Next Generation Data Integration for the Life Sciences

(short version of the 3h tutorial ICDE 2011)

Sarah Cohen-Boulakia
Université Paris Sud – Orsay

AMIB group INRIA Saclay

Ulf Leser
Humboldt-Universität zu Berlin

Data Integration for the Life Sciences in 1993

- Robbins, R. J. (1994). "Report of the invitational DOE Workshop on Genome Informatics I: Community Databases." [Rob94a]
  - DOE funded large parts of the HGP starting end of the 80ties
- "Continued HGP progress will depend in part upon the ability of genome databases to answer increasingly complex queries that span multiple community databases. Some examples of such queries are given in this appendix."
- "Note, however, until a fully atomized sequence database is available (i.e., no data stored in ASCII text fields), **none of the queries in this appendix can be answered**. The current emphasis of GenBank seems to be providing human-readable annotation for sequence information. Restricting such information to human-readable form is totally inadequate for users who require a different point of view, namely one in which the sequence is an annotation for a computer-searchable set of feature information."

Twelve Queries Unanswerable in 1993

- 1. Return all sequences which map 'close' to marker M on chrom. 19, are put. members of the olfactory receptor family, and have been mapped on a contig
  - Multidatabase: Chromosome maps from GDB, sequence-contig in GenBank, annotation from elsewhere
- 2. Return the map location, where known, of all *alu* elements having homology greater than "h" with the *alu* sequence "S".
  - Only needs GenBank and a similarity search
- 3. Return all h. gene sequences for which a putative functional homologue has been identified in a non-vertebrate organism
  - Human: GenBank, non-vertebrates: species databases; how to describe function?
- 4. Return the number and a list of the distinct human genes that have been sequenced
  - What is a gene? Semantic heterogeneity and scientific uncertainty
- 5. Return all publications from the last two years about my favorite gene "…"
  - Synonyms & homonyms; naming conventions, disambiguation

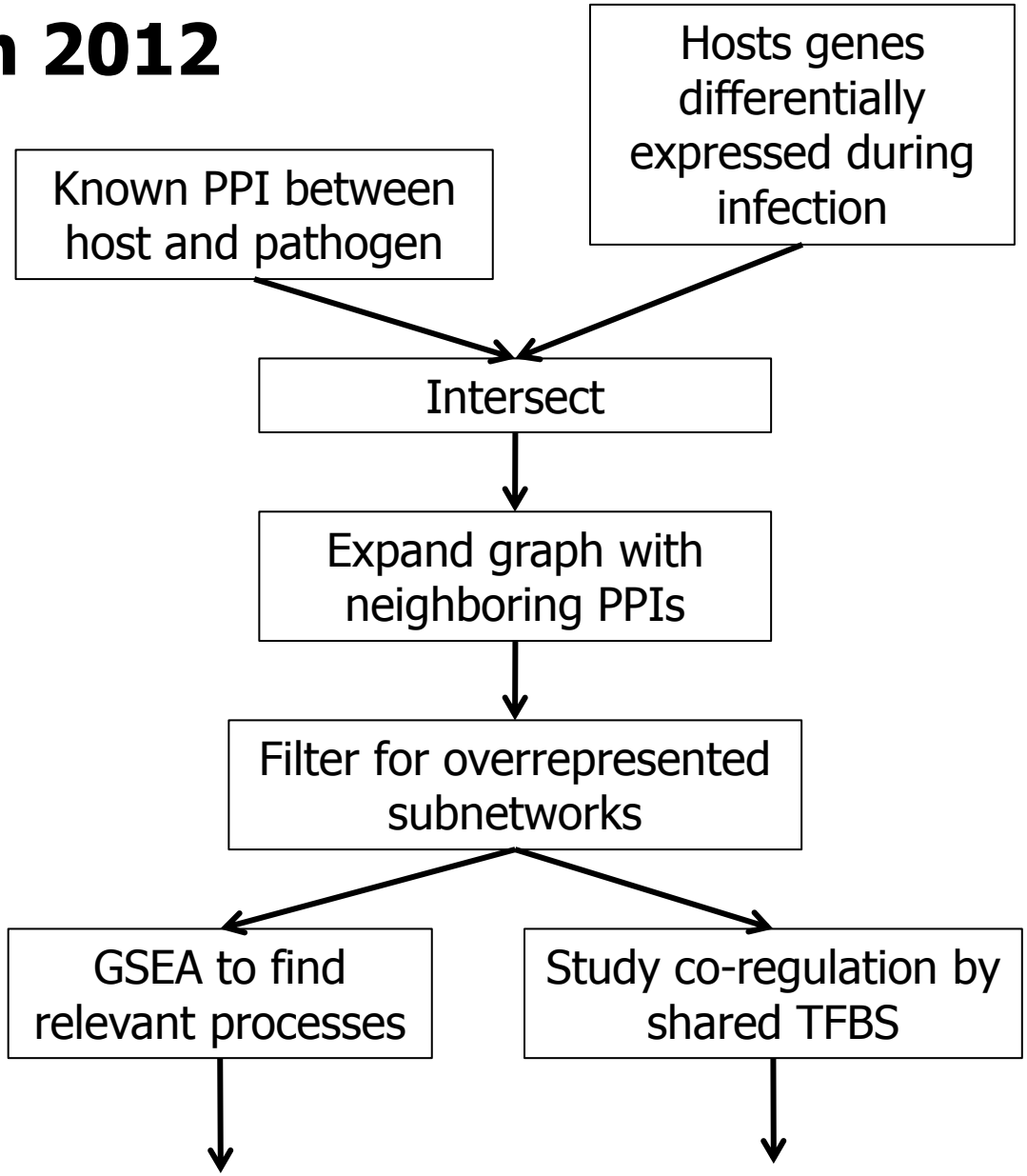
Take Home Message

- The classical problems are all there already

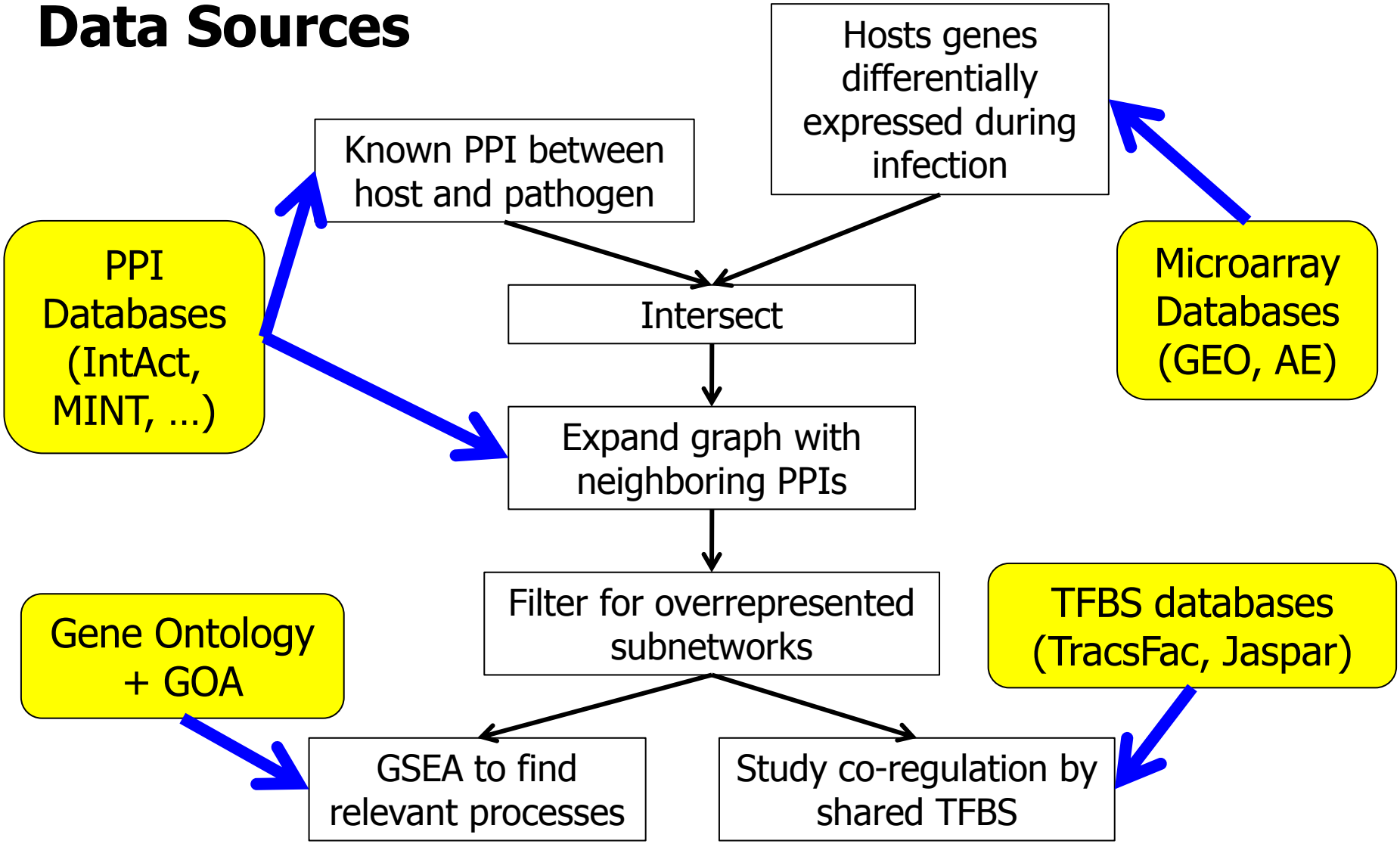
- Distributed information
  - Plenty of distributed DB
- Semantic heterogeneity
- Scientific uncertainty and evolving concepts
  - some scientific discoveries break databases: Gene → Protein
- Naming conventions on the object level
  - Impossible to *a priori* know that two sequences belong to the same gene!
- Naming conventions on the concept level
  - Definition of gene (ORF, coding sequence,…)
- Inclusion of non-standard processing
  - Sequence similarity, homology (evolution), …

