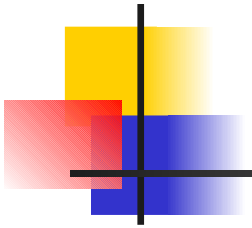
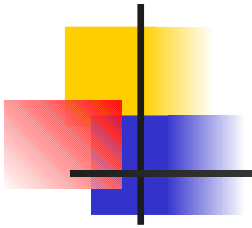


Data integration in the Life Sciences

problems and some current approaches

Presented at
Hasso-Plattner Institute, Potsdam
May 2008

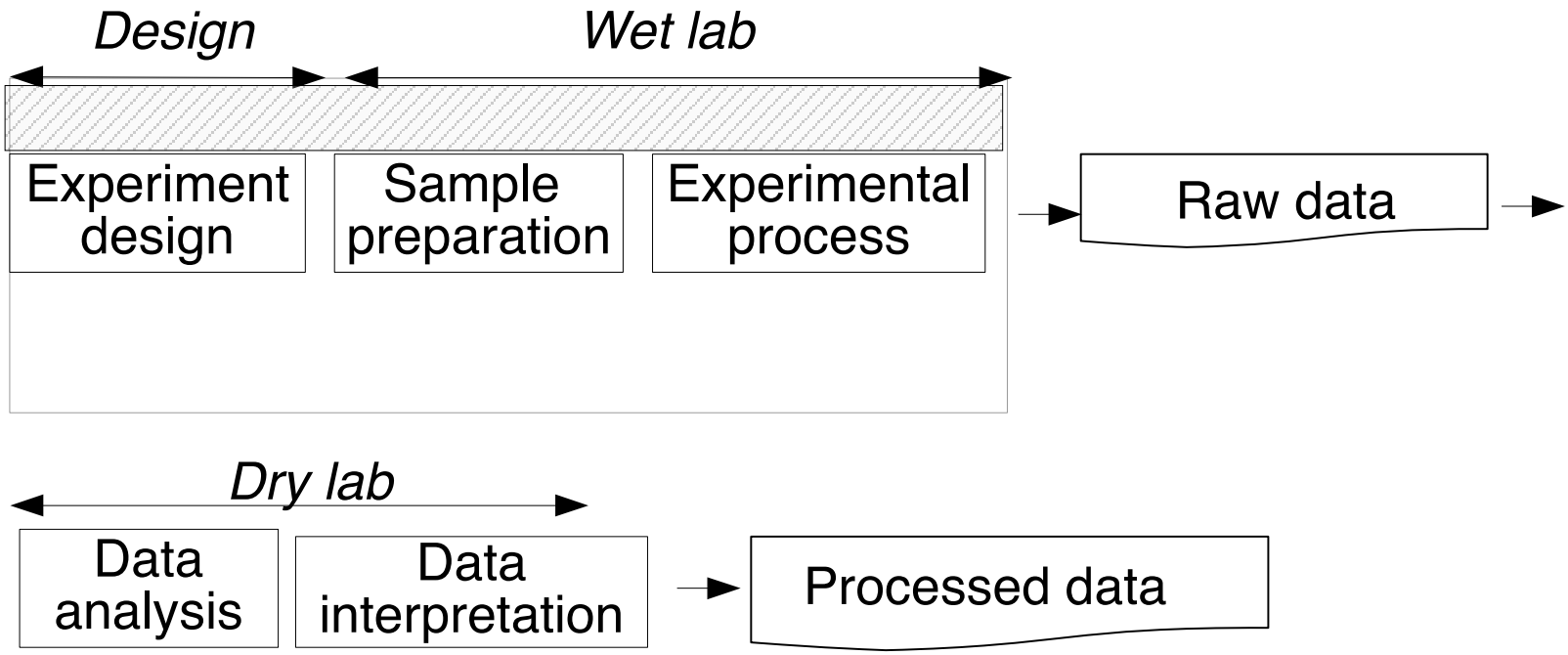
Paolo Missier
Information Management Group
University of Manchester, UK



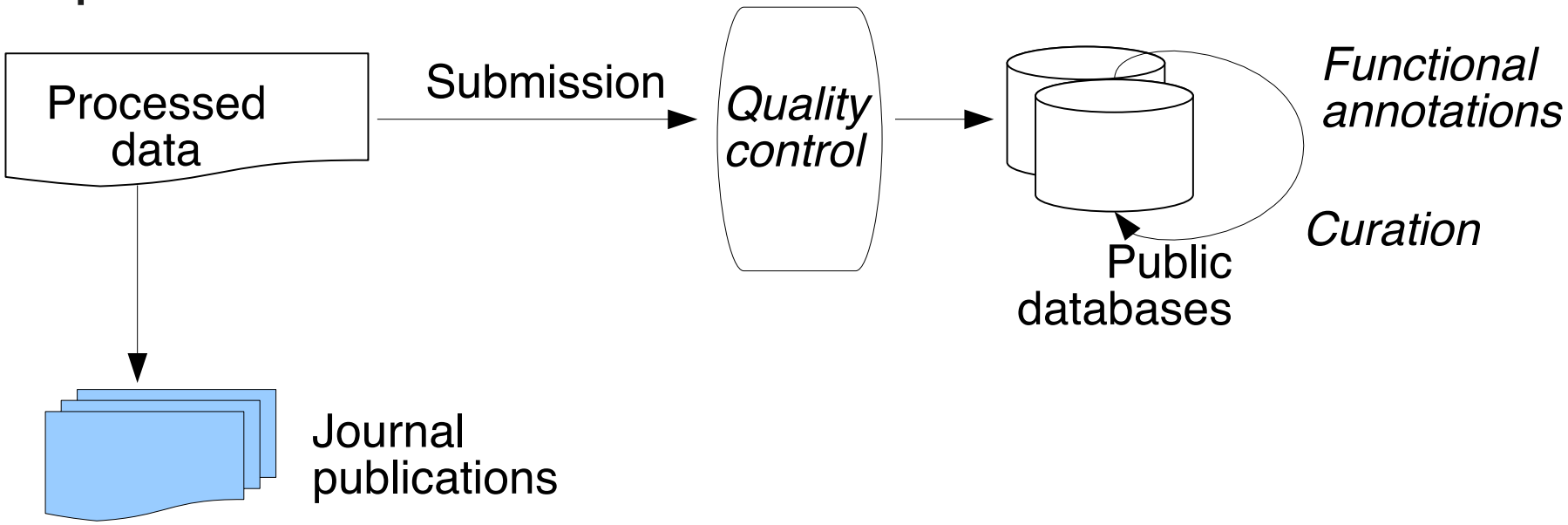
Outline

- The Life Sciences data landscape
- Why integrate?
  - the need for multi-source, cross-domain exploration and analysis
  - ex.: web-based navigation from proteins to pathways
- Challenges to integration
- Case studies
  - Tambis, SNPit, e-fungi, AutoMed
- Data integration through service integration

"Wet" and "dry" biology



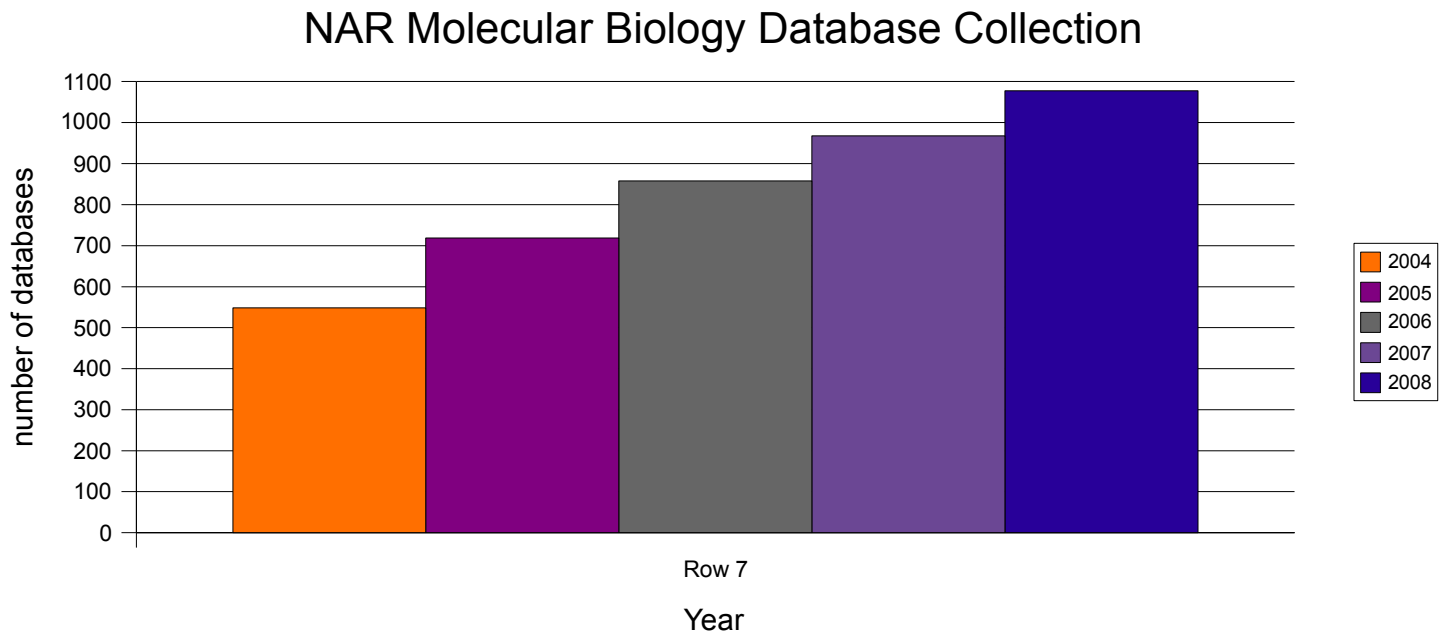
The public submission model

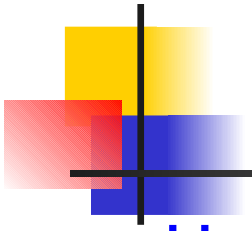


- The majority of LS DBs are community-based:
  - accept public submissions
  - offer free access to its content
- Some are manually curated

How many data sources?

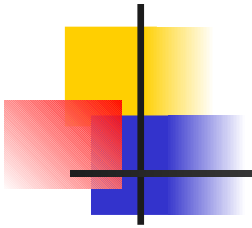
- *The Molecular Biology Database collection*
  - *representative of a large portion of Life Sciences DBs*
  - *updated yearly by the NAR journal (Nucleic Acid Research)*





A smaller set of databases...

