

# Perspectives of the application of semantic technologies in Bioinformatics

Paolo Romano

([paolo.romano@istge.it](mailto:paolo.romano@istge.it), skype:p.romano)

Bioinformatics, National Cancer Research Institute, Genoa

**"The advantages of the Semantic Web lie  
in its ability to present and  
provide access to complex knowledge  
in a  
standardized form  
making  
interoperability between distributed databases and  
middleware achievable."**

*Semantic Web: revolutionizing knowledge discovery in the Life Sciences,  
C. Baker and K.-H. Cheung (eds), Springer, 2007*

# Outline

- Characteristics of biology information systems
- Data integration issues in biology
- Past, current and maybe future data integration tools:
  - SRS !
  - Workflow management systems !!
  - Semantic Web ?
- Perspectives of the adoption of SW
- The HCLS IG of W3C

# Huge data size

Biomedical research produces an increasing quantity of **new data and new data types**

Genomics is producing an immense quantity of data

Emerging domains, like mutation and variation analysis, polymorphisms, metabolism, as well as new high-throughput technologies, e.g., microarrays, will also contribute with huge amounts of data

Analysis software must interoperate with databases

Databases as input for software

Results as new data to store and analyze

# Heterogeneity of databanks

**A few dbs are managed in a homogenous way (nucleotide sequences at EBI, NCBI, DDBJ)**

**Majority of systems have own data structure**

- **Secondary databases** are of the highest quality (good and extended annotation, quality control)
- Many databases are highly **specialized**, e.g. by gene, organism, disease, mutation
- Many databases are created by **small groups** or even by single researchers

**Databanks are distributed:**

- Different DBMS, data structures, query methods
- Same information, different syntax and semantics

# Goals of the integration and automation

In this context, **data integration** and **process automation** are needed to:

- **Automatically** carry out analyses and/or searches involving **more** databases and software
- **Effectively** perform analyses involving **large data sets**
- Achieve a better and **wider** view of available information
- Carry out a real **data mining** and **discover** new information
- The ultimate goal being to **understand** biological phenomena

This can be done by computers, but...

# What is needed

## Integration and automation need stability

- o Standardization.....
- o Good domain knowledge
- o Clearly defined data and scope
- o Clearly identified goals

## Integration and automation fear changes

- o Heterogeneity of data and systems
- o Uncertain domain knowledge
- o Fast evolution of data
- o Highly specialized data
- o Evolving needs and goals

# What about biological information

## In biology:

- A pre-analysis and reorganization of information is very difficult, because knowledge and related data change very quickly
- Complexity of information makes it difficult to design data models which are valid for different domains and over time
- Goals and needs of researchers evolve very quickly according to new theories and discoveries

Integration must therefore be carried out by using flexible systems that are easy to adapt and to extend

# Integration methods

From syntactical to semantic methods:

- Explicit links (cross references)
- Implicit links (use of common terms)
- Common terminology (shared vocabularies)
- Common models (shared data models and schemas)
- Common semantics (ontologies)

# Explicit links

## Between records of distinct databases:

- o Use of a direct link, unique, non reciprocal
- o Use unique db records' ID
- o Links are expressed in (de facto) standard formats

## Has limits:

- o Must be predefined
- o Manual, hand coded annotation
- o Semantics of the link is implicit, not specified

# Shared terms and vocabularies

Between records of distinct databases, by text search:

- o Implicit link, non unique, reciprocal
- o Automatically defined, no prior human intervention
- o Based on terms from controlled vocabularies

Has limits:

- o Sharing of vocabularies is needed
- o The context where terms appear must be specified
- o Text mining may be needed (distinguish terms that also are common words, copying with synonyms and words with many meanings)
- o Semantics of the link is not specified

# Shared data models

Between records of distinct databases, by query:

- o Semantics and context clearly defined
- o Automatically defined, no prior human intervention
- o Search through a standard abstract interface
- o Results returned in standard formats
- o Adoption of common data models and schema

Has limits:

- o Sharing of data models is requested
- o Setting up links requires prior knowledge
- o Semantics is embedded in the software
- o Requires expertise and computer skills

# Ontologies

An ontology is a formal specification of knowledge in a defined domain, usually limited.

