKD 1997)
  - Mc Callum et al. (SIGKDD 2000)
  - Cohen, Richman (SIGKDD 2002)

- decision trees
  - Verykios et al. (Information Sciences 2000)
  - Tejada et al. (Information Systems 2001)
- support vector machine
  - Bilenko, Mooney (SIGKDD 2003)
  - Minton et al. (2005)

Similarity-based grouping*

- Goal: Detect and group duplicate (very similar) data entries
- Sequential procedure
  - Specification of grouping rules: Which similarity functions (also combinations) for which attributes
  - Pairwise grouping: Computing the similarity and comparing data entries based on selected / specified grouping rules
  - Grouping of pairs of data entries into cliques based on
    - Total number of groups
    - Number of data entries in a group
    - Disjoint / overlapping groups
  - Analysis and evaluation of generated groupings



*Jakoniene, V; Rundqvist, D.; Lambrix, P.: *A method for similarity-based groupig of biological data*. Proc. DILS, 2006

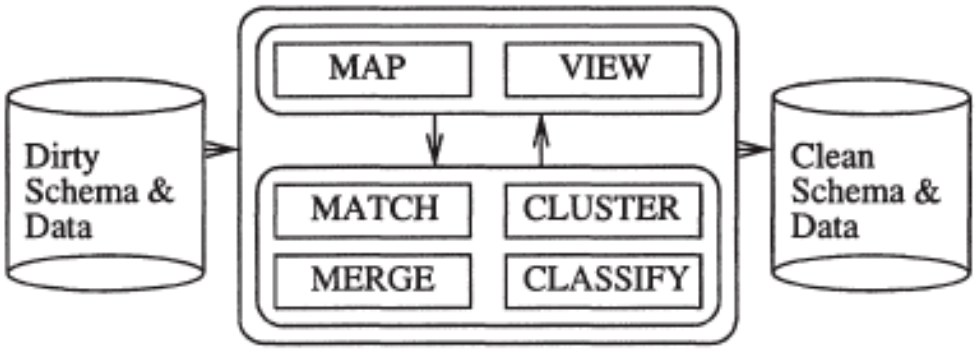
Similarity-based grouping: Test cases

- Test: Group selected proteins into classes using
  - Annotations, e.g., attributes like product, definition
  - Protein sequences
  - Associations to GO ontology
- Results
  - Best grouping by using GO associations
  - Annotation-based: Too many groups
  - Sequence alignments: Too specific for grouping

Test case	Grouping rule	n^e	n^g	n^c	p	1-E	F	MI
1	$SemSim(GO_{ann})> 0.95$ GO_{ann} for component, process, function domains	71	23	24	0.90	0.93	0.88	0.86
2	$SemSim(GO_{ann})> 0.95$	67	26	23	1.00	1.00	0.97	0.91
3	$SemSim(GO_{ann} + GO_{sw})> 0.95$	75	23	24	0.80	0.87	0.79	0.79
4	$SemSim(GO_{ann} + GO_{ec})> 0.95$	92	26	25	1.00	1.00	0.99	0.88
5	$SemSim(GO_{ann} + GO_{sw} + GO_{ec})> 0.95$	93	26	25	0.86	0.93	0.88	0.81
6	$SemSim(GO_{ann} + GO_{sw} + GO_{ec})> 0.95$; parent GO terms removed	93	26	25	0.86	0.93	0.88	0.81
7	$SemSim(GO_{ann})> 0.95$ or $SemSim(GO_{sw})> 0.95$ or $SemSim(GO_{ec})> 0.95$	93	14	25	0.48	0.65	0.51	0.59
8	$SemSim(GO_{ann})> 0.95$ or $SemSim(GO_{ec})> 0.95$	92	26	25	1.00	1.00	0.99	0.88
9	$SemSim(GO_{ann} + GO_{ec})= 1$	92	26	25	1.00	1.00	0.99	0.88
10	$SemSim(GO_{ann} + GO_{ec})> 0.85$	92	21	25	0.70	0.78	0.71	0.68
11	$SemSim(GO_{ann} + GO_{ec})> 0.95$ grouping algorithm: cliques	92	29	25	1.00	1.00	0.84	0.88
12	$EditDist(definition)> 0.9$, for $GO_{ann} + GO_{ec}$	92	67	25	1.00	1.00	0.59	0.77
13	$EditDist(definition)> 0.7$, for $GO_{ann} + GO_{ec}$	92	55	25	0.96	0.97	0.66	0.76
14	$SeqSim(sequence)> 0.85$, for $GO_{ann} + GO_{ec}$	92	44	25	1.00	1.00	0.74	0.81
15	$EditDist(definition)> 0.85$	190	94	28	0.97	0.98	0.54	0.57
16	$EditDist(product)> 0.85$	190	105	28	0.99	0.99	0.49	0.57
17	$EditDist(definition)> 0.7$	190	68	28	0.81	0.87	0.56	0.50
18	$EditDist(product)> 0.7$	190	78	28	0.95	0.98	0.64	0.58
19	$EditDist(definition)> 0.9$ or $EditDist(product)> 0.9$ or ($EditDist(definition)> 0.6$ and $EditDist(product)> 0.6$)	190	64	28	0.94	0.96	0.70	0.58
20	$SeqSim(sequence)> 0.85$	190	59	28	0.99	0.99	0.66	0.62