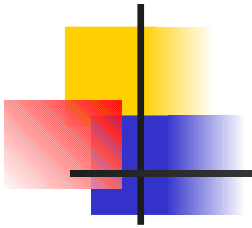
- Uniprot: Protein sequence databases
  - formerly Swiss-Prot
- Mouse Genome Database: Genomics Databases
- KEGG: Metabolic and Signaling Pathways
- Ensembl: Human genome databases, maps and viewers
- dbSNP (NCBI): polymorphism databases



Some example databases

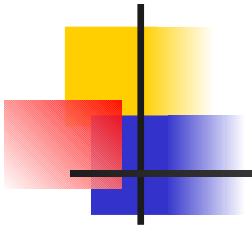
- Uniprot: (statistics)
  - protein knowledgeBase
  - incorporates Swiss-Prot
  - over 360K protein entries, manually curated (Swiss-Prot only)

- MGI: Mouse Genome Informatics
  - Jackson Lab (statistics)
  - mouse genome



More DBs

- KEGG (statistics)
  - Japanese project
  - Genomes, genes, metabolic pathways
- dbSNP (statistics)
  - Single Nucleotide polymorphism
  - an NCBI database
    - National Center for Biotechnology Information (US)
- Ensembl
  - EMBL-EBI (UK/EU) and the Sanger Institute, UK
  - primarily vertebrate genomes

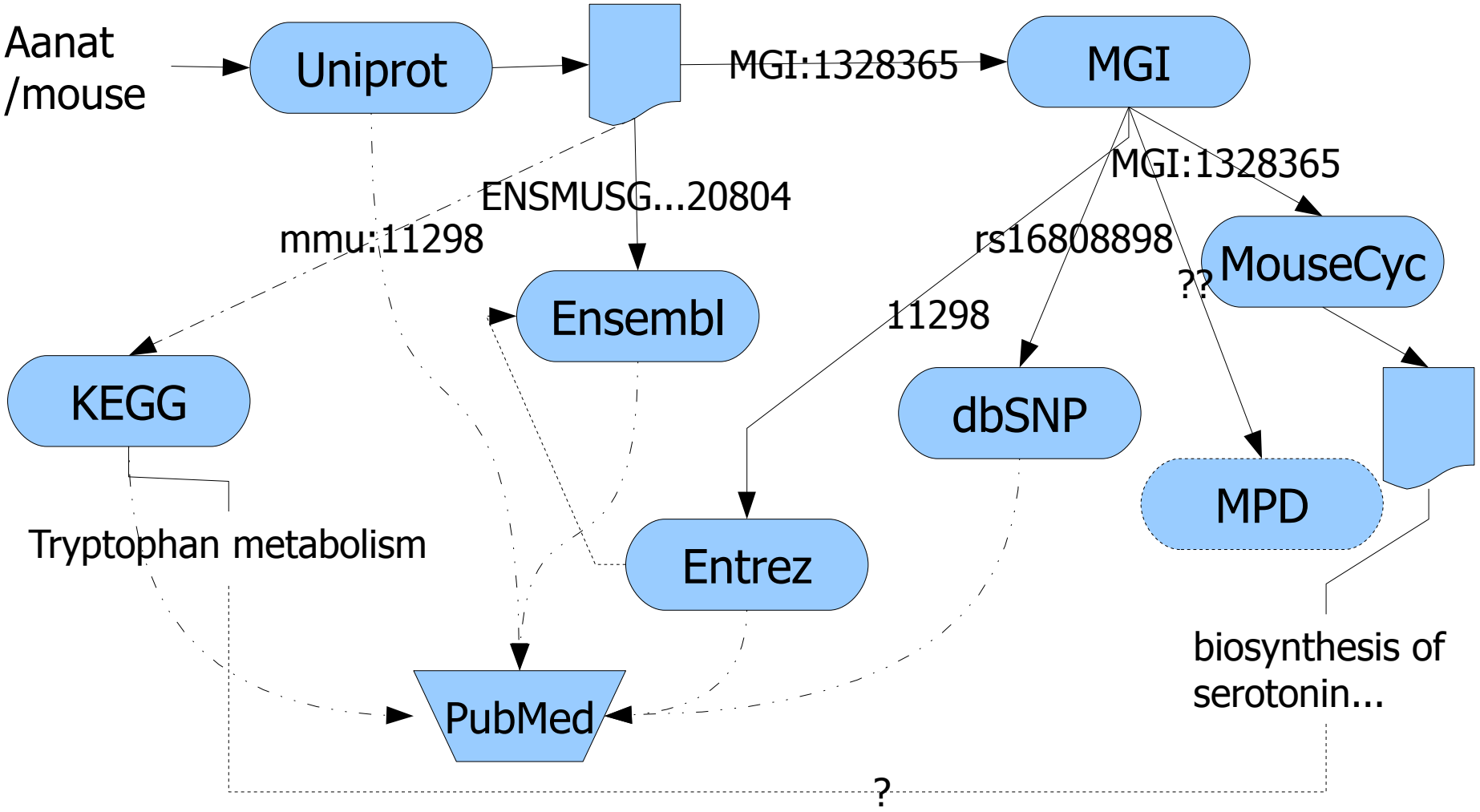


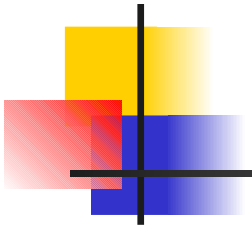
A multi-source search

Starting point: Aanat (Mus Musculus) in Uniprot

- found: Serotonin N-acetyltransferase
- cross-references to
  - KEGG entry mmu:1298 --> pathway mmu00380
    - Tryptophan metabolism
  - MGI:1328365
    - --> MouseCyc pathway: "biosynthesis of serotonin and melatonin"
    - SNPs (tens of SNPs for this gene)
      - -> dbSNP
      - -> MPD (Mouse Phenome Database)
    - ENSEMBL: 20804 ...

Cross-reference graph

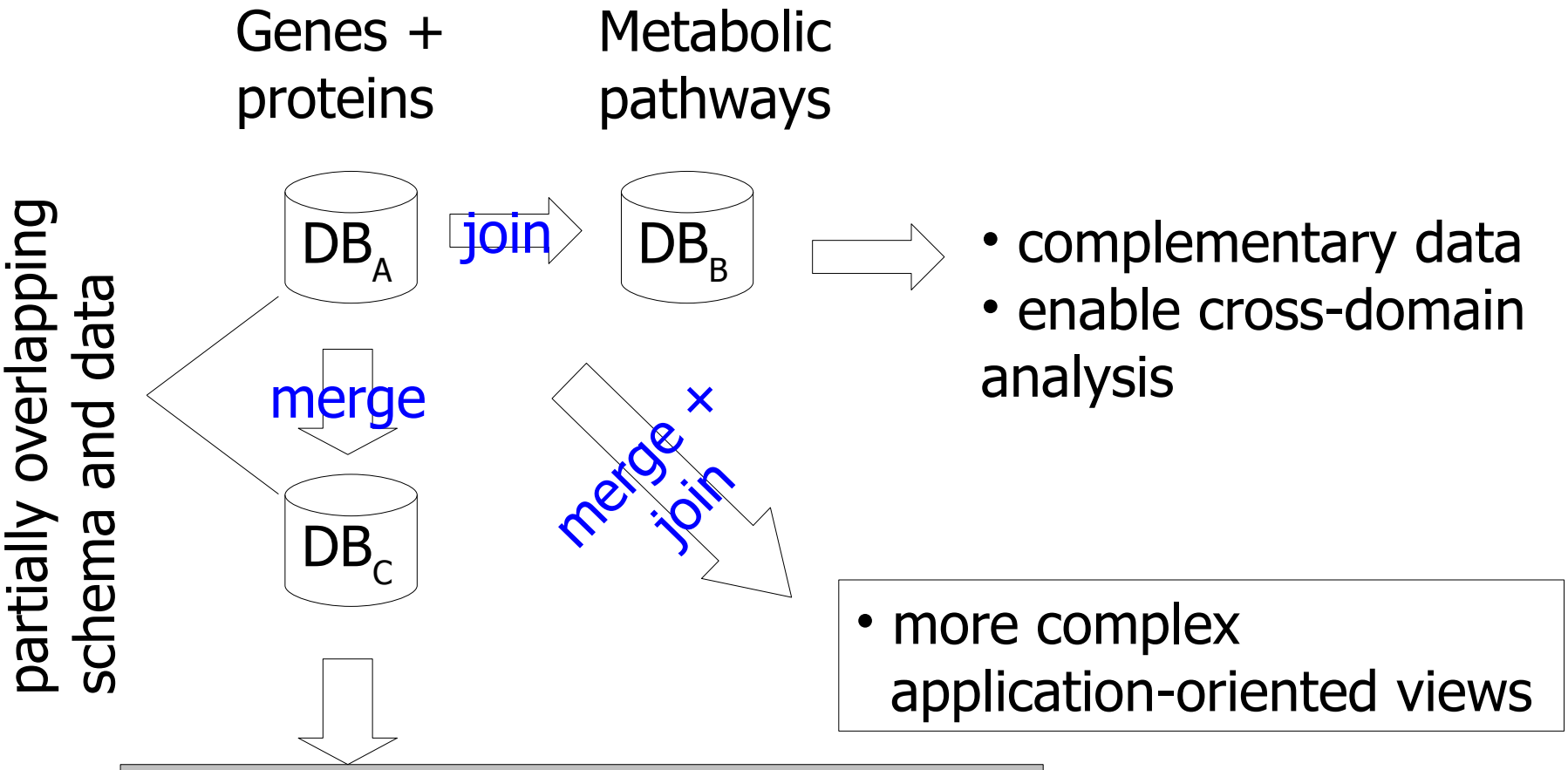


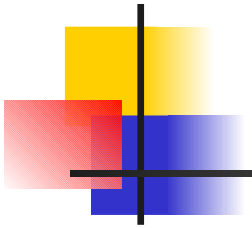


Web-based navigation

- Suitable for manual, small-scale exploration
  - Presentation layer
  - No integrated query capability
- Respectful of sites autonomy
  - only identifiers are shared

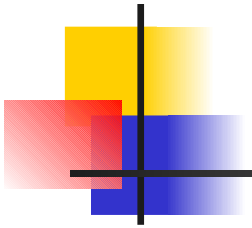
Integration scenarios





Main issues

- Overlapping sources for merge:
  - schema compatibility
  - data version alignment
  - inter-source data consistency --> fusion
- Complementary sources for join:
  - availability of consistent identifiers across sources