It consists of:

- a series of concepts,
- a controlled vocabulary to express concepts with,
- typed relationships among them.

An ontology can be used to:

- add semantic contents to a database,
- improve access to data,
- make data integration easier.

It allows researchers to understand the meaning of data by specifying related concepts and software to manage data in a coherent way.

# Ontologies

## Basic ontologies

- Gene Ontology (GO)
- MGED Ontology (MO)

## Ontologies derived from basic ones

- Cell Ontology (CO)
- Ontology for Biomedical Investigations (OBI)

## Upper-level ontologies (key concepts, reusable)

- Foundational Model of Anatomy (FMA)
- Galen Bio Upper Ontology

## Ontologies for future developments

- Phenotype, Attribute and Trait Ontology (PATO)
- Clinical Bioinformatics Ontology (CBO)

# Ontologies

Between records of distinct databases, by metadata annotation:

- o Semantics and context clearly defined
- o Automatically defined
- o No human intervention for setting the link
- o Manual annotation of the record, not of the link

Has limits:

- o Sharing of ontologies is requested
- o Existing ontologies should be interlinked
- o New ontologies must be defined
- o Still requires expertise and computer skills

# Past, current and future technologies

SRS (Sequence Retrieval System)

Workflow management systems

Semantic Web technologies and tools

# SRS - Sequence Retrieval System

SRS is a well known, effective, easy to use, widely adopted system for local data integration of heterogeneous databases.

Its approach is based on the following points:

- o databases are available locally
- o special syntax rules are defined to enable indexing and data extraction
- o links to both internal and external databases available
- o linked data can be displayed together
- o transparent integration with analysis tools

Its limits derive from its features:

- o databases must be downloaded and locally managed
- o syntax rules must be refined when data structure is changed
- o data visualization is limited
- o analysis tools must be managed, updated
- o access to new services is difficult and must be programmed