BIO-AJAX*

- Framework for biological data cleaning
- Operators
  - MAP: translates the data from one schema to another schema.
  - VIEW: extracts portions of data for cleaning purposes.
  - MATCH: detects duplicate or similar records
  - MERGE: combines duplicate records or similar records into one record



*Herbert, K.G.; Gehani, N.H.; Piel, W.H.; Wang, J.T.-L.; Wu, C.H.: *BIO-AJAX: An Extensible Framework for Biological Data Cleaning.* SIGMOD Record 33(2), 2004

Further data cleaning frameworks

- Research prototypes
  - AJAX (Galhardas et al., VLDB 2001)
  - IntelliClean (Lee et al., SIGKDD 2000)
  - Potter's Wheel (Raman et al., VLDB 2001)
  - Febrl (Christen, Churches, PAKDD 2004)
  - TAILOR (Elfeky et al., Data Eng. 2002)
  - MOMA (Thor, Rahm, CIDR 2007)
- Commercial solutions
  - DataCleanser (EDD), Merge/Purge Library (Sagent/QM Software), MasterMerge (Pitnew Bowes) ...
  - MS SQL Server 2005: Data Cleaning Operators (Fuzzy Join / Lookup)

Agenda

- Motivation
- General data integration alternatives
- Warehousing of large biological data collections
- Virtual integration of molecular-biological data
- Data quality aspects
- Matching large life science ontologies
- Conclusions and further challenges

Overall conclusions

- Diverse data characteristics
  - Large amounts of experimental data produced by different chip technologies
  - Integration / management of clinical data
  - Huge amount of inter-connected web sources
  - High amount of text data
- Comprehensive standardization efforts needed: object ids / formats, preprocessing routines of chip data, shared vocabularies / ontologies
- Need to support explorative workflows across different sources
- Different data integration architectures needed
  - Data Warehousing
  - Virtual and mapping-based integration approaches
  - Combinations

Overall conclusions cont.

- Warehousing for integration of large collections of biological data
  - Ideal for analysis / data mining on huge data sets, e.g. experimental chip data
  - Comprehensive data preprocessing
  - Support for consistent annotations needed
  - Integration of external data for enhanced analysis
- Mapping-based data integration (e.g., BioFuice)
  - Utilization of instance-level mappings to traverse between sources and fuse objects
  - Set-oriented navigation + structured queries + keyword search
  - Programmability / workflow orientation
- Ontology matching
  - Metadata vs. instance-based matching, combined approach
  - Key problem: validation of mappings by domain experts
  - More research needed