The identifiers issue

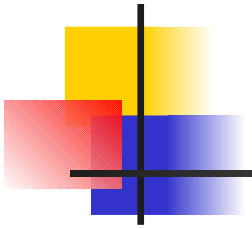
- Correct resolution of cross-references
  - weak "foreign key constraints"
- Different organizations use different / multiple identifiers
  - known and long-standing problem

Uniprot's "DR line"

```
DR   EMBL; AB013358; BAA31526.1; -; mRNA.
DR   EMBL; AF004108; AAD09408.1; -; mRNA.
DR   EMBL; U83462; AAD08637.1; -; Genomic_DNA.
DR   RefSeq; NP_033721.1; -.
DR   UniGene; Mm.418559; -.
DR   UniGene; Mm.42233; -.
DR   HSSP; Q29495; 1CJW.
DR   SMR; O88816; 23-194.
DR   PhosphoSite; O88816; -.
DR   Ensembl; ENSMUSG00000020804; Mus musculus.
DR   GeneID; 11298; -.
DR   KEGG; mmu:11298; -.
DR   MGI; MGI:1328365; Aanat.
DR   CleanEx; MM_AANAT; -.
DR   GermOnline; ENSMUSG00000020804; Mus musculus.
DR   GO; GO:0004059; F:aralkylamine N-acetyltransferase activity;
IEA:EC.
DR   InterPro; IPR016181; Acyl_CoA_acyltransferase.
DR   InterPro; IPR000182; GCN5-rel_AcTrfase.
DR   Gene3D; G3DSA:3.40.630.30; Acyl_CoA_acyltransferase; 1.
DR   Pfam; PF00583; Acetyltransf_1; 1.
DR   PROSITE; PS51186; GNAT; 1.
```

Uniprot primary accession number: O88816

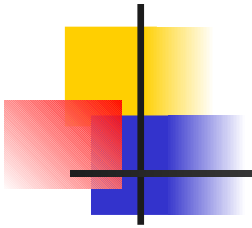
See compete list and x-ref rules



Indexing services

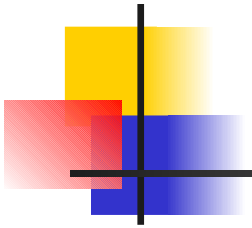
- Who maintains the mappings?
  - needs periodic sync points / updates

- DBGET is an integrated retrieval system maintained alongside KEGG (Japan)
  - its scope spans the major DBs for the most used data domains



Three approaches to integration

- Multidatabase queries, mediator systems
  - Propagate a global query to local data sources
- Warehousing
  - Focused, multi-source bioinformatics analysis
    - schema transformation, materialized views
    - expensive to create and maintain
- Ad hoc: scientific dataflows
  - user-defined processes that effectively perform on-the-fly, on demand integration
  - burden on user
  - useful new paths across sources graph

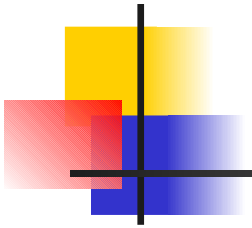


Example: TAMBIS

Problem: Querying over multiple, diverse data sources as if there was only one data source

Example analysis objective:
Select patterns in the proteins that invoke an immunological response and participate in programmed cell death, that are similar in their sequence of amino acids to the protein that is associated with triggering cell death in the white cells of the immune system.

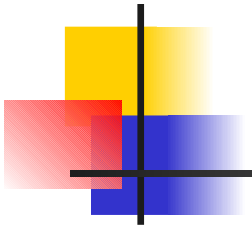
Query: Select motifs for antigenic human proteins that participate in apoptosis and are homologous to the lymphocyte associated receptor of death (also known as lard).



Integration challenges in Tambis

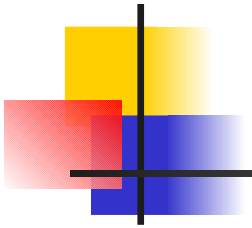
- The data resources are frequently not databases in the conventional sense
  - no explicit schema containing their meta-data
  - Need to construct own view of the meta-data in each source (the intension) and the instances covered by that source (the extension)

- Structural and content heterogeneity:
  - the global unique identifier of a protein is its accession number, but these are inconsistent between sources
  - resolving any semantic heterogeneities between the sources
    - SWISS-PROT covers some information on proteins
    - PROSITE covers motifs on protein sequences
    - BLAST is a tool for matching sequences



Heterogeneity

- Of format:
  - Construct the various parts of the request in the different formats and terms required by the different sources
  - Ad hoc data transformation to map across formats
- Of access paths:
  - Different access paths to data, no SQL interface
  - Most are tools, processes (e.g., sequence alignment), or proprietary flat file structures
  - The sources have complete autonomy



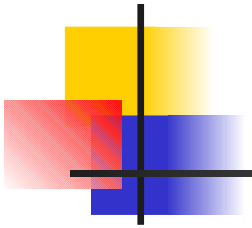
TAMBIS: ontology-based integration

Transparent Access to Multiple Bioinformatics Data Sources (1996-2001)

- an architecture for interactive, declarative expression of queries over heterogeneous sources

Distinctive feature:

- queries are expressed over an ontology of bioinformatics concepts
- rationale: Bioinformatics researchers have recognized that semantic schema and data matching could be aided by a comprehensive thesaurus of terms or a reusable reference ontology of biological concepts.



The TAMBIS ontology

- Ontology: a very expressive data model
  - around 1800 biological concepts and their relationships
  - expressed using a logic language
    - A particular Description Logic (precursor to OWL)
  - compositional constraints enable automated inference of further concepts
  - domain coverage: proteins and nucleic acids, their motifs, protein structure and structural classification, biological processes, and functions

The example query

```
Motif which
<isComponentOf (Protein which
<hasOrganismClassification Species
 FunctionsInProcess Apoptosis
 HasFunction Antigen isHomologousTo
 Protein which <hasName
                    ProteinName>)>)>
```

Species="human"
ProteinName="Lard"

concept

conjunctive operator

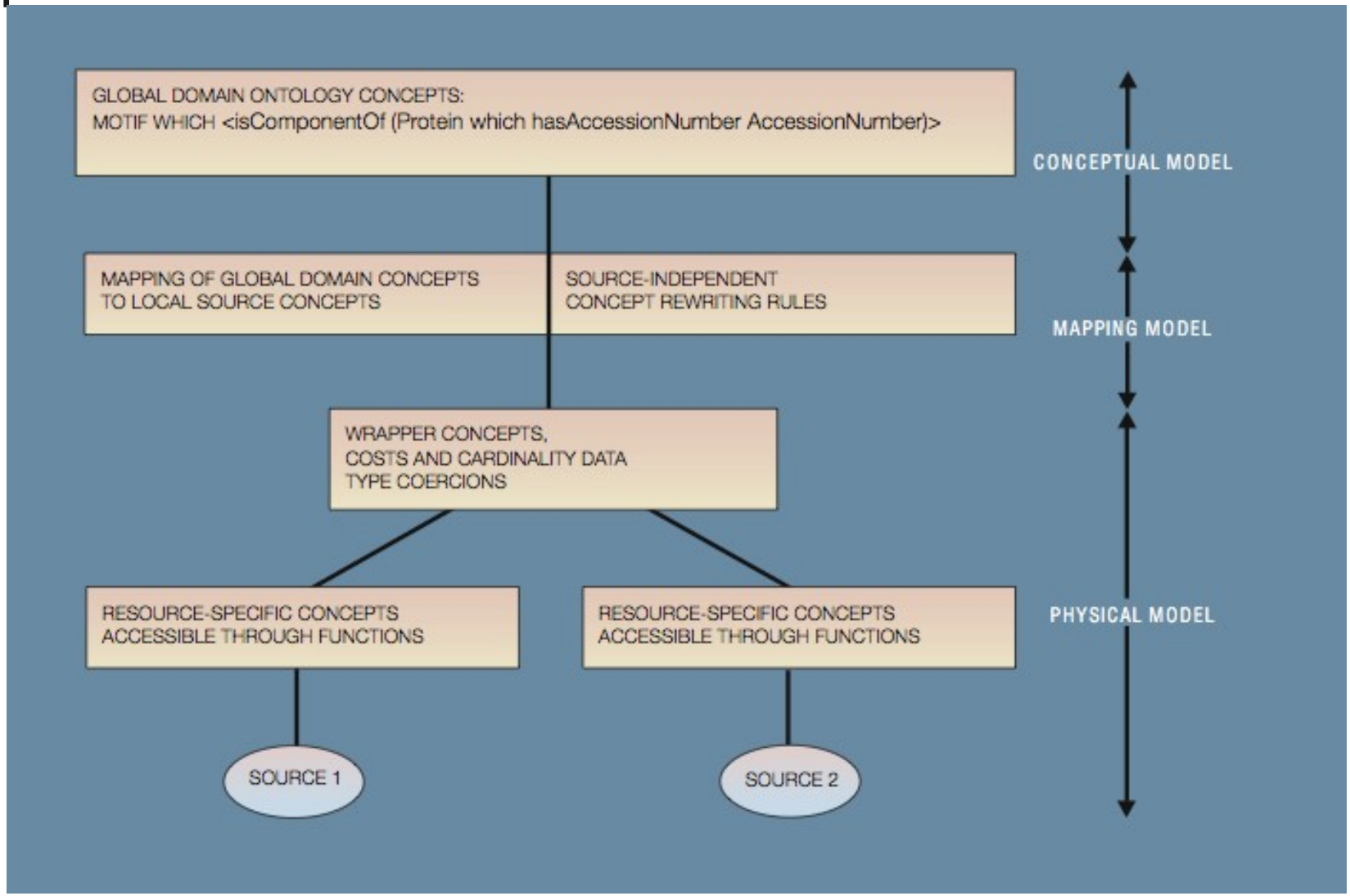
role

role filler

Informal query plan:

- Select proteins with protein name "lard" from SWISS-PROT
- Execute a BLAST sequence alignment process against results
- Check the entries for apoptosis process and antigen function
- Pass the resultant sequences to PROSITE to scan for their motifs

Models for query answering



Mapping model

- Goal: map concepts and roles in the conceptual model to functions that apply to the actual data

```
<concept: Protein which hasFunction
      CatalysisProtein which catalyzes Reaction,
 function: < name: ''get-all-enzyme-entries'',
             arguments: [ ],
             resultType: ''enzyme_record'',
             cardinality: 5000,
             cost: 200,
             source: ''Enzyme''
           >
>
```

CPL function

CPL-specific information

How many mappings??