# SRS: Links

SRS links can be defined either explicitly or implicitly

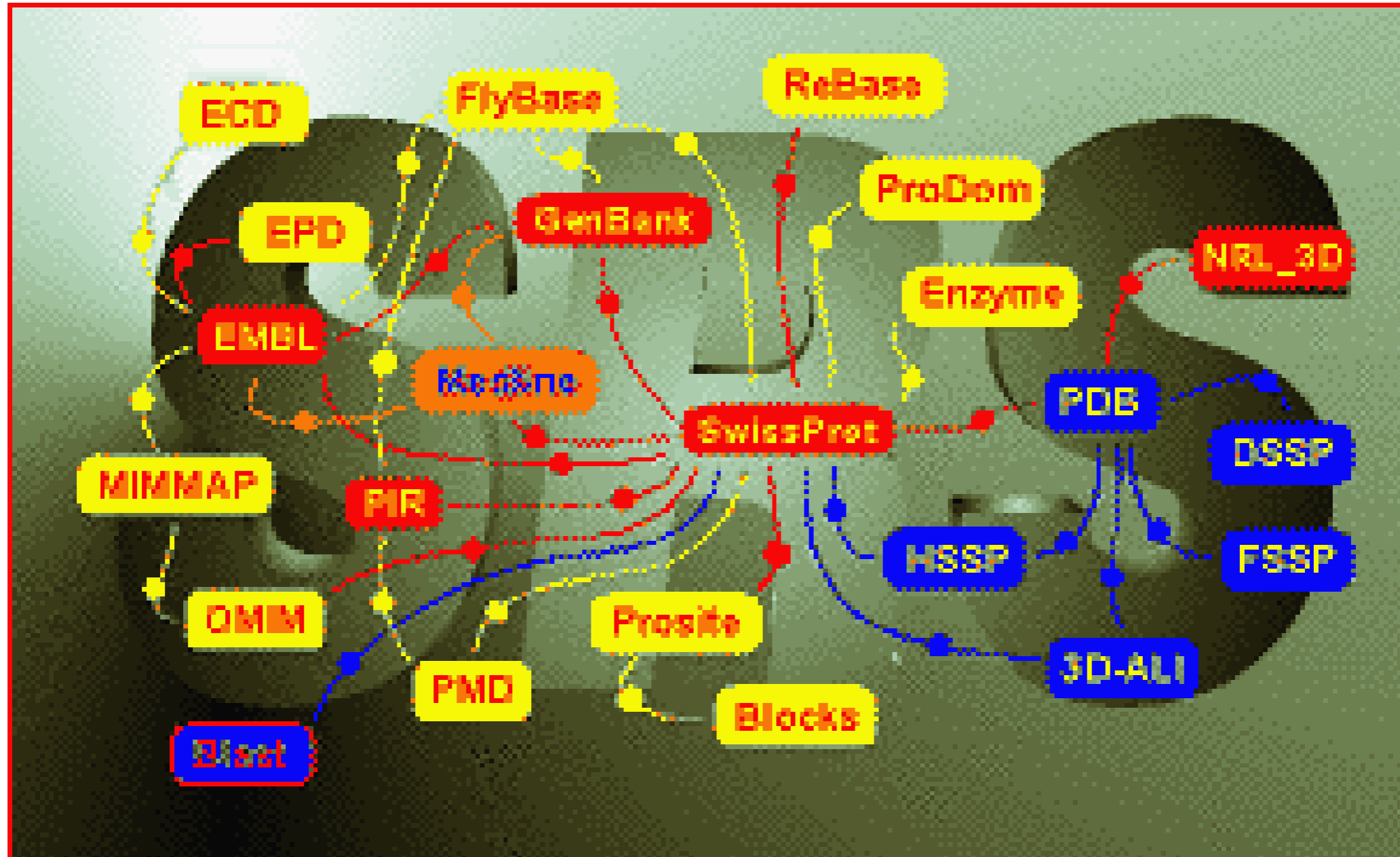
Explicit links: db IDs are inserted in databases and SRS can be instructed to recognize them

- Other\_collection\_numbers CCUG 34964; NCIB 12128
- Literature DSM ref.no. 72; DSM ref.no. 1300
- EMBL: X52289

Implicit links: common terms in specified fields

- TargetGene: APOE
- Constructed\_from pMB1, pSC101 and Tn3
- Name *Gluconacetobacter xylinus* subsp. *xylinus*, (Brown 1886) Yamada, Hoshino and Ishikawa 1998 VL
- Literature Nucleic Acids Res 1990;18:4967 [PMID: 2395673]

# SRS: map of links



# SRS: Link operators

SRS link operators allow to exploit SRS links within queries

- Retrieve all records from 'swissprot' having a link to 'EMBL'  
swissprot < EMBL
- Retrieve all records from 'EMBL' having a link to 'swissprot'  
EMBL < swissprot
- Retrieve all records in 'swissprot' having a link to records in 'EMBL' that satisfy specified conditions  
swissprot < [EMBL-id: X52289]
- Retrieve all records from 'EMBL' that satisfy a condition and have a link to records in 'medline' that also satisfy a condition  
[EMBL-organism:human] < [medline-pmid:3137981]

# SRS: link operator in action - subsets

EMBL-EBI EB-eye Search All Databases Enter Text Here **Go** Reset **Give us feedback**

Databases Tools EBI Groups Training Industry About Us Help Site Index **HELP** Job Status

Quick Search Library Page Query Form Tools Results Projects Views Databanks

**Reset**

Query: ( emb1 < [uniprot-acc:P15711])  
found 2 entries

EMBL	Primary Accession (Links to SVA)	Accession List	Description	Sequence Length
<input type="checkbox"/> <a href="#">EMBL:M29954</a>	<a href="#">M29954</a>	<a href="#">M29954</a>	T.parva 104 kDa microneme-rhoptry antigen gene, complete cds.	4584
<input type="checkbox"/> <a href="#">EMBL:AAGK01000004</a>	<a href="#">AAGK01000004</a>	<a href="#">AAGK01000004</a> <a href="#">AAGK01000000</a>	Theileria parva strain Muguga chromosome 4 ctg_529, whole genome shotgun sequence.	1835834