Future challenges

- Clinical data management: many organizational issues, data privacy
- Bridging different workstyles and research goals: computers scientists vs. biologists vs. clinicians
- Make data integration easier and faster, e.g. by a mashup-like paradigm
  - Enable biologist/users to extract, clean, integrate and analyze data themselves
  - Make it easier to develop and use data-driven workflows
- Annotation and ontology management
  - Creation, evolution, matching, merging of ontologies
  - Utilization of generic and domain-specific approaches
- Data quality: object matching and fusion, provenance, …
- Data integration in new application fields, e.g. systems biology
  - e.g., management of metabolic ~, regulatory pathways, protein-protein-interaction networks
  - Combination of data of wet-lab experiments with cell-based simulation (in silico experiments)

Literature: Surveys, Overviews

- T. Hernandez, S. Kambhampati: *Integration of biological sources*: *current systems and challenges ahead*. SIGMOD record, 33(3):51-60, 2004.
- Z. Lacroix: *Biological data integration: wrapping data and tools.* IEEE Trans. Information Technology in Biomedicine. 6(2), 2002
- Z. Lacroix, T. Critchlow (eds.): *Bioinformatics – Managing scientific data*. Morgan Kaufmann Publishers, 2003.
- B. Louie, P. Mork, F. Martin-Sanchez, A. Halevy, P. Tarczy-Hornoch: *Data integration and genomic medicine*. Journal of Biomedical Informatics, 40:5-16, 2007.
- L. Stein: *Integrating biological databases*. Nature Review Genetics, 4(5):337-345, 2003.

- H.-H. Do, T. Kirsten, E. Rahm: *Comparative evaluation of microarray-based gene expression databases*. Proc. 10th BTW Conf., 2003.
- M.Y. Galperin: *The molecular biology database collection: 2006 update*. Nucleic Acids Research, 34 (Database Issue):D3-D5, 2006.

Literature: Warehousing of biological data

- A. Brazma et al.: *Minimum information about a mircoarray experiment (MIAME) – toward standards for microarray data*. Nature Genetics, 29(4): 365-371, 2001
- A. Kasprzyk, D. Keefe, D. Smedley et al.: *EnsMart: A generic system for fast and flexible access to biological data*. Genome Research, 14(1):160-169, 2004
- T. Kirsten, J. Lange, and E. Rahm: *An integrated platform for analyzing molecular-biological data within clinical studies*. Proc. Intl. EDBT Workshop on Information Integration in Healthcare Applications, 2006.
- V.M. Markowitz et al.: *The Integrated Microbial Genomes (IMG) System: A Case Study in Biological Data Management .* Proc. VLDB 2005
- R. Nagarajan, M. Ahmed, A. Phatak: *Database challenges in the integration of biomedical data sets*. Proc. 30th VLDB Conf., 2004.
- E. Rahm, T. Kirsten, J. Lange: *The GeWare data warehouse platform for the analysis of molecular-biological and clinical data*. Journal of Integrative Bioinformatics, 4(1):47, 2007.
- K. Rother, H. Müller, S. Trissl et al.: *Columba: Multidimensional data integration of protein annotations*. Proc. 1st DILS Workshop, 2004.