Data Integration in 2012

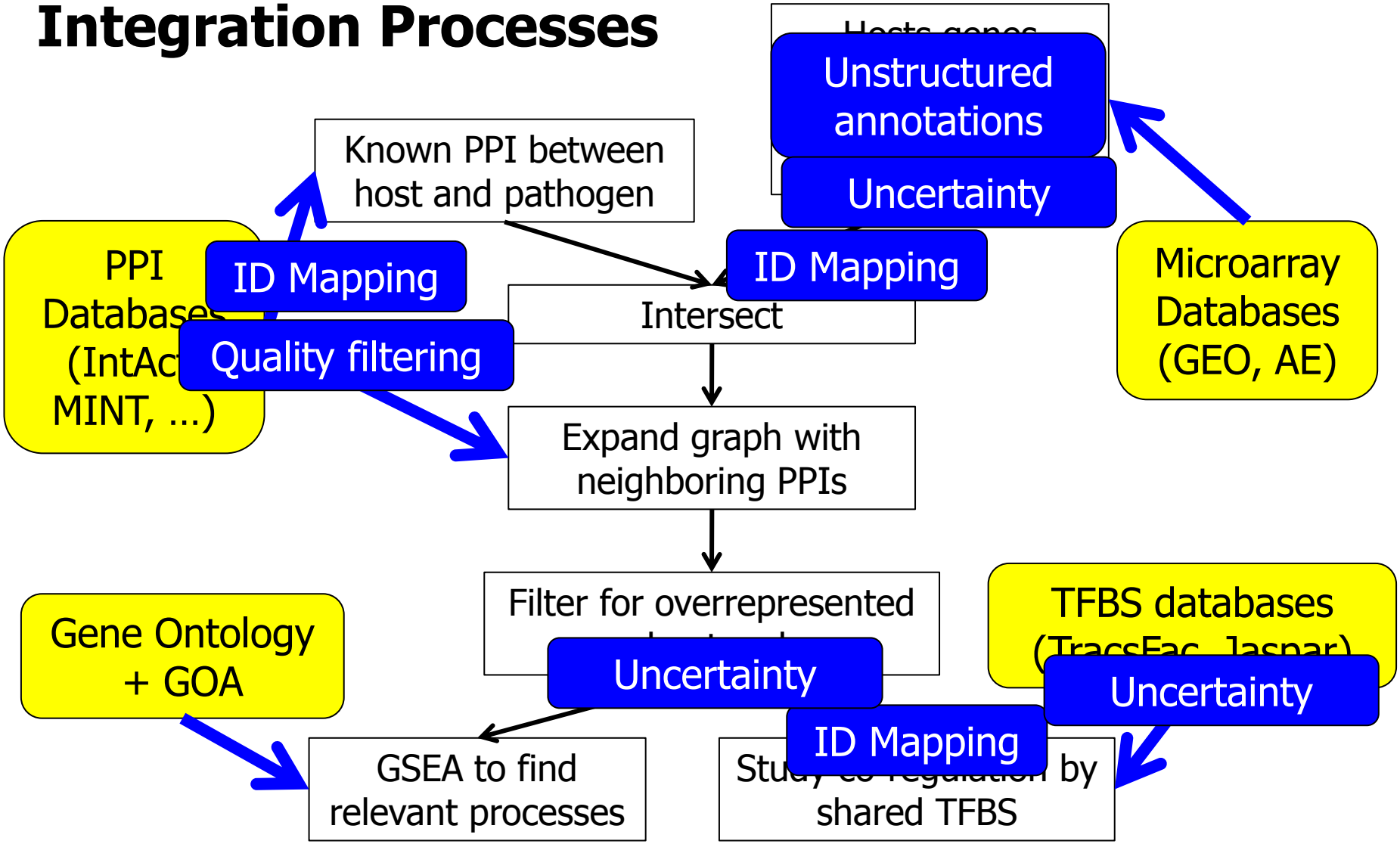
- *Query*: Which are the genes that play a central role in the response of a host to a pathogen?
  - Bacteria / viruses must attach to cells to have an influence
  - Attachment is a physical binding of proteins (PPI)
  - This binding provokes a reaction in the cell, transmitted by more PPI (e.g. transient signaling)