**Apply Options to:**

selected results only

unselected results only

**Result Options**

Launch analysis tool:  
NCBI BLASTN **Launch**

Show tools relevant to these results: **Tools**

Link to related information: **Link**

Save results: **Save**

**Display Options**

View results using:  
EMBLSeqSimpleView

Show 30 results per page

Printer friendly view

**Apply Display Options**

# SRS: link operator in action - views

EMBL-EBI EB-eye Search All Databases Enter Text Here [Go](#) [Reset](#) [Give us feedback](#)

Databases Tools EBI Groups Training Industry About Us Help Site Index [HELP](#)

Quick Search Library Page Query Form Tools Results Projects Views Databanks [Job Status](#)

[Reset](#) `( embirelease < [uniprot-acc:P15711])`

Query found 1 entries

EMBL (Release)	ID	Organism	UniProtKB	EntryName	GeneName	MolWeight
<input type="checkbox"/> <a href="#">EMBL (Release):M29954</a>	M29954	Theileria parva	<a href="#">UniProtKB:104K_THEPA</a>	104K_THEPA		103626

**Apply Options to:**

selected results only  
 unselected results only

**Result Options**

Launch analysis tool:  
 NCBI BLASTN [Launch](#)

Show tools relevant to these results: [Tools](#)

Link to related information: [Link](#)

Save results: [Save](#)

**Display Options**

View results using:  
 test

Show 30 results per page

Printer friendly view

[Apply Display Options](#)

# Workflow managements systems

An integration methodology based on data standardization:

- XML schemas: creation of the models of the information
- XML based languages: data representation and storage
- Web Services: data exchange, software interoperation
- Ontology annotation of data types and analysis
- Computerized workflows: definition and execution of analysis processes
- Portals: accessibility and usability of workflows by all

Models lack, while XML languages, Web Services and WMS are more and more available.

# Workflow

"A computerized facilitation or automation of a business process, in whole or part". (Workflow Management Coalition)

Its main goal is the implementation of data analysis processes in standardized environment

Its main advantages relate to:

- **effectiveness:** being an automatic procedure, it frees bio-scientists from repetitive interactions with the web and it supports good practice,
- **reproducibility:** analysis can be replicated over time,
- **reusability:** intermediate results can be reused,
- **traceability:** the workflow is carried out in a transparent analysis environment where data provenance can be checked and/or controlled.

# Status of software components

## Models: a few

BioSQL (bio\* initiative)

Helmholtz Open BioInformatics Technology (HOBIT) XML schemas

MAGE Object Model (MAGE-OM)

## XML dialects: more and more

Sequences (BSML, Agave), Proteins (SPML), NCBI outputs (BlastXML)

Microarray (MAGE-ML)

Systems Biology Markup Language (SBML), Cell System Markup Language (CSML)

Polymorphism Markup Language (PML), Biological Variation Markup Language (BVML)

## Web Services: a lot

EMBOSS, XEMBL, Interpro (EBI)

eUtils (NCBI)

GeneCruiser, Biosphere (microarray)

CABRI (biological resources), TP53 mutations (gene mutations)

bioMOBY (directory), Soaplab (tools), Distributed Annotation System (DAS)

# Status of software components

## **Workflow management systems: some**

- **Software libraries:** Add-on to development tools, need programming efforts.
- **Standalone systems:** Normally implemented on personal computers for accessing distributed services.
- **Client/server systems:** May implement some functions of the WMS at the client side. Services can be local to the server or distributed.  
Maybe Grid enabled

## **Languages for Workflows: too many**

- **Proprietary:** Developed for a WMS, optimized to its goals, Commercial software often not standardized
- **Standard:** Standard can be very general, not really goal-oriented and specific. Different organizations, different standards
- **Different availability schemes:** Commercial / Public domain or use / Open source

# Status of software components

## **Ontologies for bioinformatics data types and tasks: some**

- **Specify both data types** that can be exchanged between Web Services and **elaboration tasks** performed by services
- In an automated analysis process, **all data exchange** is carried out through **Web Services** that should therefore use a shared ontology of bioinformatics data types and tasks.