Literature: Virtual & mapping-based integration

- H.-H. Do, E. Rahm: *Flexible integration of molecular-biological annotation data: The GenMapper approach*. Proc. EDBT Conf., 2004.
- T. Etzold, A. Ulyanov, P. Argos: *SRS: Integration retrival system for molecularbiological data banks*. Methods in Enzymology, 266:114-128, 1996.
- L. Haas et al.: *Discoverylink: A system for integrating life sciences data*. IBM Systems Journal 2001
- D. Hull et al.; *Taverna: a tool for building and running workflows of services*. Nucleic Acid Research 2006
- T. Kirsten, E. Rahm: *BioFuice: Mapping-based data integration in bioinformatics*. Proc. 3rd Intl. Workshop on Data Integration in the Life Sciences, 2006.
- B. Ludaescher at al.: *Scientific Workflow Management and the Kepler System*. Concurrency and Computation: Practice & Experience, 2005
- A. Prlic, E. Birney, T. Cox et al.: *The distributed annotation system for integration of biological data*. Proc. 3rd Workshop on Data Integration in the Life Sciences, 2006.
- S. Prompramote, Y.P. Chen: *Annonda: Tool for integrating molecular-biological annotation data*. Proc. 21st ICDE Conf., 2005.
- E. Rahm, A.Thor, D. Aumüller et al.: *iFuice – Information fusion utilizing instance-based peer mappings*. Proc. 8th WebDB Workshop, 2005.
- R. Stevens et al.: *Tambis - Transparent Access to Multiple Bioinformatics Information Sources*. Bionformatics 2000
- J. Saltz, S. Oster, et al.: *caGRID: Design and implementation of the core architecture of the cancer biomedical informatics grid. Bioinformatics*, 22(15):1910-1916, 2006.

Literature: Ontologies and ontology matching

- S. Schulze-Kremer: *Ontologies for molecular biology*. Proc. 3rd Pacific Symposium on Biocomputing, 1998.
- O. Bodenreider, M. Aubry, A. Bugrun: *Non-lexical approaches to identifying associative relations in the Gene Ontology*. Proc. Pacific Symposium on Biocomputing, 2005.
- O. Bodenreider, A.Bugrun: *Linking the Gene Ontology to other biological ontologies*. Proc. ISMB Meeting on Bio-Ontologies, 2005.
- J. Euzenat, P. Shvaiko: *Ontology matching*. Springer Verlag, 2007.
- T. Kirsten, A. Thor, E. Rahm: *Matching large life science ontologies*. Proc. 4th Intl. Workshop on Data Integration in the Life Sciences. 2007.
- P. Mork, P. Bernstein: *Adapting a generic match algorithm to align ontologies of human anatomy*. Proc 20th ICDE Conf., 2004.
- S. Myhre, H. Tveit, T. Mollestad, A. Laengreid: *Additional Gene Ontology structure for improved biological reasoning*. Bioinformatics, 22(16):2020-2037, 2006.
- P. Lambrix, H.Tan: *Sambo – A system for aligning and merging biomedical ontologies*. Journal of Web Semantics, 4(3):196-206 , 2006.

Literature: Data quality aspects

- A.K. Elmagarmid, P.G. Ipeirotis, and V.S. Verykios: *Duplicate Record Detection: A Survey*. IEEE Transactions on Knowledge and Data Engineering 19(1), 2007.
- K.G. Herbert et al: *BIO-AJAX: An Extensible Framework for Biological Data Cleaning*. SIGMOD Record 33(2), 2004
- K.G. Herbert, J. Wang: *Biological data cleaning: A case study*. International Journal of Information Quality, 1(1):60-82, 2007.
- V. Jakoniene, D. Rundqvist, and P. Lambrix: *A method for similarity-based grouping of biological data*. Proc 3rd Intl. Workshop on Data Integration in the Life Sciences, 2006.
- J. Koh, M. Lee, A. Khan et al.: *Duplicate detection in biological data using association rule mining*. Proc Workshop on Data and Text Mining in Bioinformatics, 2004.
- A. Monge C. Elkan: *An efficient domain-indepent algorithm for detecting approximatively duplicate database records*. Proc. SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, 1997.
- H. Müller and J.-C. Freytag: *Problems, Methods and Challenges in Comprehensive Data Cleansing*. Technical Report HUB-IB-164, Humboldt University Berlin, 2003.
- F. Naumann, J.-C. Freytag, and U. Leser: *Completeness of integrated information sources*. Journal of Information Systems, 29(7):583-615, 2004.
- E. Rahm, H.-H. Do: *Data cleaning: Problems and current approaches*. IEEE Bulletin of the Technical Committee on Data Engineering, 23(4):3-13, 2000.

Online Bibliographies

http://dc-pubs.dbs.uni-leipzig.de/
http://se-pubs.dbs.uni-leipzig.de/