- It is not necessary to create a mapping for each possible combination of concepts

Ex. no mapping for

(1) "Protein which hasFunction Receptor"

but the concept

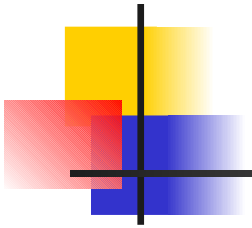
(2) "Protein which hasFunction biologicalFunction"

subsumes (1):

- biologicalFunction is more general than Receptor

and there is a mapping for (2)

- subsumption is used to select the most specialized mapping available
- the ontology guides the selection



Physical model

The original query expressed in CPL:

- Collection Programming Language

P. Buneman, S. B. Davidson, K. Hart, C. Overton, and L. Wong
"A Data Transformation System for Biological Data Sources,"
Proceedings of VLDB, Zurich (September 1995).-

```
\protein3 <- get-sp-entries-by-de("lard"),
\protein2 <- do-blastp-by-sq-in-entry(protein3),
Check-sp-entries-by-kwd("apoptosis",protein2),
check-sp-entries-by-de("antigen",protein2),
Check-sp-entry-for-species("human",protein2),
\motif1 <- do-ps-scan-by-sq-in-entry(protein2)}
```

- CPL functions provide a physical level of mapping
- the ontology coordinates intersource management

TAMBIS - Principal Results

- An ontology for bioinformatics, describing biological concepts and the informatics analyses to which they could be subjected.
- An ontology driven query construction interface that guided users through the ontology, avoiding the construction of biologically meaningless queries.
  - Goble, C.A., Stevens, R., Ng, G., Bechhofer, S., Paton, N.W., Baker, P.G. Peim, M. and Brass, A., *Transparent access to multiple bioinformatics information sources*, *IBM Systems Journal*, Vol 40, No 2, 534-551, 2001.[online]

Spin-out: biological ontologies took off, giving rise to a line of work on expressive bio-ontologies

i.e. the myGrid ontology

Source: N. Paton, DILS 2008 keynote

Principal Problems/Criticisms

- *High development cost*: adding a new source involves extending the ontology, developing extensive metadata and wrappers.
- *Too slow*: data is retrieved at a cost somewhat slower than the slowest integrated sources.
- *Too weird*: description logic ontologies seem out there, and user interface is like nothing anyone has ever seen before.
- *Too restrictive*: manageable implementation leans on simple description logic and query mapping.
- *Not useful*: no biologist ever learned anything they didn't already know by running a query over TAMBIS.

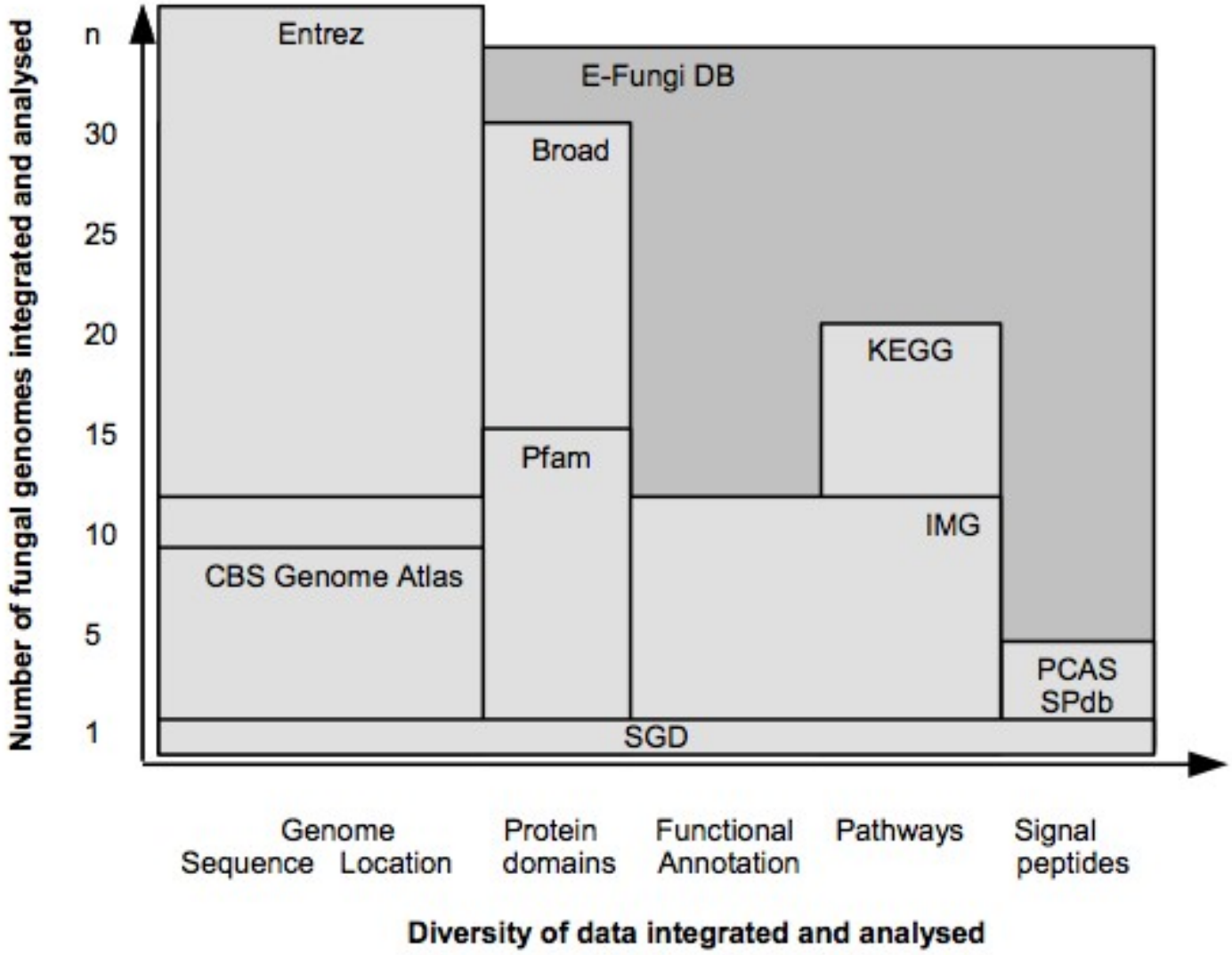
Source: N. Paton, DILS 2008 keynote

Warehousing: the e-Fungi case

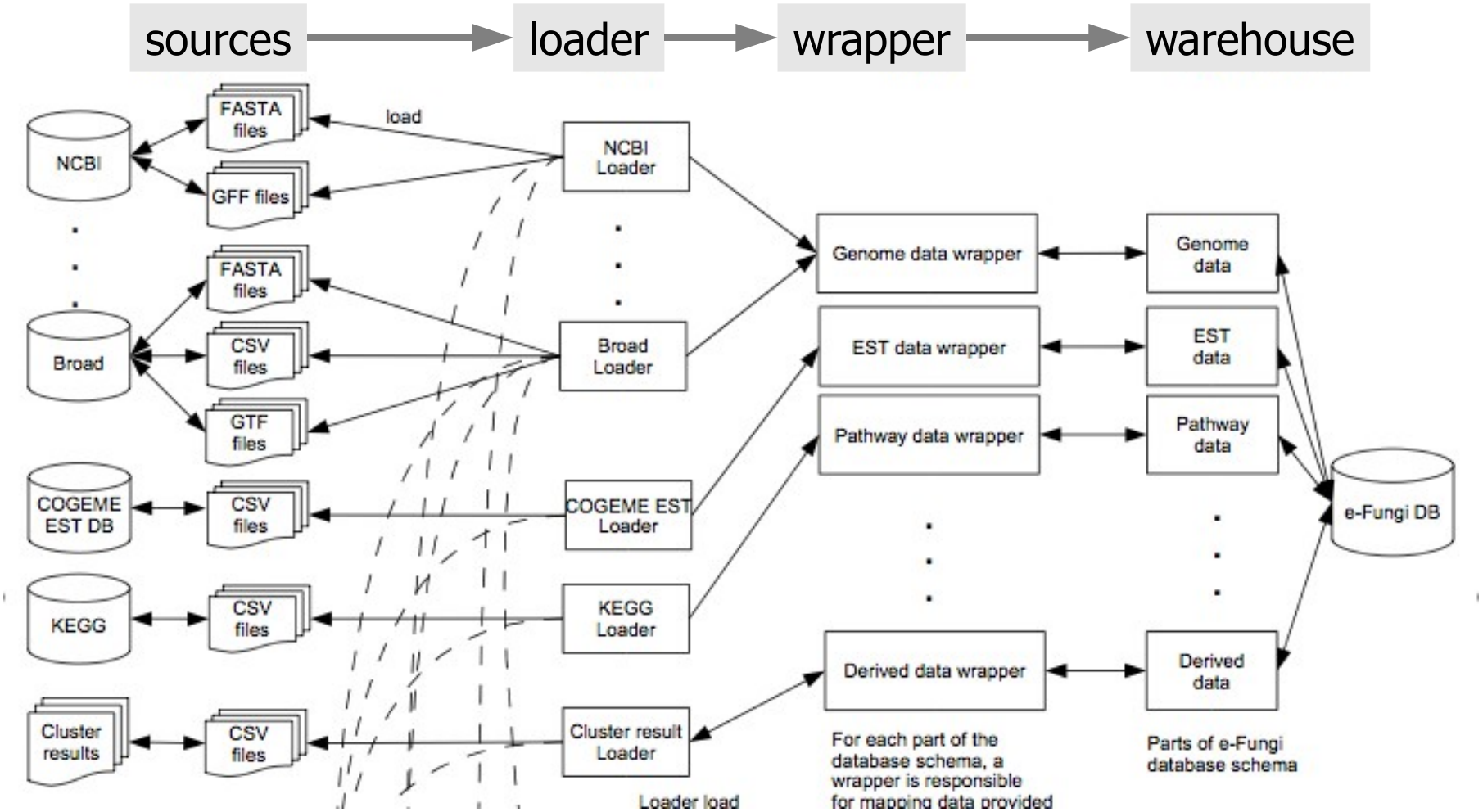
- Comparative functional genomics in the fungi (2004-2008)
- *Biological problem*:
  - Limited understanding of the functional relationships between sequenced fungal genomes
- e-Fungi is a data warehouse that materialises the results of a collection of computationally intensive analyses over sequence data
- supports comparative analyses downstream

Source: N. Paton, DILS 2008 keynote

Sources and their relative contribution



e-Fungi warehousing architecture



Principal Problems/Criticisms

- *High population cost*: populating a new version of the warehouse involves weeks of analysis on a campus grid.
- *Labour intensive*: ever changing underlying genomes and formats make for cumbersome manual cleaning.
- *Unstable results*: ever changing genome analyses and genomes available mean that derived data differs between releases.
- *Low flexibility*: the data in the warehouse is never exactly what is required, and incremental loading is problematic.
- *Deferred gratification*: the queries over the warehouse provide results of complex analyses that are tricky to interpret.