Data Integration?
Data Sources



Data Integration? Integration Processes

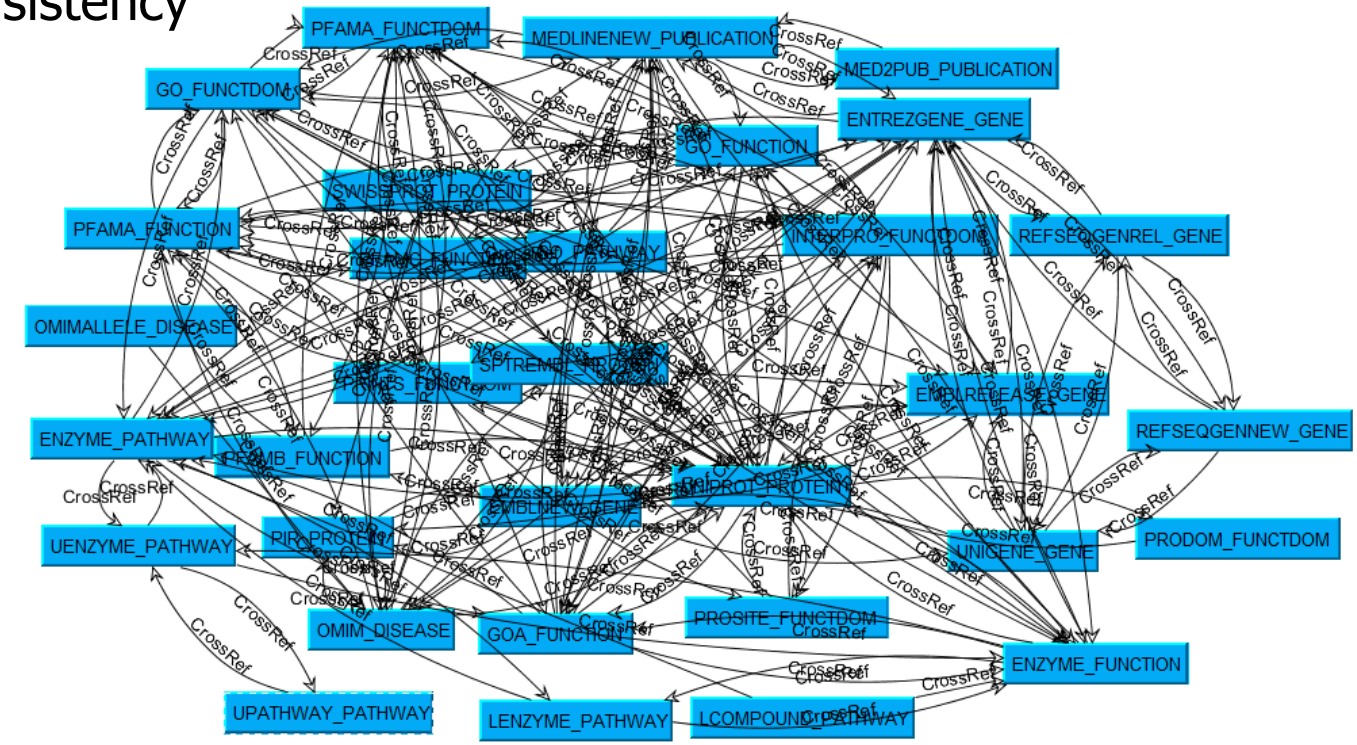


Biological Databases (1/2): contain expertise

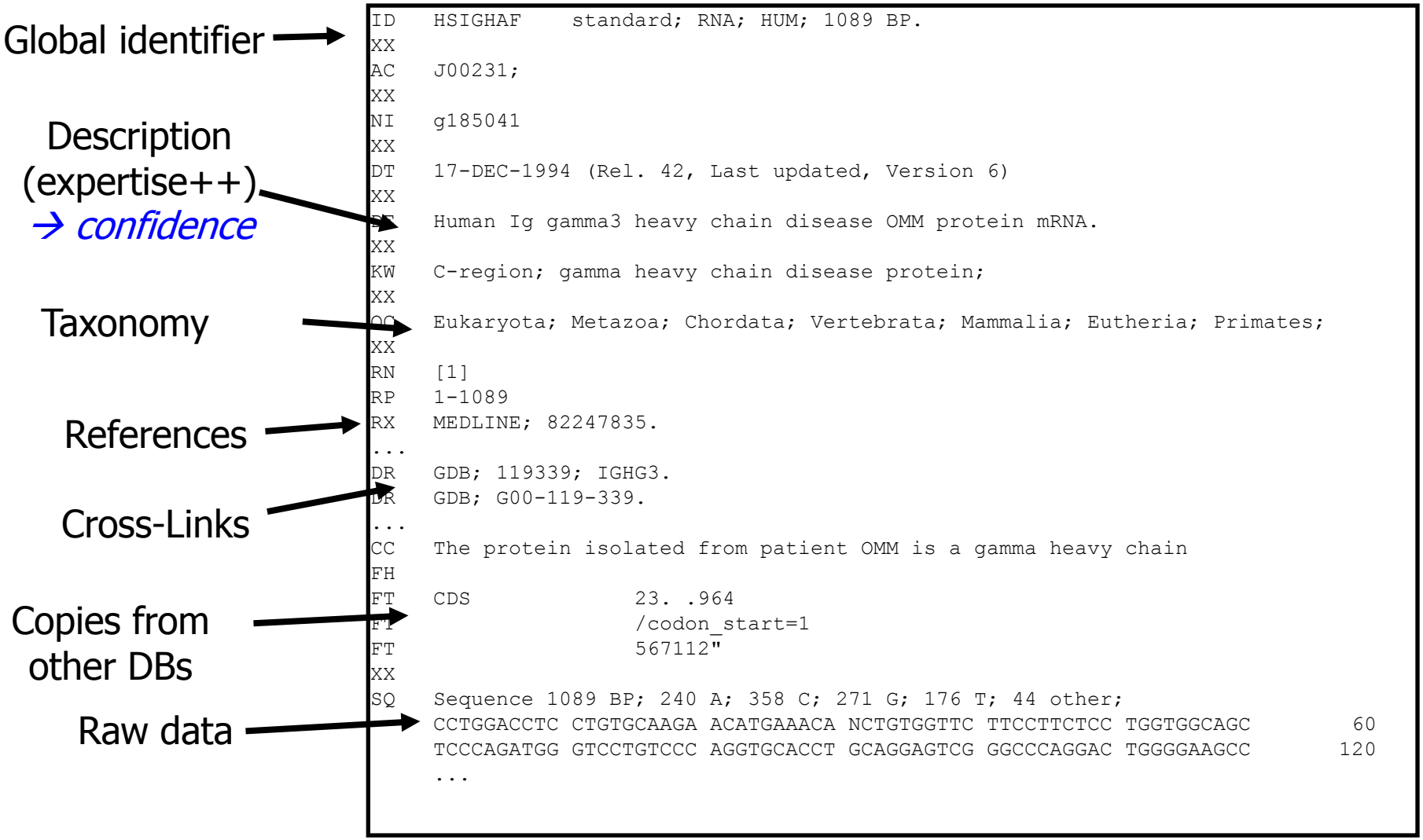
- Primary – secondary
  - Taking experimental data (sequences) are conclusions drawn from experiments (genes)
  - Relatively few primary (20), many secondary (100reds): >1 000 total !
- Curated or not
  - Most secondary databases are created manually
  - Many of them by reading and summarizing , copy-pasting / modifying (curation)
  - Issues: Consistency, completeness, quality assurance, objectivity, …
  - data = data + origin (otherwise impossible to interpret it)
- Some primary databases are international de-facto standard
  - Sequences: Genbank, proteins: UniProt, structures: PDB, …

Biological data (2/2): Links, links, links…

- BDB maintain links to many other BDBs
  - Instance level - external IDS, web browsing support
  - d1 → d2: "authors of d1 think that d2 *is related to* d1"
- No central authority for ID or links
- No consistency



A Biological Database



Take Home Message

- The complexity of the questions to be answered has increased a lot
- The diversity of the sources (and data) has increased a lot
- The number of sources to be used has increased a lot

Emergence of New Trends

- The complexity of the questions to be answered has increased a lot
  - Integration requires analysis and analysis requires integration
  - Scientific workflows

- The diversity of the sources has increased a lot
  - Inclusion of quality as a first-class citizen
  - Ranking of integrated search results

- The number of sources to be used has increased a lot
  - Scalability of integration in number of sources
  - One major goal of the Semantic Web

This Tutorial

- Part I – Integrating Life Sciences data
- Part II – Past and Presence
- Part III – Current Trends
- Part IV – Conclusions

Approaches in the 2000's

	SRS	Kleisli	Discoverylink	Tambis
Global schema	No	No	Views	Yes, Description Logics & ontology
Distributed data	No (later added)	Yes	Not in focus	Yes (Kleisli)
Virtual	Yes	Yes	Yes	Yes
Global data model	(XML like)	Nested collections	Relational	Description Logics
Data handling	No	No	No	No
Process integration	Limited	No	No	No

Impact in the Life Sciences

→ Except SRS / Entrez, systems were ignored in the LS community

→ None of the DB-drive systems still in use today!

Possible Explanations…

- Focused on schemas, while biologists focus on data
  - Content is king
- Virtual integration prevents changing the data
  - Statistical integration often needs to manipulate/mine data
- Transparency hides provenance as indicator for quality
- Approaches to data integration all were domain independent
  - Genes cannot be compared with the same methods as person names – different error models, different primary data, different additional data, different types of "equality"
- DR projects target discrete integration, while Life Scientists think in statistical integration
  - Schema, queries, mappings, …
  - Sequence alignment, normal distribution, error models, …

The Presence (from LS perspective)

- XML + (Perl | Java | Python) + (MySQL | Oracle | PostGreSql)
- Big role of open source libraries (BioSQL) + Ontologies

→ Probably >95% of integration projects use materialization

- Successful systems implemented by domain scientists, little participation of DR
- Very little semantic integration, very little query optimization, very little data fusion, very little schema matching / schema integration
- Full provenance information

This Tutorial

- Part I – Integrating Life Science data
- Part II – Past and Presence
- Part III – Next Generation Data Integration
  - Data integration workflows
  - Semantic Web
  - Ranking in integrated datasets
- Part IV – Conclusions

Lesson's Learned – 4 Criteria

Effort	Integrating dozens of data sources still requires considerable effort
Provenance	Provenance beats transparency
Analysis	99% of users are not happy with queries (but integrated data sets are a pre-requisite for integrated analysis)
Quality	Often, the problem is „finding the right answer", not „finding any answer"

Three Trends

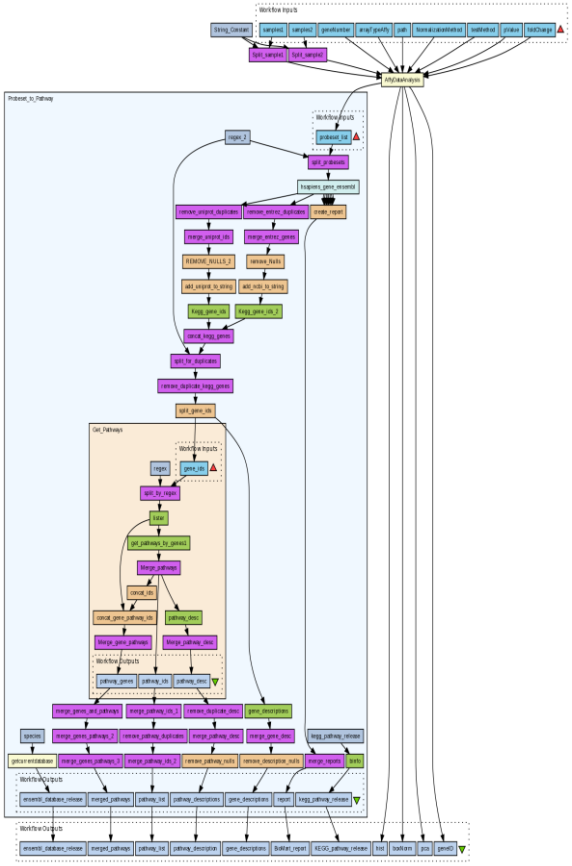
DI workflows	• Integration means analysis, and analysis means integration • No schemas, no explicit semantics • Scientific workflow systems	Effort: Analysis: Provenance: Quality:	~ ++ ++ +
Ranking	• Report results in a biologically meaningful order • Stays with queries, adds ranking • Requires a DI system in place	Effort: Analysis: Provenance: Quality:	~ - ++ ++
Semantic Web	• Reduce upfront cost of DI • No schemas, explicit semantics • Semantic Web tech. (RDF, SPARQL)	Effort: Analysis: Provenance: Quality:	++ - + -