## **Portals for execution of workflows: a few**

- Simplify execution of **predefined workflows** by non skilled users
- **Users are registered**, authenticated and own a personal space
- Results of **previous executions** can be saved and re-analyzed
  
- Workflows are kept **up-to-date** with new and revised Web Services
- Procedures can be **standardized** in a company/institute

# WfMS: Taverna Workbench

A standalone workbench for life sciences workflows definition and execution. It allows to:

- build complex analysis workflows
- make access to both remote and local processors
- define alternative processors
- run workflows
- display results in various formats

It includes the myGRID ontology for bioinformatics data types and tasks

Requirements: java, Windows or Linux

Open source: <http://taverna.sourceforge.net/>

Current version: 1.7.1 (stable, 2.0 (beta))

# WfMS: Taverna Workbench

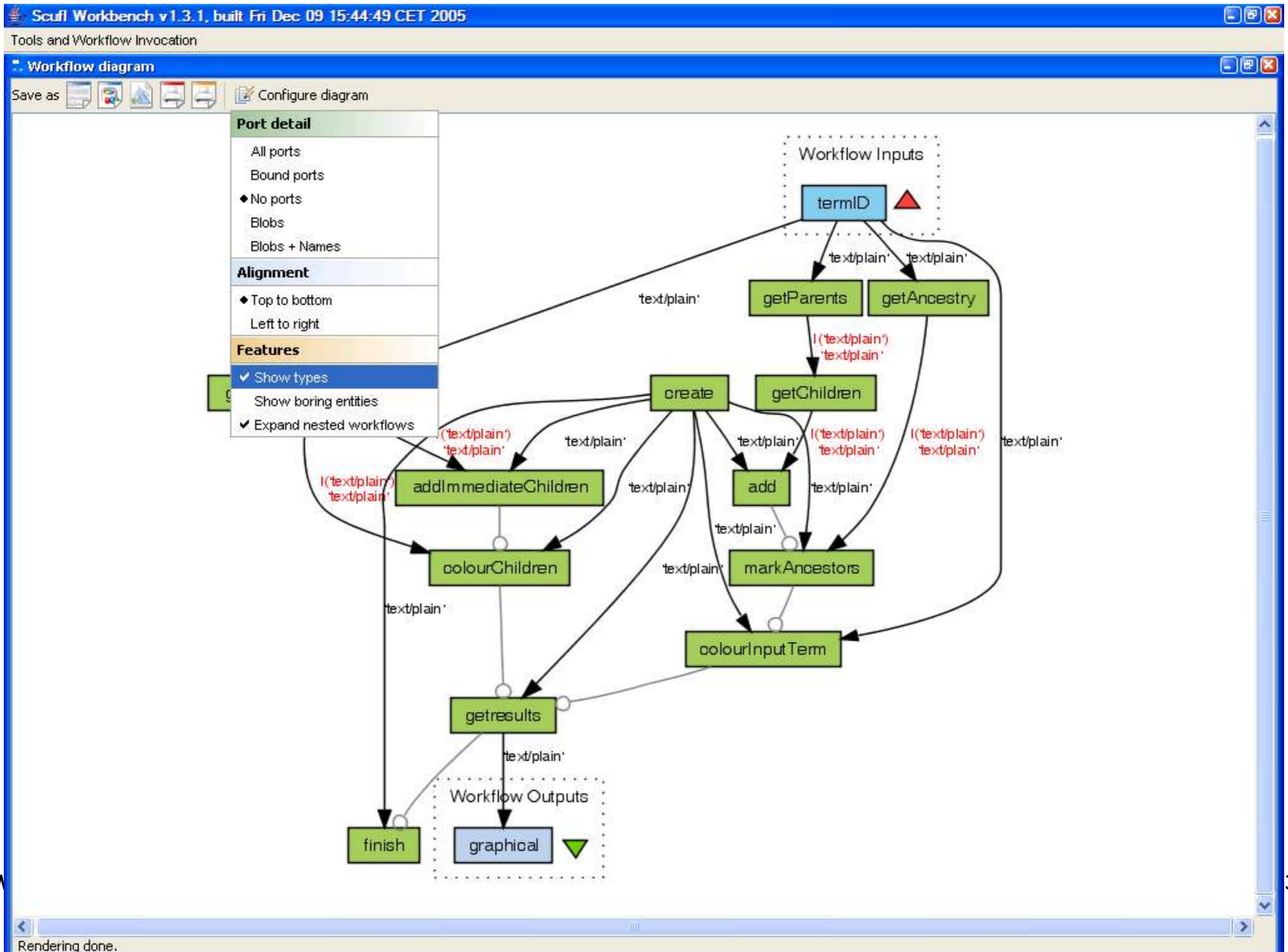
## Multi-windows GUI including:

- Advanced Model Explorer for workflow composition
- Workflow diagram display
- List of available services
- Workflow execution and results display window

## Rich web services management:

- Standard WSDL services
- Soaplab services
- BioMOBY registries
- XScufl Workflows
- Biomart databases
- Local processors: list/string, r/w, constants, beanshell scripts

# WfMS: Taverna Workbench data links



Workflow ▼ Metadata for 'libs'Ontology Description MIME Types

Pick from ontology

Find from regex :

Available ontologies :

- [-] root:Types
  - [-] task
    - filtering
    - retrieving
    - grouping
    - manipulating
    - calculating
    - merging
    - parsing
    - joining
    - summarising
    - removing
    - splitting
    - inserting
    - translating
    - searching
    - displaying
    - distinguishing
    - [-] aligning
  - [-] bioinformatics\_concept
    - [-] bioinformatics\_application
    - [-] bioinformatics\_data
    - [-] bioinformatics\_metadata
    - [-] bioinformatics\_database
    - [-] bioinformatics\_algorithm
    - [-] bioinformatics\_diagram
      - 3D\_plot\_of\_Gene\_Ontology\_lattice
      - nucleotide\_sequence\_feature\_diagram
      - ABI\_graph\_plot
      - sequence\_alignment\_dot\_plot
      - 2D\_alignment\_quality\_graph\_plot

Select from ontology or manually edit term below

[http://www.mygrid.org.uk/ontology#CABRI\\_cell\\_lines\\_catalogue](http://www.mygrid.org.uk/ontology#CABRI_cell_lines_catalogue)

# WfMS: Limitations and perspectives

WfMS are a promising methodology to solve data integration issues in biology

Nevertheless, they still have limitations.

If Semantic Web technologies can overcome such limitations, they can become the best tool.

Main limitations refer to:

- Abstraction in the definition of the tasks
- Performances of automated analysis
- Repeatability of the analysis

# WfMS: abstraction

## A greater abstraction is needed

Scientists' activity is oriented toward **scientific results**  
Building workflows and coping with services is a burden

Knowledge of services, data formats and programming skills  
are still required: *GUIs don't help*

A semantic interface can help if it includes:

- metadata management,
- annotation of services and databases,
- automatic format conversions,
- discovery of most appropriate services for user needs,
- automatic workflows composition.

The **best interface** should allow researchers to **build workflows** by describing the **required analysis** (almost) in **natural language**.

# WfMS: performances

## Performances

Scientists need **best results in the shortest time**, regardless which database, site or supercomputer is used (**complete transparency**); for this, are useful:

- **reuse** of intermediate results,
- **policy for distribution** of tasks (avoid huge data transfer)
- adding **metadata** to available services
- **automatic selection** of best services (fastest, most reliable)

Scientists need not to cope with **systems' faults**: high traffic networks, crash of networks and sites; for this, it may be useful:

- **transparency** of sources,
- use of **alternative services** for the same elaboration,
- ability to **identify failures, retry failed analysis**, suspend execution and restart stopped workflows

# WfMS: repeatability

## Repeatability of experiments

**Repeatability of experiments** is a fundamental requirement in biology. **Traces of executions** should be stored to enable it.