Source: N. Paton, DILS 2008 keynote

Lessons learnt

- Hard to construct a warehouse for more than one task
- The bottleneck quickly became result *interpretation*, not *population*.
  - Many analyses generate less than conclusive outcomes, leading to much detailed scrutiny and literature trawling.

Source: N. Paton, DILS 2008 keynote

The SNPit integrated DB

Goal:
to build a flexible data infrastructure to support current biology research involving gene polymorphism (SNP)

- Add value to existing public SNP databases
- Support multiple experimental added-value SNP analysis packages

Core application:

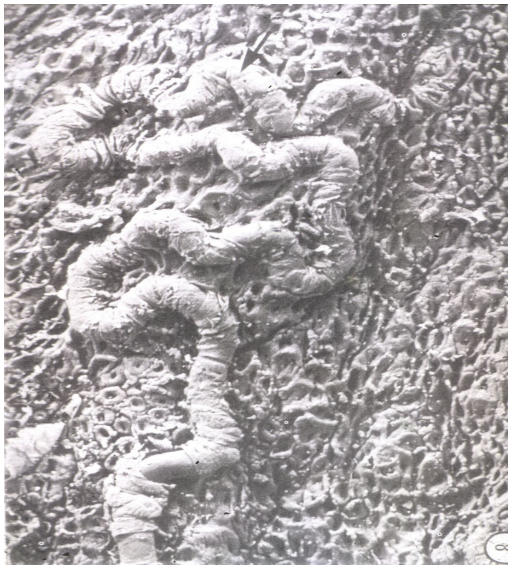
- Analysis of genetic factors in observed phenotypes
  - resistance / susceptibility to a certain disease
  - life span, weight, …

Diseases and phenotypes

Trichuris trichiura

- Same life cycle
- Natural parasite of mice

Trichuris muris



Genetic Component to Susceptibility to *Trichuris trichiura:* Evidence from Two Asian Populations

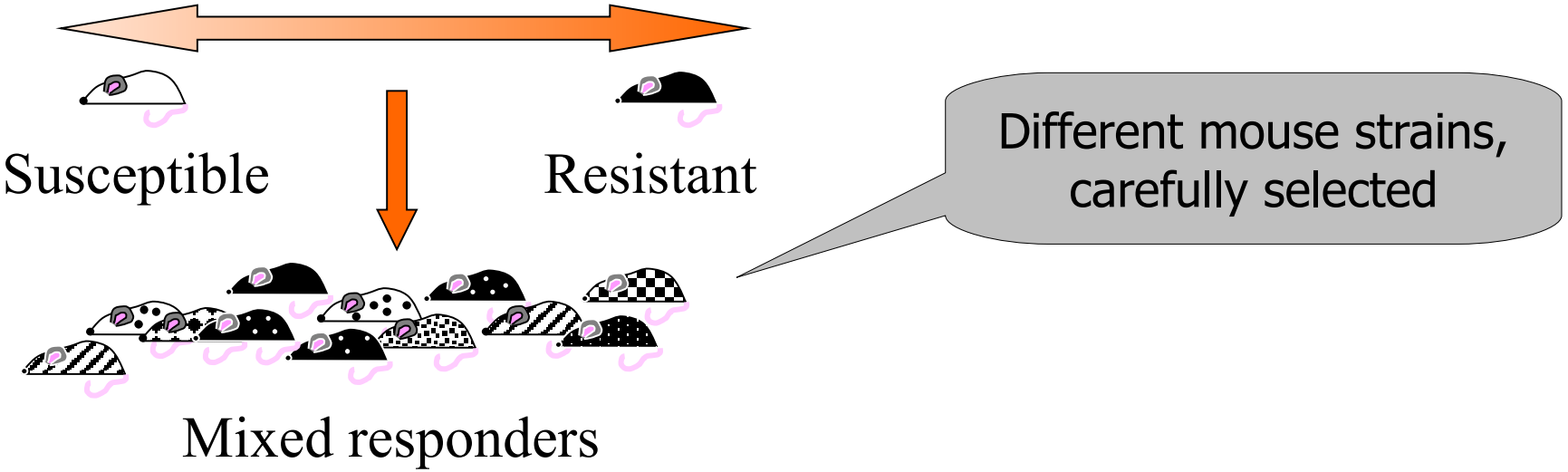
S. Williams-Blangero et al. - Genetic Epidemiol. 2002 22 (5):254

''……28% of the variation in Trichuris trichiura loads was attributable to genetic factors in both populations."

Genetic causes of known phenotypes

Association mapping

- Experimental method to correlate quantitative phenotype with genotype
- Associates a region on the chromosome to a specific phenotype through complex in-breeding schemes

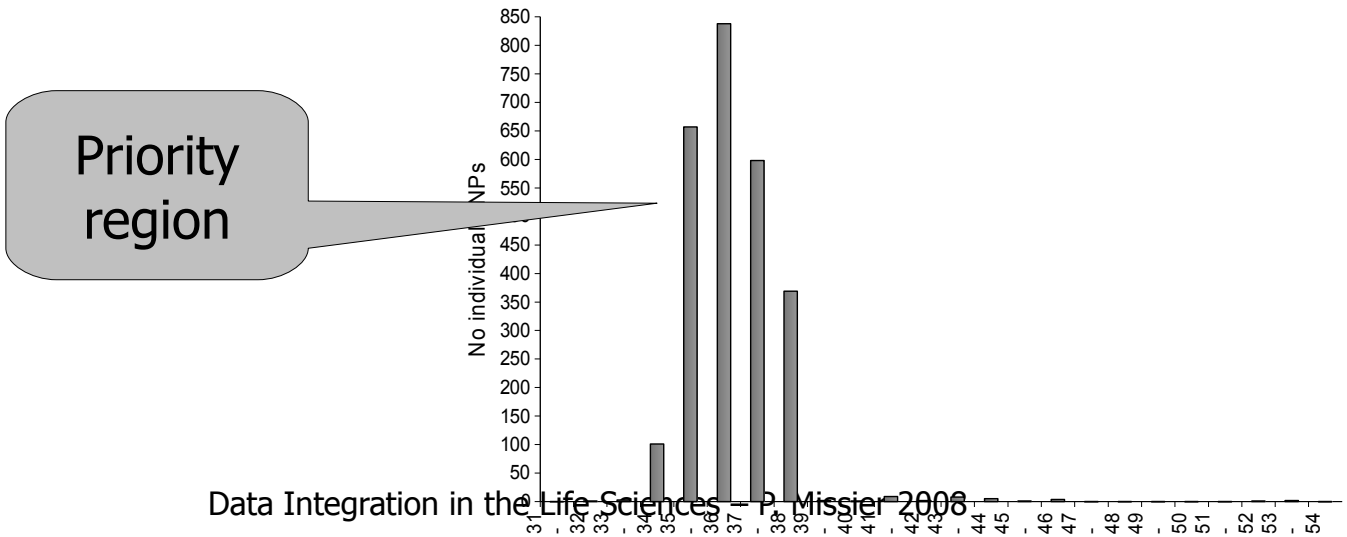


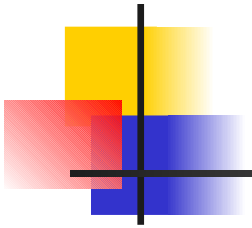
Role of SNPs in association mapping

SNP: Single Nucleotide Polymorphism

- single-base change in a strain relative to a reference mouse strain

- Bioinformatics approach
  - Identify areas of greatest difference between resistant / susceptible strains
  - Prioritize candidate gene search using the density of highly differentiated gene regions

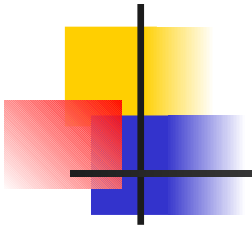




The SNPit project

- A "lightweight" SNP database designed to support association mapping studies

- SNPit is a *secondary* DB
  - It draws from three primary sources:
  - **Ensembl (EBI), dbSNP (NCBI), Perlegen**



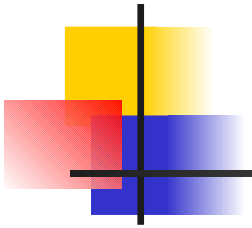
SNPit application challenges

Supports interactive exploratory analysis over large genomic regions

- Over 50Mb –200K SNPs per source / *per query*

Typical flow: (see SNPit web interface)

- Region selection (or gene set)
- Source selection (multiple)
- Strain group selection – per-session basis
- Compute score for each SNP in the region – on the fly
- (filter by gene polymorphism)
- Rank SNPs by score, gene polymorphism – in-memory sorting
- Plot density of high-score SNPs over the selected region
- Change parameters and repeat…



Why multiple SNP DBs

- SNP databases differ
  - Partially overlap in structure and content
  - Different update policy and frequency
- Biologists like to choose their sources
  - Based on experience, prior usage, confidence
- The SNPit application offers an explicit choice
- It exploits complementary features and content of the DBs