This Tutorial

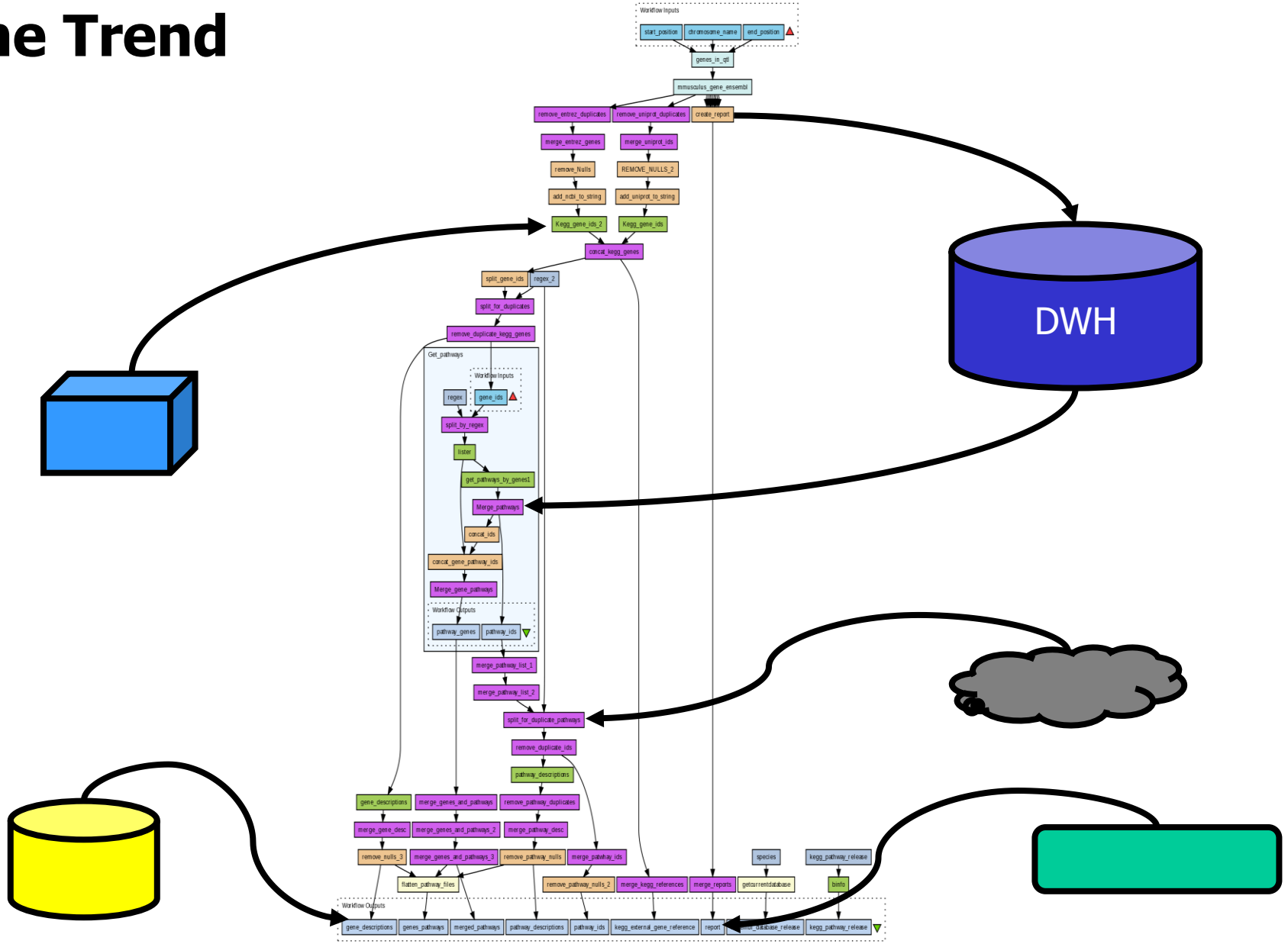
- Part I – Data Integration for the Life Sciences
- Part II – Past and Presence
- Part III – Current Trends
  - Data integration workflows
  - Semantic Web
  - Ranking in integrated datasets
- Part IV – Conclusions

Data Integration Workflows

- No separation between integration and analysis
- Complex pipelines broken into
  - tasks (accessing DB, using bioinformatics tool)
  - output to input ports connections
- Data flow driven
- Scientific Workflow Management System perform
  - scheduling and parallelization, logging, debugging, provenance (reproducibility), integration of web services, recovery…
- Plenty of systems available: Taverna, Kepler, Galaxy, Triana, Wings, …
- Independent of materialization/virtualization – pick what is best
- Tasks can be executed locally or distributed (web services)



The Trend

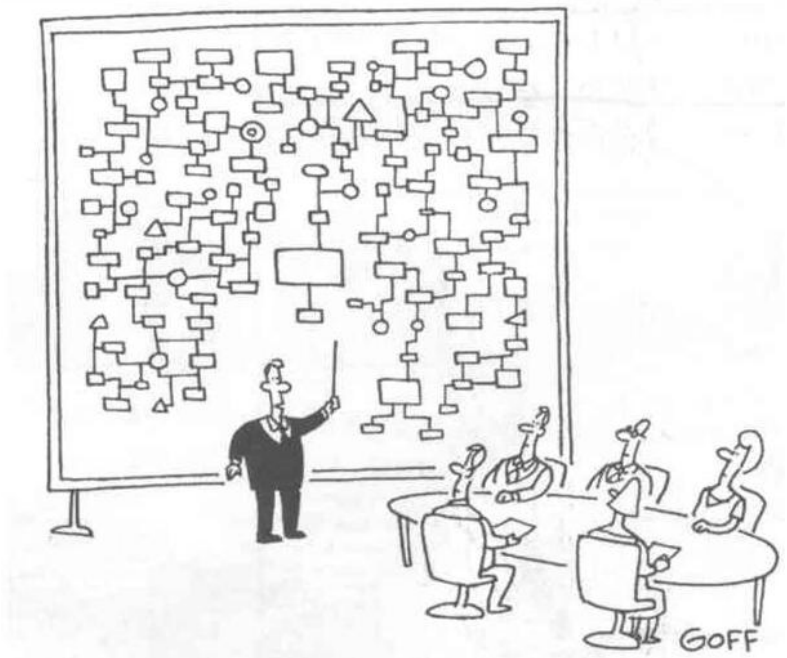

Workflow Inputs
start_position
chromosome_name
end_position
genes_in_qtl
mmusculus_gene_ensembl
remove_entrez_duplicates
remove_uniprot_duplicates
create_report
merge_entrez_genes
merge_uniprot_ids
remove_Nulls
REMOVE_NULLS_2
add_ncbi_to_string
add_uniprot_to_string
Kegg_gene_ids_2
Kegg_gene_ids
concat_kegg_genes
split_gene_ids
regex_2
split_for_duplicates
remove_duplicate_kegg_genes
Get_pathways
Workflow Inputs
regex
gene_ids
split_by_regex
lister
get_pathways_by_genes1
Merge_pathways
concat_ids
concat_gene_pathway_ids
Merge_gene_pathways
Workflow Outputs
pathway_genes
pathway_ids
merge_pathway_list_1
merge_pathway_list_2
split_for_duplicate_pathways
remove_duplicate_ids
pathway_descriptions
gene_descriptions
merge_genes_and_pathways
remove_pathway_duplicates
merge_gene_desc
merge_genes_and_pathways_2
merge_pathway_desc
remove_nulls_3
merge_genes_and_pathways_3
remove_pathway_nulls
merge_pathway_ids
species
kegg_pathway_release
flatten_pathway_files
remove_pathway_nulls_2
merge_kegg_references
merge_reports
getcurrentdatabase
binfo
Workflow Outputs
gene_descriptions
genes_pathways
merged_pathways
pathway_descriptions
pathway_ids
kegg_external_gene_reference
report
kegg_pathway_release
DWH

But …

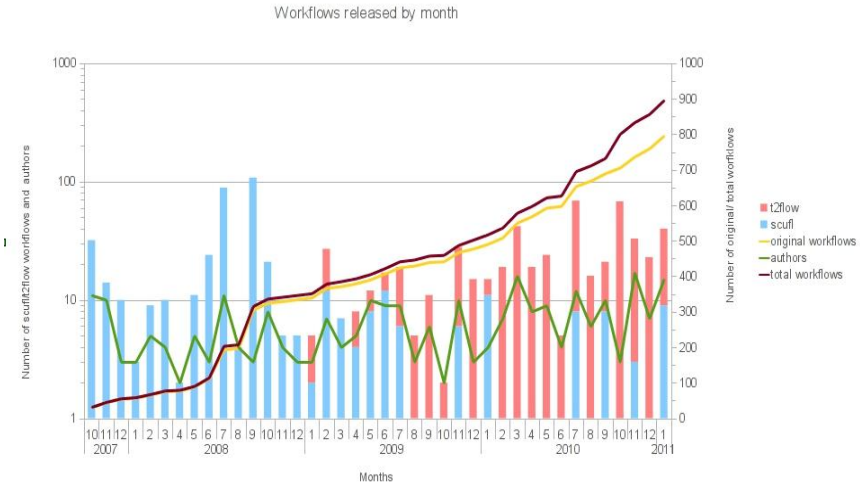
- What do we save compared to Perl?
  - Workflows are not easier to read than Perl programs (plumbing workflows)

- Obviously, Perl doesn't do
  - Automatic logging of all steps
    - Reproducibility, credibility
  - Automatic scheduling on available hardware
  - Automatic restart in case of failure
  - Tasks grouping opportunities (sub-workflows)