**NB!** In-silico elaborations are prone to frequent updating and evolution of databases. Perfect repeatability is hard!

The trace should include execution metadata:

- **trivial** information (workflow description, inputs used, processing software and sites accessed),
- **non-trivial** data (software and databases versions, operating systems of computers that run software),
- such data should be provided by Web Services

WfMS are increasingly providing data provenance features.

# Capacities of Semantic Web

"The advantages of the Semantic Web lie in its ability to present and provide access to **complex knowledge** in a **standardized form** making **interoperability between distributed databases** and middleware achievable."

The Semantic Web can address these issues since it supports:

- **integration of heterogeneous information systems**,  
by operating as a meta-database over heterogeneous information sources
- **a distributed environment**,  
reducing problems related to download of databases and to the evolution of data
- **evolving domain knowledge**,  
by adopting ontologies (instead of implicit semantics either biologists' common sense or semantics encoded in database structures and software tools)

So there is a chance: **what is needed** to adopt SW technologies in biology?

# Adoption of SW: ontologies

## Shared definitions of knowledge domains (ontologies)

Many efforts are now made towards this direction.

- **Association of ontological terms and concepts to existing data is still in its infancy and it refers to a few recognized and well known ontologies, e.g. Gene Ontology.**
- **Annotation of the huge amount of data with ontologies concepts really is a big task.**
- **Definition of new ontologies and their interlinking with existing ones are required to properly cope with the majority of information systems**
- **Anyway, the addition of semantic contents in current databases would give an essential contribution to integration of distributed biological information.**

# Adoption of semantics technologies

## RDF stores

There are difficulties

- biological databases stored in RDF or OWL are starting to appear as demonstration systems or prototypes
- a huge amount of data is still **only available in unstructured** or partially structured formats, accessible only through web interfaces.
- this is due to the requirement of **keeping production systems running** and accessible by means of current data analysis tools.

but

- most recent database implementations include XML based releases
- **automatic conversion of XML data to RDF** could constitute a crucial key for the exploitation of Semantic Web tools in this domain.

# Adoption of semantics technologies

## Data search and mining

- **search tools** able to **exploit biological ontologies**, information metadata and RDF stores have **not yet** been implemented.
- **search tools** based on **current query standards** (SPARQL, RDF Query Language) and **reasoners** are already **in place**
- so, once RDF stores are set up, the problem should mainly consist in the definition of **search proper interfaces**
- **interfaces** should be **user-friendly**, by hiding SPARQL queries to the users, and **flexible**, by allowing disparate queries, at the same time
- Is there space for a **controlled natural language** for biology?

# The W3C Interest Group on Semantic Web for Health Care and Life Science and its activities

# HCLSIG: history and mission

The interest Group for Health Care and Life Sciences (HCLSIG) was created in the context of the **Semantic Web activities** of the **World-Wide Web Consortium (W3C)**.

First discussed in October 2004, **announced** in November 2005, **confirmed and restarted** in 2008, with new charter and chairs.

**Mission** (from 2008 charter):

**"to develop, advocate for, and support the use of Semantic Web technologies for health care and life science, with focus on biological science and translational medicine"**

Currently, HCLSIG:

- provides an **open forum** for collecting application and implementation experiences
- addresses valuable **use cases**
- **disseminate** implementations (communications, tutorials and courses)

# HCLSIG: Biological sciences

Three main application domains considered:  
biology, translational medicine and health care.

For **biological sciences**, the focus is on semantic integration of main data repositories. In this context, activities aim at:

- **assisting** researchers, developers and publishers to **make information accessible** using Semantic Web technologies
- **promoting** use of **ontologies** for data integration
- **showing** how current bioinformatics analysis tools can use biological data structured by using **Semantic Web standards**

# HCLSIG: Translational medicine

## Translational medicine

Translational medicine relates to the possibility of **delivering personalized treatments** based on patients' molecular characterization.