Data integration architecture

SNPit Web app

SNPit Web Service

Core Data processing

Score 1

Score 2

…

SNPit DB

Ensembl

dbSNP

Perlegen

rsId

ssId

load

load

load

Ensembl SNP

dbSNP

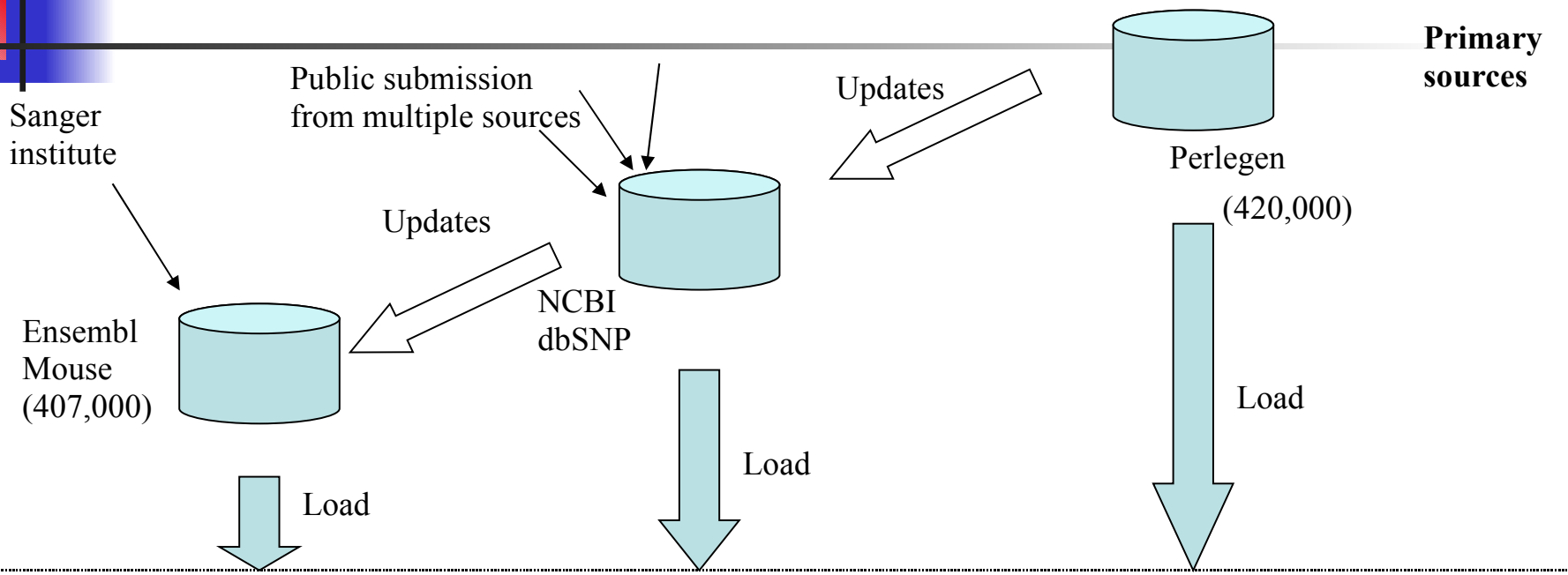
Perlegen

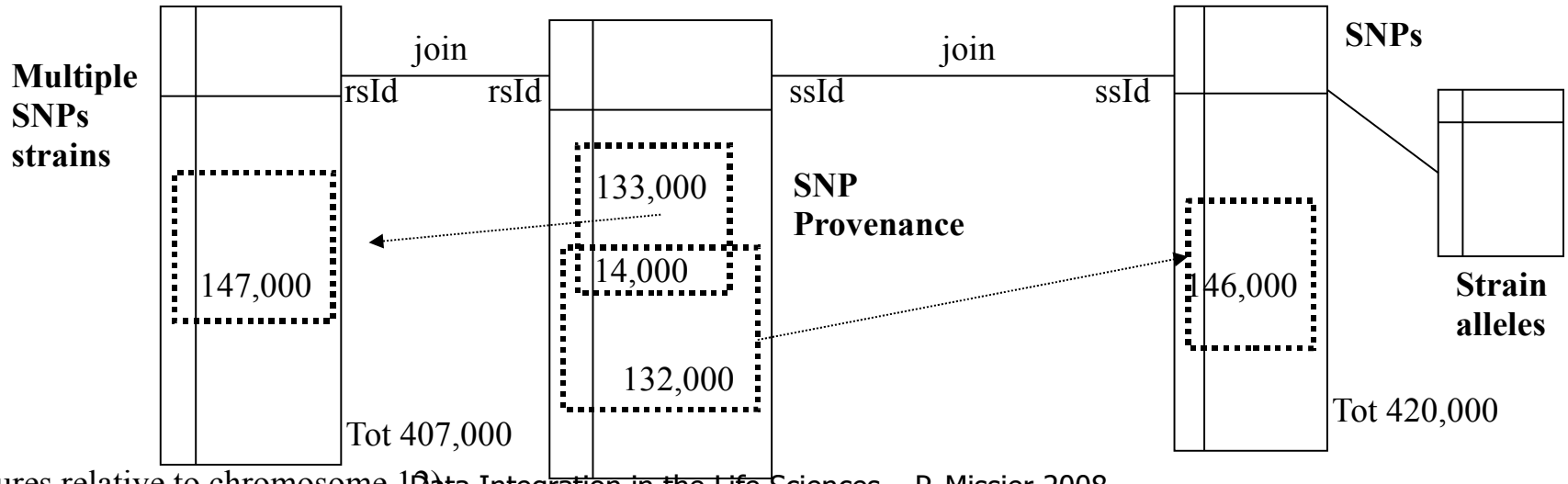
Interdependent materialized views

Periodic updates

- no single global schema
- queries against the views
- some queries can be directed to more than one DB
- End-user web app
- Web Service accessible as a workflow processor (Taverna)

SNP DB dependencies



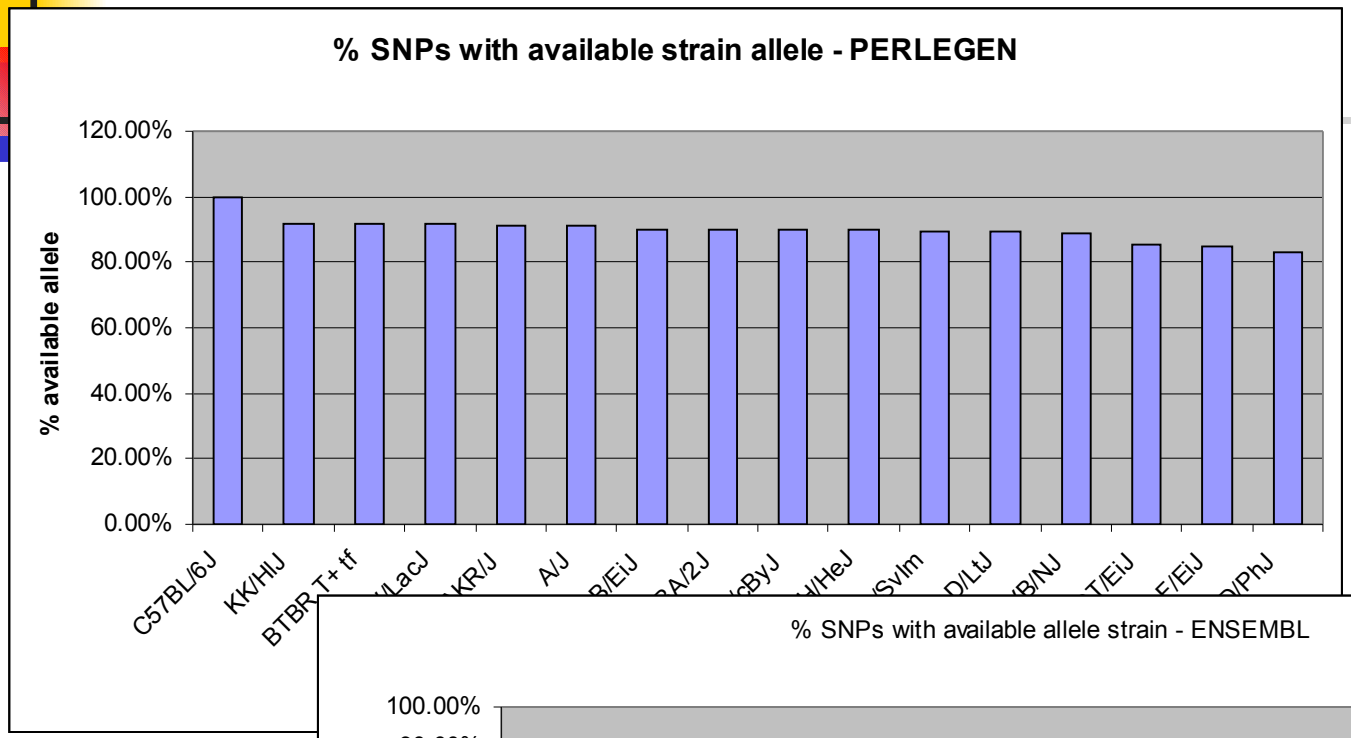


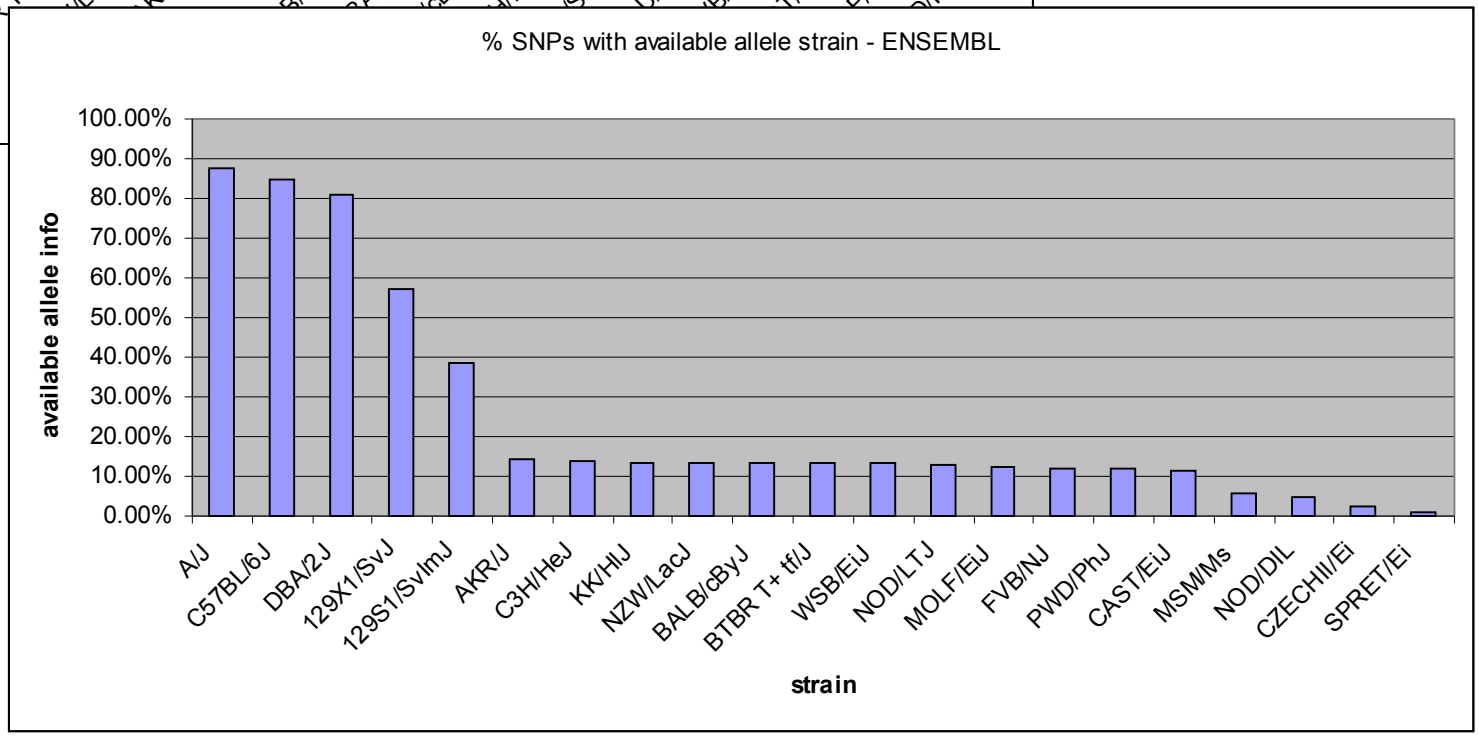
(all figures relative to chromosome 1)