  → Sharing opportunities



- > 1,900 workflows available for immediate download
- Cross-system: Taverna, Triana, Kepler
- Social functionality: Tagging, rating, usage statistics
- But reuse could be improved [SCB+12]



Opportunities (and Untackeled Problems)

1. Support for workflow sharing
2. Support for execution scheduling (parallelism)
3. Support for typical integration tasks
   - Reducing amount of repeated work

- Towards ultimate scientific justification of results
  - Upload your analysis runs

1. Finding the Right Workflow

- Current repositories can be searched only with IR-style queries on workflow metadata / documentation
- Open question: Query languages for workflow repositories
  - Given a high-level description of a (integration) task – a sketch
  - Given a input and/or and output format/type
  - Given a workflow – find similar workflows
  - Find workflows (global similarity) or sub-workflows (local similarity)
  - Search across workflow models (Taverna and Kepler)
- Core of the problem: Workflow similarity
  - Metadata similarity (information retrieval)
  - Topological similarity (graph matching)
  - Semantic similarity (using (formal) task and type descriptions)
- Becomes a practical topic only now: Large repositories are available

Related Work

- Using only topological properties [GLG06]
  - Ignoring WF metadata and task descriptions
- Topological similarity in series-parallel graphs [ZCBD+09]
  - Captures a large class of workflow graphs
  - Can be solved in polynomial time
- Query languages from the business workflow community
  - BPQL, BPMN-Q [AS10], BP-QL [BEKM08], …
  - Do not include notion of similarity not local (sub-workflow) matches
  - Bound to workflow specification languages (BPEL)
- Query languages for repositories of workflow runs [KSB10, MPB10]
  - Querying the log of a workflow execution to find, e.g., the lineage of a specific result /trace
- Queries for filtering workflow runs [BCB+08]
  - Definition of views to filter relevant from irrelevant

2. Enhance parallel executions

- Scientific workflows form huge graphs that need to be executed
- Multi-core, clusters, grids and clouds should be much more exploited

- Top-down approach (improving existing workflows)
  - Decompose existing workflows into little tasks that can be parallelized
  - … While providing ways to hide the complexity! (groups of tasks)

- Bottom-up approach (better design from scratch)
  - help scientists design workflows easier to parallelize
  - Provide tasks implemented to be parallelized and then group them

Related Work

- Around Cloud computing
  - Map/Reduce, PACT, and second order functions
  - Stratosphere project [cf Volker Markl tutorial]
- Skeleton programming, Workflow patterns
  - Bioinformatics [AIG12]
  - Business [AHK+03]
- Libraries of bioinformatics tasks implemented for parallelism
  - Developed in the context of grids
  - Dedicated to domains: sequence comparison, PPI, …

- Graphs structures e.g. series-parallel graphs [ZCBD+09]
  - Towards workflows rewriting from DAGs to SP-graphs

3. Supporting DI Tasks

- Integration tasks are typically data-intensive and time-consuming
- Especially during WF development, such tasks need to be executed again and again
- Storing and reusing intermediate results can be of high benefit
  - Transparent materialization and reuse (caching)
- Open problem: Savepoints in SWFS
  - How to define (language, graphical)?
  - Who places them into a WF (manual or automatic)?
  - Mapping of results to workflow steps?
  - Efficiently storing and reusing the data

- Note: Results depend on concrete data (executions), workflows do not
- Note: Storing inputs together with results also enhances reproducibility

Related Work on "Smart Recomputation"

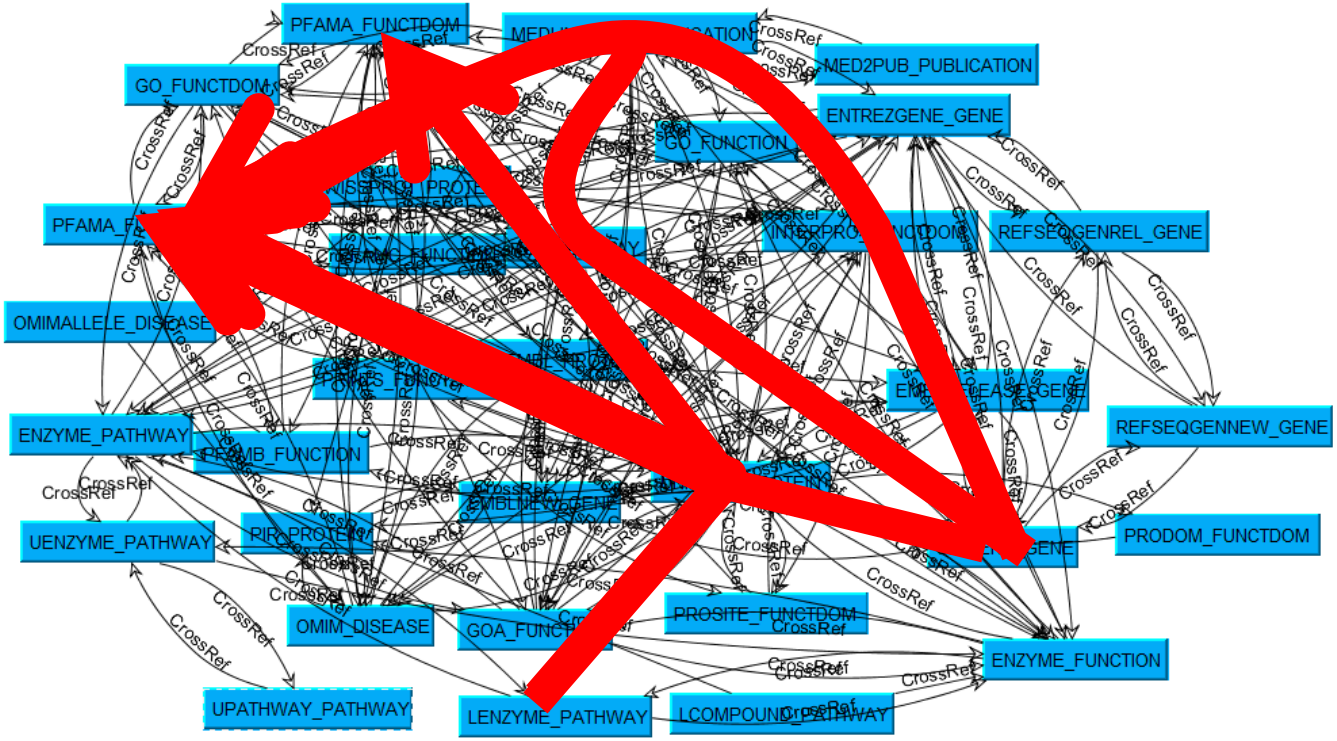
- Caching in general
- Strong Links [KSB+10]
  - Mapping of files using signature of "upstream" workflow
  - Support for post-WF analysis (which runs have used this file?)
- Smart re-computation [LAF+06]
  - Moves responsibility to the file system
  - Requires tight integration with SWFS

- Nice summary in "Managing Scientific Data", CACM 2010, [AKD10]

Three Trends

DI workflows	• Integration means analysis, and analysis means integration • No schemas, no explicit semantics • Scientific workflow systems	Effort: Analysis: Provenance: Quality:	~ ++ ++ +
Ranking	• Report results in a biologically meaningful order • Stays with queries, adds ranking • Requires a DI system in place	Effort: Analysis: Provenance: Quality:	~ - ++ ++
Semantic Web	• Reduce upfront cost of DI • No schemas, explicit semantics • Semantic Web tech. (RDF, SPARQL)	Effort: Analysis: Provenance: Quality:	++ - + -

Recall: Links



Report all GO annotation for a given protein

Ranking of Search Results

- Basis 1: Most types of objects are represented in multiple sources
- Basis 2: Sources link to each other (extensively yet unsystematically)
- For any query X->*->Y, there usually exist multiple paths
  - Which paths are the best?
  - Which paths produce results with the highest "relevance" to the query?
- Each path from a query X->*->Y produce an excessive number of results
  - Which results are the best
  - Which results have the highest relevance to the query?