Such personalized treatments would include **right dose, right time** (as usual) and **right drug**.

The **same disease** (phenotype) can be **better treated** by knowing the **individual characteristics** (genotype) of the patients.

This implies linking genomic information with clinic data ("bench to bedside").

**NB! Biological and medical data** have seldom been linked and they do not share nomenclatures at all.

HCLSIG activities in this domain therefore focus on:

- **connecting pre-clinical and clinical trial data** with **clinical decision support knowledge**
- **creation of dashboards** for enabling **integration of heterogeneous and disparate data** in support to treatment and therapy selection.

# HCLSIG: Health care

## Health care

HCLSIG goals aim at improving quality of care and at supporting clinical research.

Integration of medical records and clinical research systems (clinical trials) is meant to be one of the first applications.

This implies efforts for:

- **standardizing and harmonizing medical data**
- **developing ontologies for clinical medicine and investigations**
- **developing mappings between terminologies**

# HCLSIG: task forces

Five task forces have been defined:

BIORDF (Structured Data to RDF)

Scientific Publishing

Ontologies

Adaptive Healthcare Protocols and Pathways

Drug Safety and Efficacy

Active participation is only open to:

W3C partners

Invited experts

Many documents are freely available, mailing lists are open

# HCLSIG: SenseLab

The goal of the BIORDF group is the exploration of the effectiveness of **current SW tools for making biomedical data available as RDF.**

A use case referring to an **integrated neuroscience data environment (SenseLab)** was developed, including three databases converted to RDF and OWL-DL formats.

It allows to **query information** and to **run simulations** pertaining to the function of neurons both in healthy and disease states.

- **NeuronDB**: descriptions of anatomic locations, cell architecture and physiologic parameters of **neuronal cells**.
- **BrainPharm**: descriptions of **actions** of pathological and **pharmacological agents**.
- **ModelDB**: computational neuroscience **models** and **simulations**, annotated with references to NeuronDB.

# Gleaning Resource Descriptions from Dialects of Languages (GRDDL)

HCLSIG is supporting adoption of GRDDL, a technique for **extracting data from XML documents and build RDF triples**. Transformation rules must be specified for each schema.

GRDDL is particularly relevant for biomedical information since:

- it is often expressed in XHTML (evolution from plain HTML)
- XML biomedical stores are more and more being adopted

Using GRDDL, researchers can

- transform data in coherent RDF archives
- apply Semantic query languages and reasoners on these archives,
- integrate data from diverse sources by using reference ontologies.

# Banff demo

A **demonstration** of possibilities offered by Semantic Web technologies was given at **WWW2007** in Banff.

A **knowledge base integrating 15 distinct data sources** was built

- most data **translated to RDF** and managed in a triple store
- other data **maintained** in original database and **mapped to RDF**
- a **reasoner** was used to **infer triples** to increase expressiveness
- **SPARQL queries** and visualization tools were defined

The demonstration showed how data on signal transduction pathways, CA1 Pyramidal Neurons (CA1PN), their genes, and gene products, were integrated and queried to identify drug target candidates for Alzheimer's Disease.

# Banff demo: data sources

Information sources included, among the others:

- ontologies (GO, Galen, all OBO ontologies),
- literature and nomenclature (Medline, MeSH),
- genomic sources (HomoloGene, GOA)
- Alzheimer specific sources (Semantic Web Applications in Neuromedicine, SWAN).

Incorporated data for a total of:

- 350M triples
- 20Gb when stored in RDF.

# Banff demo: the question

Scientific question:

**"What genes are involved in signal transduction that are related to pyramidal neurons?"**

The query involved four data sources:

- MeSH to retrieve keywords related to pyramidal neurons,
- Medline for extracting papers and determining involved genes
- Entrez Gene for collecting data of involved genes
- Gene Ontology for restricting genes to those showing to perform a signal transduction function.

Results showed that many genes were linked to Alzheimer disease through the activity of the gamma secretase (presenilin) protein.

This is the end

Thank you!