Qualitative differences

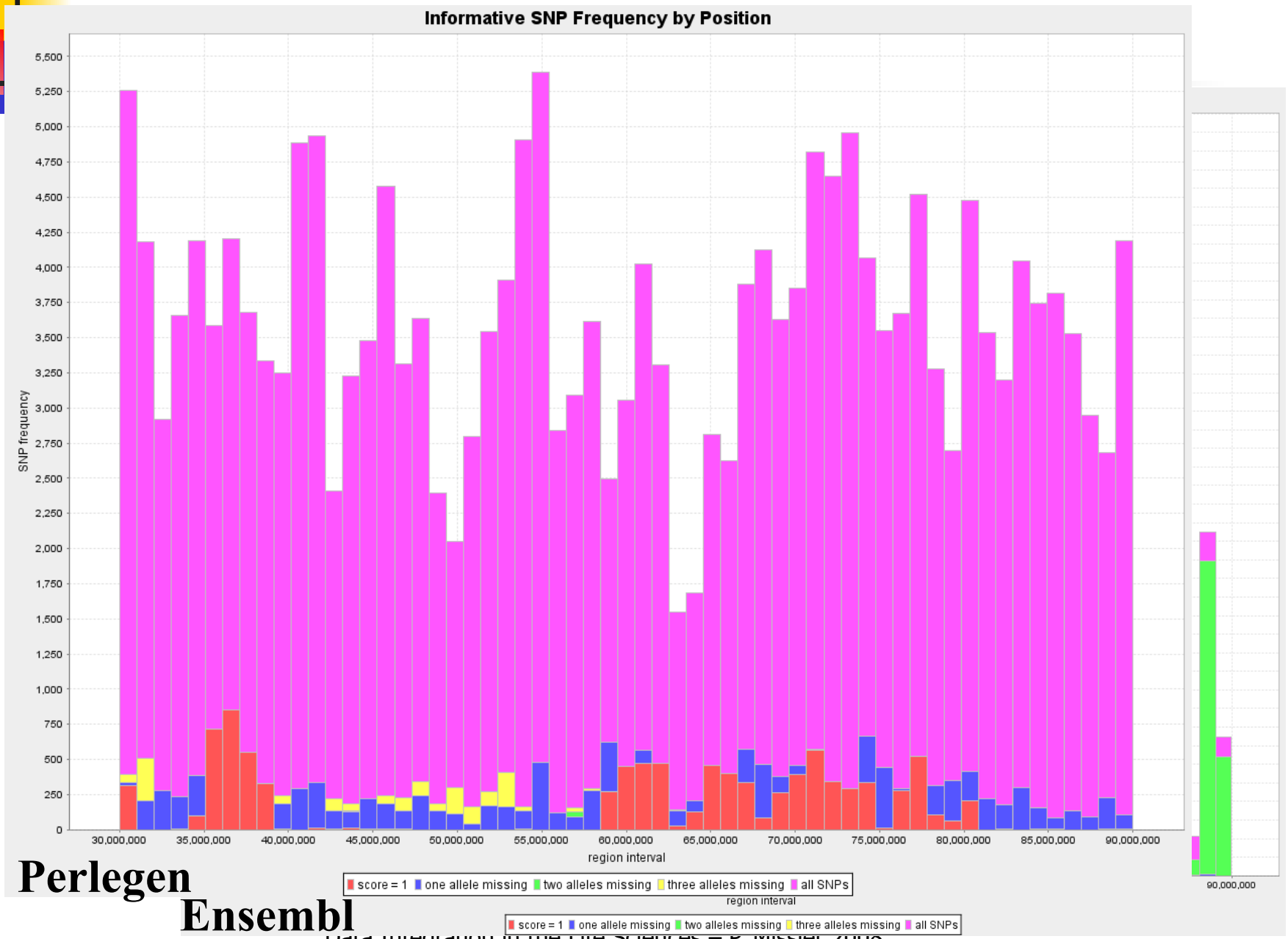
	Strengths	Weaknesses	Strain info
Ensembl	• Curated SNPs • Evolving • SNP location info (exonic, intronic) • Multiple reputable sources •Controlled submission	Low timeliness	About 60 strains Not very complete
dbSNP	• Submitter info • Update history (provenance)	•Multiple sources •Low quality control on public submission •Timely	Not used
Perlegen	• Good quality control • High reputation	•No SNP location •Not evolving	16 strains (ref + 15) Fairly complete

Missing strains – chr 17



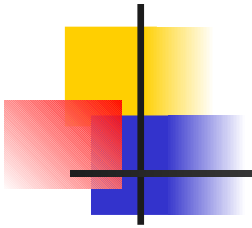


Effect of source selection



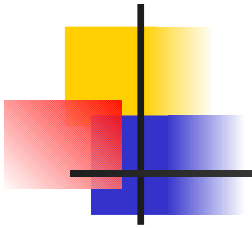
Perlegen

Ensembl



SNPit – lessons learnt

- Useful, but open service architecture idea needs to be explored further
  - single-purpose syndrome
- Periodic materialization needed
- Useful for on-the-fly source selection
  - but, more sources needed...
- hard to justify maintenance resources
  - competes with MGI, Mouse Phenome Database...



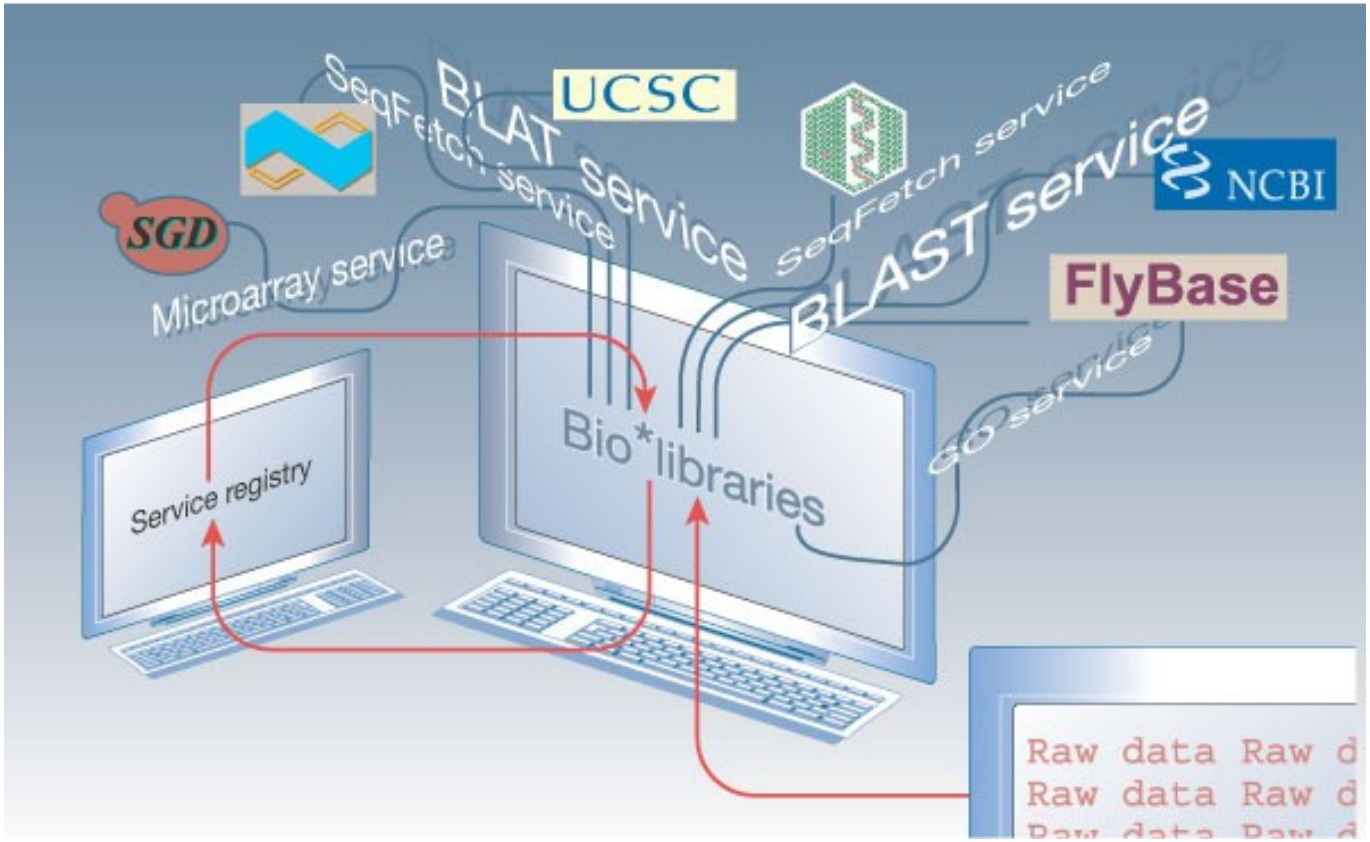
The "bioinformatics nation" vision

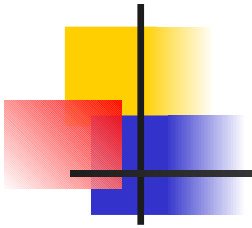
Commentary by L. Stein on Nature 417(9 May 2002)