- Alternative to data fusion: present best choice
  - In the LS, there are no "true" answers but diverging experiments

Relevant for Relevance

User provided	• Assessment of quality of used data sources • Assessment of quality of links • Freshness, completeness, trust, …
Query dependent	• Number of paths • Length of paths
Domain specific	• Similarity of linked sequences • Quality of matching leading to a link • …
Graph intrinsic	• Topology of the graph
Technical issues	• Execution time (joins, distributed query optimization) • Budget-based optimization • Best-effort optimization

Current approaches in Bio data ranking

- Systems require an underlying data access (integration) infrastructure
  - Can be central (BioZon, Columba) or distributed (BioGuide)
- Systems require different degrees of human intervention
  - Fully automatic (BioZon, Columba) to human-driven (BioGuide, BioRank)
- Ranking considerations may have an influence on query execution
  - Not here, but other projects, e.g. [BLM+04, NLF99]
- To date, no in-use DI system implements sensible a ranking method

Opportunities

1. Obtaining confidence scores
2. Consensus rankings
3. Comparable, objective evaluation strategies

1. Obtaining Scores

- We need (confidence, probabilistic) scores for data sources, link sources, objects, links
- Quality of biological databases is a much discussed issue, but difficult to map into a single value
- Link sets are incomplete, inconsistent, outdated
- Computed scores cannot be used directly
  - Sequence similarity of 40% in proteins -> very likely same function
  - Sequence similarity of 40% in genes -> no statement about function
- User-defined preferences are hard to specify and to obtain

Work in this Direction

- Quality of biological databases [BCF+07, BBF+01, MNF03]
  - Often completeness / currentness
  - Measuring „degree of truth" is notoriously difficult – different experiments, different results
- Quality criteria / user preferences [NLF99, BFL+04]
- Learning user preferences from relevance feedback [TJM+08]
  - Based on BioGuide system
- Robustness of ranking [DGL+09]
  - With respect to small derivations in preference scores
- Interactive search processes are generally under-researched

2. Consensus rankings and frameworks

- Ranking can be performed using various criteria
- No one ranking function is the best
- Consensus ranking aims to make the most of a set of rankings
  - Highlight their common points
  - Minimize their disagreements
- Computing the median of a set of rankings under the Kendell-tau distance is NP-difficult
  - Finding the right distance
  - Considering ties (partial order)
  - Heuristics, approximate algorithms

- Various ranking solutions exist but it is still difficult to compare them
  - Variations of PageRank/objectRank…

3. Evaluation

- Probably the hardest problem
  - See evaluation in information retrieval
- Problem: To what and how should results be compared?
- How: Choice of metrics
  - Precision at k, average precision, ROC, …
- What option 1: Expert opinion
  - Favors the certain, ignores the surprising
  - Subjective (inter-annotator agreement?)
  - Not scalable
- What option 2: Gold standard data sets
  - No generally accepted gold standards exist - everybody uses its own

- Calls for a competition

Related Work

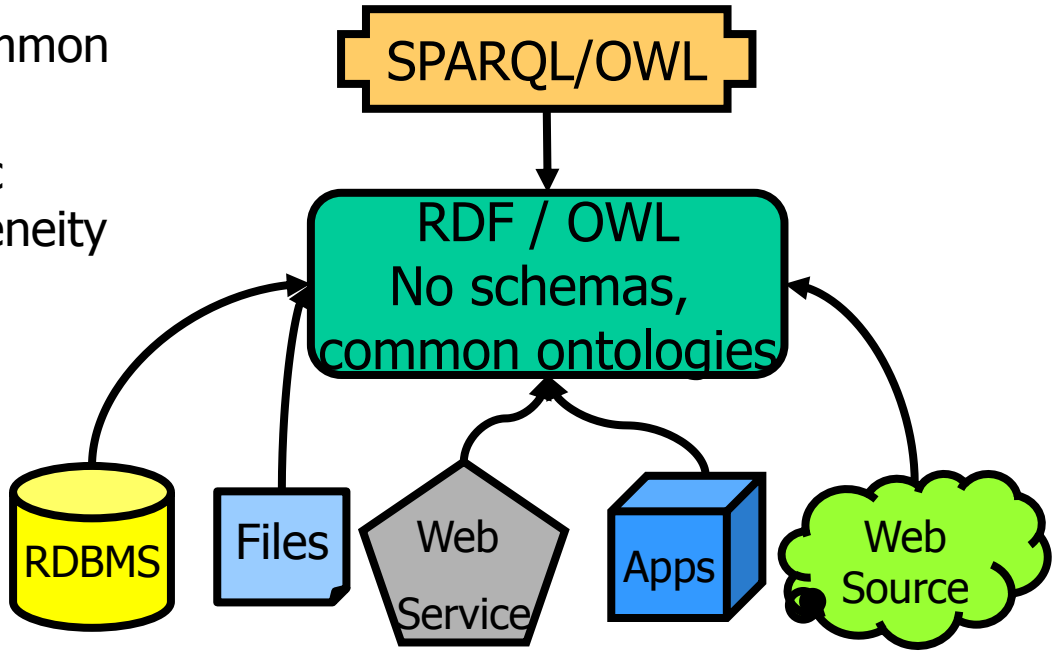
- Ranking in IR, especially on the web
  - Also combine textual with topological evidence
  - But: Unstructured entities, no entity classes, semantic-free links, different query types (no paths)
- Keyword searches in relational databases
  - Also consider paths through a data graph
  - Also may use class information
  - But: Different query types (subgraphs)
- Long tradition in AI research
  - Bayesian networks, fuzzy logic, Dempster-Shafer Theory of Believe, …
- Probabilistic databases
  - Highly similar setting
  - Also research on different semantics of uncertainty and on different methods for uncertainty propagation through a query network

Three Trends

DI workflows	• Integration means analysis, and analysis means integration • No schemas, no explicit semantics • Scientific workflow systems	Effort: Analysis: Provenance: Quality:	~ ++ ++ +
Ranking	• Report results in a biologically meaningful order • Stays with queries, adds ranking • Requires a DI system in place	Effort: Analysis: Provenance: Quality:	~ - ++ ++
Semantic Web	• Reduce upfront cost of DI • No schemas, explicit semantics • Semantic Web tech. (RDF, SPARQL)	Effort: Analysis: Provenance: Quality:	++ - + -

SemWeb for Data Integration

- Usual approach
  - RDFify everything: RDF as common data model
  - Ontologies cope with semantic instance and schema heterogeneity
  - SPARQL is the query language
- Sometimes: Use OWL for inferencing
  - Consistency of data sets, inference of new triples
  - Almost exclusively used: class, subclassOf, sameAs



- Examples in LS

  RDFifying is simple, many bio DBs are available in RDF: BioDash [NQ06], Rio2RDF [BNT+08], Chem2Bio2RDF [CDJ+10]….

Opportunities

1. RDF as data model
2. Extensions to SPARQL

2. RDF as Common Data Model

- RDF actually was meant to be a model for representing metadata
  - Discrete, certain facts
  - Geared towards logical inference
  - Numerical values not considered as such (no data types)
- But LS data is dirty
  - Dealing with uncertainty, contradictions, noise, …
  - Need for data fusion and efficient n-ary relationships
- But LS data can be voluminous
  - Experimental data
  - Need for hybrid approach
    - RDF for representing information (derived facts)
    - Links to original data sets in other format

3. Extensions to SPARQL

- Given the level of heterogeneity in merged RDF data sets, a powerful query language is a pre-requisite for comprehensive analysis
- However, SPARQL lacks
  - … grouping and aggregation for in-query de-duplication and data fusion
  - … user-defined predicates for implementing non-standard DI functions
  - … an understanding of class hierarchies to exploit semantic structures
  - … general transitive predicates to cope with heterogeneous schemas
  - … a sensible way to access multiple distributed RDF databases
  - … methods to cope with confidence / probabilities
- Early work in these directions has started
- None was applied to LS yet, maturity / scalability unclear

Work in these Directions

- Distributed SPARQL optimization
  - DARQ: Query rewriting based on predicate mappings [QL08]
  - Avalanche: SPARQL over Linked Open Data [BA10]
- Statistical aggregation in SPARQL [KT08]
  - Ad-hoc syntactic extension to SPARQL
- Using ontology mappings in query processing
  - Query rewriting using graph pattern rewriting [CSM+10]
  - SPARQL query rewriting using (relational) views [CWWM07]
- Transitive predicates for SPARQL [KAC+02, KJ07]
- Of course, using OWL for query rewriting
  - Not scalable [ZAV+07]

This Tutorial

- Part I – Data Integration for the Life Sciences
- Part II – Past and Presence
- Part III – Current Trends
- Part IV – Conclusions

Wrap-Up: Three Trends

- Scientists have an increasingly large support in IT (developers, computers, clusters…) in their own labs
- Several opportunities of research for database researchers
- DI workflows emphasize data analysis and may support DI by sharing
- Ranking focuses on providing meaningful answers despite questionable data quality
- Semantic Web approaches strive for cost reduction for initial DI phases

Acknowledgements

Ulf Leser

Taverna: Paolo Missier, Norman Paton

Kepler: Bertram Ludäscher, Shawn Bower

VisTrails: Juliana Freire

Susan Davidson, Johannes Starlinger, members of the AMIB/bioinfo group

Surveys

[DOB95] Davidson, S., Overton, G. C. and Buneman, P. (1995). "Challenges in Integrating Biological Data Sources." *Journal of Computational Biology* ***2(4): 557-572.***

[GS08] Goble, C. and Stevens, R. (2008). "State of the nation in data integration for bioinformatics." *J Biomed Inform* ***41(5): 687-93.***

[HK04] Hernandez, T. and Kambhampati, S. (2004). "Integration of Biological Sources: Current Systems and Challenges Ahead." *SIGMOD Record* ***33(5): 51-60.***

[Karp94] Karp, P. D. (1994). "Report of the Workshop on Interconnection of Molecular Biology Databases", SRI International Artificial Intelligence Center, Stanford, California.

[LC03] Lacroix, Z. and Critchlow, T., Eds. (2003). "Bioinformatics - Managing Scientific Data". San Francisco, California, Morgan Kaufmann Publishers.

[Sea05] Searls, D. B. (2005). "Data Integration: Challenges for Drug Discovery." *Nature Reviews Genetics* ***4: 45-58.***

[Ste03] Stein, L. D. (2003). "Integrating biological databases." *Nat Rev Genet* ***4(5): 337-45.***

[Ste08] Stein, L. D. (2008). "Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges." *Nature Reviews Genetics* ***9(9): 678-88.***

References

[AKD10] Ailamaki, A., Kantere, V. and Dash, D. (2010). "Managing scientific data." *Communications of the ACM* ***53(6): 68-78.***

[AHK+03] W.M.P van der Aalst, A.H.M. ter Hofstede, B. Kiepuszewski, and A.P. Barros. « Workflow Patterns. » Distributed and Parallel Databases, **14(3):5-51**, 2003.

[AIG12] Mohamed Abouelhoda, Shadi A Issa and Moustafa Ghanem. Tavaxy: Integrating Taverna and Galaxy workflows with cloud computing support. BMC Bioinformatics 2012, 13:77

[Ail10] Ailon, N. (2010). "Aggregation of Partial Rankings, p-Ratings and Top-m Lists." *Algorithmica* ***57(2): 284-300.***

[Alb09] Albrecht, A. (2009). "METL: Managing and Integrating ETL Processes". VLDB PhD workshop.

[ABE+08] Antezana, E., Blondé, W., Egana, M., Rutherford, A., Stevens, R., De Baets, B., Mironov, V. and Kuiper, M. (2008). "Structuring the life science resourceome for semantic systems biology: lessons from the BioGateway project". Workhop on Semantic Web Applications and Tools for Life Sciences, Edinburgh, UK.

[AS10] Awad, A. and Sakr, S. (2010). "Querying Graph-Based Repositories of Business Process Models ". Database Systems for Advanced Applications. pp 33-44.

[ZCBD+09] Bao, Z., Cohen-Boulakia, S., Davidson, S. B., Eyal, A. and Khanna, S. (2009). "Differencing Provenance in Scientific Workflows". EEE Int. Conf. on Data Engineering, IEEE Computer Society.

[BA10] Basca, C. and Bernstein, A. (2010). "Avalanche: Putting the Spirit of the Web back into Semantic Web Querying". Workshop on Scalable Semantic Web Knowledge Base Systems.

[BCF+07] Baumgartner Jr, W. A., Cohen, K. B., Fox, L. M., Acquaah-Mensah, G. and Hunter, L. (2007). "Manual curation is not sufficient for annotation of genomic databases." *Bioinformatics* ***23(13): i41.***

[BEKM08] Beeri, C., Eyal, A., Kamenkovich, S. and Milo, T. (2008). "Querying business processes with BP-QL." *Information Systems* ***33(6): 477-507.***

[BNT+08] Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P. and Morissette, J. (2008). "Bio2RDF: Towards a mashup to build bioinformatics knowledge systems." *Journal of Biomedical Informatics* ***41(5): 706-716.***

[BHL01] Berners-Lee, T., Hendler, J. and Lassila, O. (2001). "The Semantic Web." *Scientific American* ***284: 34-43.***

[BBF+01] Bhat, T. N., Bourne, P., Feng, Z., Gilliland, G., Jain, S., Ravichandran, V., Schneider, B., Schneider, K., Thanki, N., Weissig, H.*, et al. (2001). "The PDB data uniformity project." Nucleic Acids Res* ***29(1): 214-8.***

References

[BY06] Birkland, A. and Yona, G. (2006). "BIOZON: a system for unification, management and analysis of heterogeneous biological data." *BMC Bioinformatics* ***7: 70.***

[BCB+08] Biton, O., Cohen-Boulakia, S., Davidson, S. B. and Hara, C. S. (2008). "Querying and Managing Provenance through User Views in Scientific Workflows". 24th Int. Conf. on Data Engineering, IEEE Computer Society.

[BLM+04] Bleiholder, J., Lacroix, Z., Murthy, H., Naumann, F., Raschid, L. and Vidal, M.-E. (2004). "BioFast: Challenges in Exploring Linked Life Science Sources." *SIGMOD Record* ***33(2).***

[BL09] Böhm, C. and Leser, U. (2009). "Graph-Based Ontology Construction from Heterogeneous Evidences". Int. Semantic Web Conference (ISWC), Washington, US.

[CDJ+10] Chen, B., Dong, X., Jiao, D., Wang, H., Zhu, Q., Ding, Y. and Wild, D. J. (2010). "Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data." *BMC Bioinformatics* ***11: 255.***

[CWWM07] Chen, H., Wu, Z., Wang, H. and Mao, Y. (2006). "RDF/RDFS-based Relational Database Integration". International Conference on Data Engineeringing (ICDE), Atlanta, USA. pp 94.

[CBD+06] Cohen-Boulakia, S., Davidson, S. B., Froidevaux, C., Lacroix, Z. and Vidal, M.-E. (2006). "Path-based systems to guide scientists in the maze of biological data sources." *Journal of Bioinformatics and Computational Biology* ***4(5): 1069-1095.***

[BFL+04] Cohen-Boulakia, S., Froidevaux, C., Lair, S., Stransky, N., Radvanyi, F., Graziani, S. and Barillot, E. (2004). "Selecting Biomedical Data Sources According To User Preferences". Int. Conference on Intelligent Systems in Molecular Biology (ISMB/ECCB), Glasgow, UK.

[CSM+10] Correndo, G., Salvadores, M., Millard, I., Glaser, H. and Shadbolt, N. (2010). "SPARQL query rewriting for implementing data integration over linked data". EDBT Workshops, ACM. pp 1-11.

[CYS+08] Cusick, M. E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A. R., Simonis, N., Rual, J. F., Borick, H., Braun, P., Dreze, M.*, et al. (2009). "Literature-curated protein interaction datasets." Nat Methods* ***6(1): 39-46.***

References

[DCB+01] Davidson, S., Crabtree, J., Brunk, B. P., Schug, J., Tannen, V., Overton, G. C. and Stoecker Jr., C. J. (2001). "K2/Kleisli and GUS: Experiments in Integrated Access to Genomic Data Sources." *IBM Systems Journal* ***40(2): 512-531.***

[DGL+09] Detwiler, L., Gatterbauer, W., Louie, B., Suciu, D. and Tarczy-Hornoch, P. (2009). "Integrating and ranking uncertain scientific data". Int. Conf. on Data Engineering, Shanghai, CN, IEEE. pp 1235-1238.

[ES08] Euzenat, J. and Shvaiko, P. (2007). "Ontology matching". Heidelberg, Springer.

[FKM+06] Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D. and Vee, E. (2006). "Comparing partial rankings." *SIAM J. Discrete Mathematics* ***20(3): 628-648.***

[GBR+07] Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., Emanuelsson, O., Zhang, Z. D., Weissman, S. and Snyder, M. (2007). "What is a gene, post-ENCODE? History and updated definition." *Genome Res* ***17(6): 669-81.***

[GLG06] Goderis, A., Li, P. and Goble, C. (2006). "Workflow discovery: the problem, a case study from e-Science and a graph-based solution". Int. Conf. on Web Services, Chicago, Illinois.

[HBF+09] Hedeler, C., Belhajjame, K., Fernandes, A., Embury, S. and Paton, N. (2009). "Dimensions of dataspaces". 26th British National Conference on Databases. pp 55-66.

[HAA+09] Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., Bernabe, R. R., Bhan, M. K., Calvo, F., Eerola, I., Gerhard, D. S.*, et al. (2010). "International network of cancer genome projects." Nature* ***464(7291): 993-8.***

[HTL07] Hussels, P., Trissl, S. and Leser, U. (2007). "What's new? What's certain? Scoring Search Results in the Presence of Overlapping Data Sources". 4th Workshop on Data Integration in the Life Sciences, Philadelphia, US, Springer, Lecture Notes in Computer Science. pp 231-246.

[IBS08] Ilyas, I. F., Beskales, G. and Soliman, M. A. (2008). "A Survey of Top-k Query Processing Techniques in Relational Database Systems." *ACM Computing Surveys* ***40(4).***

References

[IPSN10] Ioannou, E., Papapetrou, O., Skoutas, D. and Nejdl, W. (2010). "Efficient Semantic-Aware Detection of Near Duplicate Resources". Extended Semantic Web Conference, pp 136-150.

[JAB+08] Jenkinson, A. M., Albrecht, M., Birney, E., Blankenburg, H., Down, T., Finn, R. D., Hermjakob, H., Hubbard, T. J., Jimenez, R. C., Jones, P., *et al. (2008). "Integrating Biological Data – The Distributed Annotation System ". Data Integration for the Life Sciences, Springer LNBI 5109, Evry, France.*

[JLP+10] Jonquet, C., LePendu, P., Falconer, S. M., Coulet, A., Noy, N. F., Musen, M. A. and Shah, N. H. (2010). "NCBO Resource Index: Ontology-Based Search and Mining of Biomedical Resources". Semantic Web Challenge, at ISWC10, Shanghai, CN.

[KAC+02] Karvounarakis, G., Alexaki, S., Christophides, V., Plexousakis, D. and Scholl, M. (2002). "RQL: a declarative query language for RDF". World Wide Web Conference, Honolulu, Hawaii, USA. pp 592-603.