- The curse of heterogeneity on a large scale
  - A plethora of web sites, services, and data formats
  - "creative chaos" failing to mature into coordinated progress
- The promise of Bio-* services
  - integration through standardized services

Interoperation through services

- Web services + service registry
  - the ultimate solution to interoperability (circa 2002)





Data integration through service integration

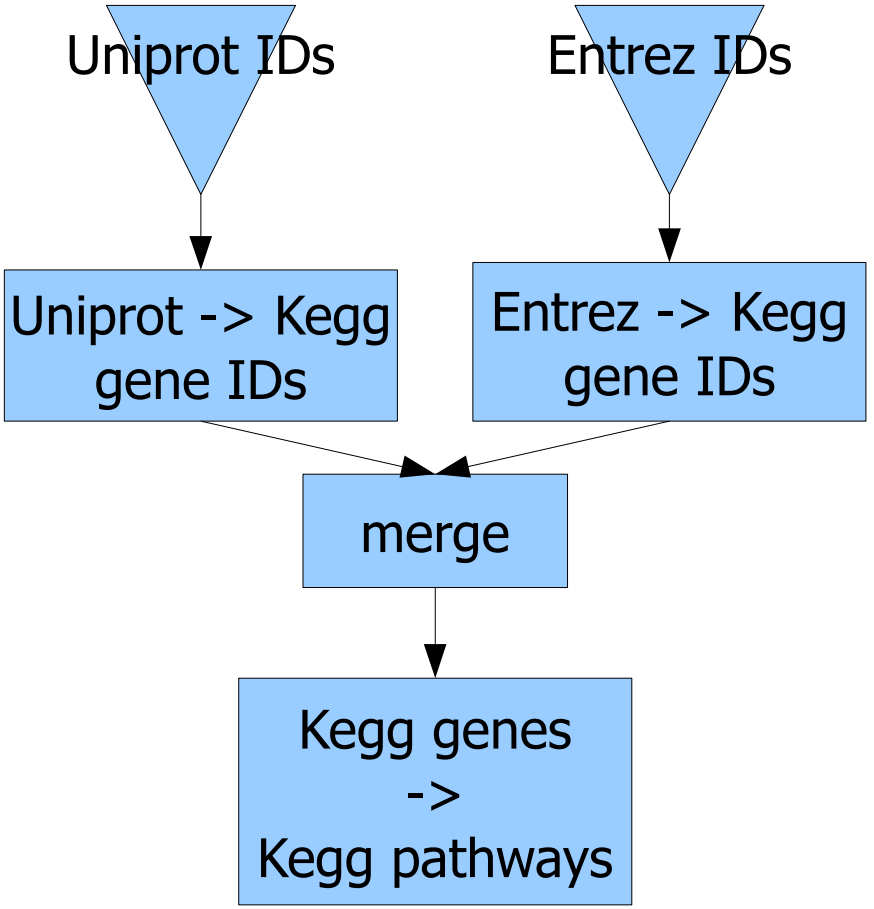
Scientific workflows are effectively used to perform ad hoc integration of data sources

Case study: myGrid

- a middleware infrastructure for e-science research
- consists of the Taverna workflow authoring and enactment infrastructure, plus associated tools for discovering and annotating services and workflows

An example Taverna workflow

- The workflow maps mouse genes from Uniprot and Entrez to Kegg pathways



• This is only a conceptual view of the process

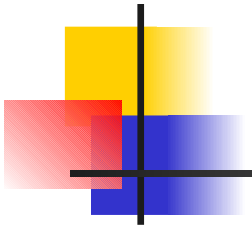
• The actual process involves many more steps

• see workflow loaded in the Taverna workbench

Principal Problems/Criticisms

- *No data integration*: underlying services go their own way on key issues like identity and formats; resolving the mess consumes much of the workflow-writing effort.
- *Instability*: underlying services come, go and change rapidly, making workflows fragile.
- *Knowledge management is expensive*: Taverna is cheap-and-cheerful; providing the associated ontologies and annotations tends to be expensive.
- *Findings are generally superficial*: Taverna is better suited to light-weight hypothesis checking or source linking than to large scale, systematic studies.

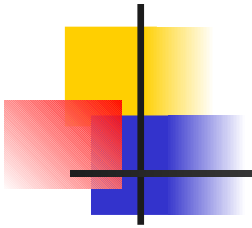
Source: N. Paton, DILS 2008 keynote



With hindsight…

- Workflows glamorise an uncomfortable truth – data integration in the life sciences is complicated and *ad hoc*.
- It is easier to talk the talk with semantic web services, than to walk the walk due to the high cost of capturing and maintaining good quality metadata.
- The first few iterations on provenance and discovery services yield publications but not users.

Source: N. Paton, DILS 2008 keynote



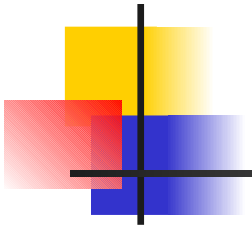
Role of standardization

Data submission content and format standards

- MIAME (The MGED society)
  - description of microarray experiments
- MIAPE (Proteomics experiments) – see next slide

Open Ontologies

- rationale: standardize on domain terminology
  - The Gene Ontology
  - Open Biomedical Ontologies
  - MGED
  - …



MIAPE

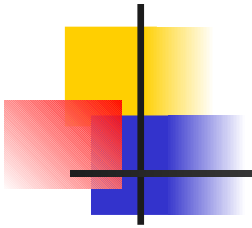
Minimum information about a proteomics experiment [paper]

Goals:

- ensure that paper submissions are matched by experimental data submission to a database
- standardize on appropriate metadata for experiment description, and its structure
  - minimize heterogeneity
  - ensure uniform abstraction level

Requirements:

- facilitate search
- encourage adoption (i.e., by providing tools)

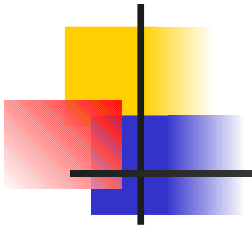


AutoMed

- see separate slides set (courtesy of Lucas Zamboulas): HBMI07

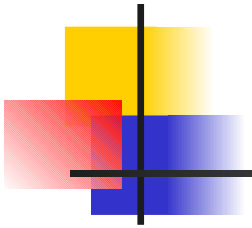
Selected further reading

- The DILS workshop series
  - Springer proceedings, DBLP, ...
- Tambis
  - Baker, P.G., Goble G.A., Bechhofer, S., Paton, N.W., Stevens, R. and Brass, A., An Ontology for Bioinformatics Applications, *Bioinformatics*, Vol 15, No 6, 510-520, 1999.
  - Goble, C.A., Stevens, R., Ng, G., Bechhofer, S., Paton, N.W., Baker, P.G. Peim, M. and Brass, A., *Transparent access to multiple bioinformatics information sources*, *IBM Systems Journal*, Vol 40, No 2, 534-551, 2001.[online]
- SNPit
  - P. Missier, S. Embury, C. Hedeler, M. Greenwood, J. Pennock, A. Brass, Accellerating Disease Gene Identification through Integrated SNP Data Analysis, Data Integration in the Life Sciences (DILS '07), July 2007, Philadelphia, USA



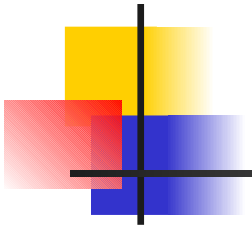
Taverna / myGrid

- T Oinn et al, Taverna: lessons in creating a workflow environment for the life sciences in Concurrency and Computation-Practice & Experience, Volume 18, Issue 10, 2006
- # Duncan Hull, Katy Wolstencroft, Robert Stevens, Carole Goble, Matthew Pocock, Peter Li, Tom Oinn Taverna: a tool for building and running workflows of services in Nucleic Acids Research June 2006
- # Yolanda Gil, Ewa Deelman, Mark Ellisman, Thomas Fahringer, Geoffrey Fox, Dennis Gannon, Carole Goble, Miron Livny, Luc Moreau, Jim Myers Examining the Challenges of Scientific Workflows, IEEE Computer, vol.40,no.12,pp. 24-32,December,2007
- K. Wolstencroft, P. Alper, D. Hull, C. Wroe, P.W. Lord, R.D. Stevens, C.A Goble, The myGrid Ontology: Bioinformatics Service Discovery, International Journal of Bioinformatics Research and Applications, Special Issue on Ontologies for Bioinformatics, Vol. 3, No. 3. (2007), pp. 303-325



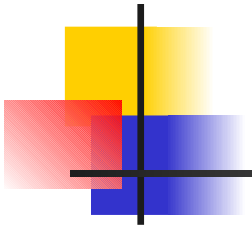
Standards for proteomics

- MIAPE:
  - Taylor C.F., Paton N.W. et al The minimum information about a proteomics experiment (MIAPE) Nature Biotech, August;25(8), 887 - 893, 2007
  - HUPO-PSY (Proteomics Standards Initiative)



e-fungi

- A comparative analyses of the evolving behaviour of genomes in the fungi:
  - Cornell M.J., Alam, I., Soanes, D.N., Wong, H.M., Hedeler, C., Paton, N.W., Rattray, M., Hubbard, S.J., Talbot, N.J., Oliver, S.G., Comparative genome analysis across a kingdom of eukaryotic organisms: Specialization and diversification in the Fungi, *Genome Research*, 17, 1809-1822, 2007.
- An architecture that supports the analyses:
  - Hedeler, C., Wong, H.M., Cornell, M.J., Alam, I., Soanes, D.M., Rattray, M., Hubbard, S.J., Talbot, N.J., Oliver, S.G., Paton, N.W., e-Fungi: a data resource for comparative analysis of fungal genomes, *BMC Genomics*, 8:426, 2007.



AutoMed / Ispider

- Lucas Zamboulis et al, Data Access and Integration in the ISPIDER Proteomics Grid in Data Integration in the Life Sciences, Lecture Notes in Computer Science Volume 4075/2006 , Springer Berlin / Heidelberg
- Lucas Zamboulis, Nigel J. Martin, Alexandra Poulovassilis: Bioinformatics Service Reconciliation by Heterogeneous Schema Transformation, in Data Integration in the Life Sciences, Lecture Notes in Computer Science Volume 4544/2007 , Springer Berlin / Heidelberg
- http://www.dcs.bbk.ac.uk/~lucas/talks/HBMI07.pps