[KKS04] Kasprzyk, A., Keefe, D., Smedley, D., London, D., Spooner, W., Melsopp, C., Hammond, M., Rocca-Serra, P., Cox, T. and Birney, E. (2004). "EnsMart: a generic system for fast and flexible access to biological data." *Genome Res* ***14(1): 160-9.***

[KTR07] Kirsten, T., Thor, A. and Rahm, E. (2007). "Instance-based matching of large life science ontologies". 4th Int. Conf. on Data Integration in the Life Sciences Philadelphia, USA. pp 172-187

[KP10] Klingstrom, T. and Plewczynski, D. (2010). "Protein-protein interaction and pathway databases, a graphical review." *Brief Bioinform.*

[KT08] Kobayashi, N. and Toyoda, T. (2008). "Statistical search on the Semantic Web." *Bioinformatics* ***24(7): 1002-10.***

[KJ07] Kochut, K. and Janik, M. (2007). "SPARQLeR: Extended Sparql for Semantic Association Discovery". European Conference on the Semantic Web, Innsbruch, Austria.

[KSD+11] Kozhenkov, S., Sedova, M., Dubinina, Y., Gupta, A., Ray, A., Ponomarenko, J. and Baitaluk, M. (2011). "BiologicalNetworks - tools enabling the integration of multi-scale data for the host-pathogen studies." *BMC Syst Biol* ***5: 7.***

References

[KSB10] Kumar, A. M., Shawn, B. and Bertram, L. (2010). "Techniques for efficiently querying scientific workflow provenance graphs". 13th Int. Conf. on Extending Database Technology. Lausanne, Switzerland, ACM.

[LPW+06] Lee, T. J., Pouliot, Y., Wagner, V., Gupta, P., Stringer-Calvert, D. W., Tenenbaum, J. D. and Karp, P. D. (2006). "BioWarehouse: a bioinformatics database warehouse toolkit." *BMC Bioinformatics* ***7: 170.***

[LLF08] Lemoine, F., Labedan, B. and Froidevaux, C. (2008). "GenoQuery: a new querying module for functional annotation in a genomic warehouse." *Bioinformatics* ***24(13): i322-9.***

[LAF+06] Liu, D. T., Abdulla, G. M., Franklin, M. J., Garlick, J. and Miller, M. (2006). "Data-preservation in scientific workflow middleware". Scientific and Statistical Database Management, Vienna, AU, IEEE. pp 49-58.

[MKP+05] Markowitz, V. M., Korzeniewski, F., Palaniappan, K., Szeto, E., Ivanova, N. and Kyrpides, N. C. (2005). "The integrated microbial genomes (IMG) system: a case study in biological data management". 31st Conference on Very Large Databases (VLDB), Trondheim, Norway.

[MPB10] Missier, P., Paton, N. W. and Belhajjame, K. (2010). "Fine-grained and efficient lineage querying of collection-based workflow provenance". Int. Conf. on Extending Database Technology, Lausanne, CH, ACM. pp 299-310.

[MSR+10] Missier, P., Soiland-Reyes, S., Owen, S., Tan, W., Nenadic, A., Dunlop, I., Williams, A., Oinn, T. and Goble, C. (2010). "Taverna, Reloaded". Scientific and Statistical Database Management Systems, Heidelberg, Germany.

[MNF03] Müller, H., Naumann, F. and Freytag, J.-C. (2003). "Data Quality in Genome Databases". Conference on Information Quality, Boston, US.

[NLF99] Naumann, F., Leser, U. and Freytag, J. C. (1999). "Quality-driven Integration of Heterogeneous Information Systems". 25th Conference on Very Large Database Systems, Edinburgh, UK. pp 447-458.

[NQ06] Neumann, E. K. and Quan, D. (2006). "BioDash: a Semantic Web dashboard for drug development". Pac Symp Biocomput, Hawai, US. pp 176-87.

[NSW+09] Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M. A., Chute, C. G.*, et al. (2009). "BioPortal: ontologies and integrated data resources at the click of a mouse." Nucleic Acids Res* ***37(Web Server issue): W170-3.***

References

[OV99] Oezsu, M. T. and Valduriez, P. (1999). "Principles of Distributed Database Systems". New Jersey, Prentice Hall, Inc.

[ORS+08] Olston, C., Reed, B., Srivastava, U., Kumar, R. and Tomkins, A. (2008). "Pig latin: a not-so-foreign language for data processing". SIGMOD Conference, Vancouver, CD, ACM. pp 1099-1110.

[PRM+07] Post, L. J., Roos, M., Marshall, M. S., van Driel, R. and Breit, T. M. (2007). "A semantic web approach applied to integrative bioinformatics experimentation: a biological use case with genomics data." *Bioinformatics* ***23(22): 3080-7.***

[QL08] Quilitz, B. and Leser, U. (2008). "Querying Distributed RDF Data Sources with SPARQL". European Semantic Web Conference (ESWC), Teneriffa, Spain.

[RKL07] Rahm, E., Kirsten, T. and Lange, J. (2007). "The GeWare data warehouse platform for the analysis of molecular-biological and clinical data." *Journal of Integrative Bioinformatics* ***4(1): 47.***

[RTA+05] Rahm, E., Thor, A., Aumueller, D., Do, H. H., Golovin, N. and Kirsten, T. (2005). "iFuice - Information Fusion utilizing Instance Correspondences and Peer Mappings". WebDB, Baltimore, USA. pp 7-12.

[Rob94a] Robbins, R. J. (1994). "Report of the invitational DOE Workshop on Genome Informatics I: Community Databases." *Journal of Computational Biology* ***3: 173-190.***

[RML05] Rother, K., Michalsky, E. and Leser, U. (2005). "How well are protein structures annotated in annotation databases?" *PROTEINS* ***60(4): 571-576.***

[RCB+07] Ruttenberg, A., Clark, T., Bug, W., Samwald, M., Bodenreider, O., Chen, H., Doherty, D., Forsberg, K., Gao, Y., Kashyap, V.*, et al. (2007). "Advancing translational research with the Semantic Web." BMC Bioinformatics* ***8 Suppl 3: S2.***

[SPLO05] Sarkans, U., Parkinson, H., Lara, G. G., Oezcimen, A., Sharma, A., Abeygunawardena, N., Contrino, S., Holloway, E., Rocca-Serra, P., Mukherjee, G.*, et al. (2005). "The ArrayExpress gene expression database: a software engineering and implementation perspective." Bioinformatics* ***21(8): 1495-501.***

References

[SV09] Savage, C. J. and Vickers, A. J. (2009). "Empirical study of data sharing by authors publishing in PLoS journals." *PLoS One* ***4(9): e7078.***

[SIY06] Shafer, P., Isganitis, T. and Yona, G. (2006). "Hubs of knowledge: using the functional link structure in Biozon to mine for biologically significant entities." *BMC Bioinformatics* ***7: 71.***

[SBJ+09] Shah, N. H., Bhatia, N., Jonquet, C., Rubin, D., Chiang, A. P. and Musen, M. A. (2009). "Comparison of concept recognizers for building the Open Biomedical Annotator." *BMC Bioinformatics* ***10 Suppl 9: S14.***

[SHX+05] Shah, S. P., Huang, Y., Xu, T., Yuen, M. M., Ling, J. and Ouellette, B. F. (2005). "Atlas - a data warehouse for integrative bioinformatics." *BMC Bioinformatics* ***6(1): 34.***

[SMB+04] Shaker, R., Mork, P., Brockenbrough, J. S., Donelson, L. and Tarczy-Hornoch, P. (2004). "The BioMediator System as a Tool for Integrating Biologic Databases on the Web". Information Integration on the Web.

[SAR+08] Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J.*, et al. (2007). "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration." Nat Biotechnol* ***25(11): 1251-5.***

[SMS+02] Stein, L. D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J. E., Harris, T. W., Arva, A.*, et al. (2002). "The generic genome browser: a building block for a model organism system database." Genome Res* ***12(10): 1599-610.***

[TJM+08] Talukdar, P. P., Jacob, M., Mehmood, M. S., Crammer, K., Ives, Z. G., Pereira, F. and Guha, S. (2008). "Learning to create data-integrating queries." *Proceedings of the VLDB Endowment* ***1(1): 785-796.***

[TRM+05] Trissl, S., Rother, K., Müller, H., Koch, I., Steinke, T., Preissner, R., Frömmel, C. and Leser, U. (2005). "Columba: An Integrated Database of Proteins, Structures, and Annotations." *BMC Bioinformatics* ***6:81.***

[YIF+08] Yu, Y., Isard, M., Fetterly, D., Budiu, M., Erlingsson, Ú., Gunda, P. K. and Currey, J. (2008). "DryadLINQ: A System for General-Purpose Distributed Data-Parallel Computing Using a High-Level Language". Symposium on Operating Systems Design and Implementation.

[ZAV+07] Zacharias, V., Abecker, A., Vrandecic, D., Borgi, I., Braun, S. and Schmidt, A. (2007). "Mind the Web". Workshop on New Forms of Reasoning for the Semantic Web, Busan, Korea.