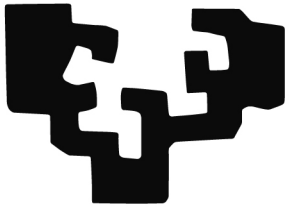
Use of Semantic Web resources for knowledge discovery

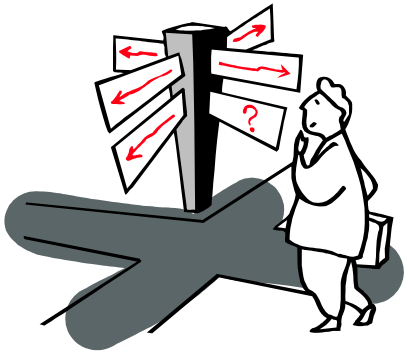
Mikel Egaña Aranguren
<mikel.egana@ehu.es>

eman ta zabal zazu



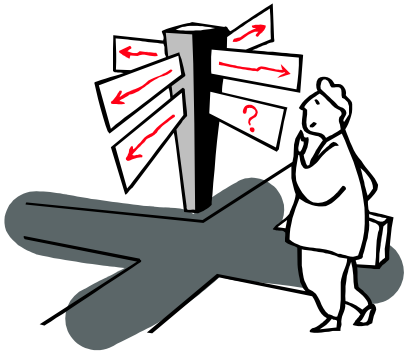
Universidad del País Vasco
Euskal Herriko Unibertsitatea

Contents



Contents

1. Introduction
   - State of affairs
   - Background
1. The Cell Cycle Ontology
2. BioGateway
3. Concluding remarks
4. Future prospects



State of affaires

- The amount of data generated in the biological experiments continues to grow exponentially (e.g. NGS)
- The shortage of proper approaches or tools for analysing this data has created a gap between raw data and knowledge
- The lack of a structured documentation of knowledge leaves much of the data extracted from these raw data unused
- Differences in the technical languages used (synonymy and polysemy) have complicated the analysis and interpretation of data
- Many of our tasks (will) require correct and meaningful communication and integration among the project information resources
- So, a major barrier to such interoperability is semantic heterogeneity: different applications, databases, and agents may ascribe disparate meanings to the same terms or use distinct terms to convey the same meaning

Strategy

Steps:

1. **Problem definition:** test bed case (*e.g.* cell cycle process, forestry, anatomy, …)
2. **Data scaffold elements:** standards, terminologies and ontologies
3. Development of **tools**
4. **Data integration** and **exploitation**
5. **Beyond** the domain: *e.g.* cell cycle → all processes in the Gene Ontology

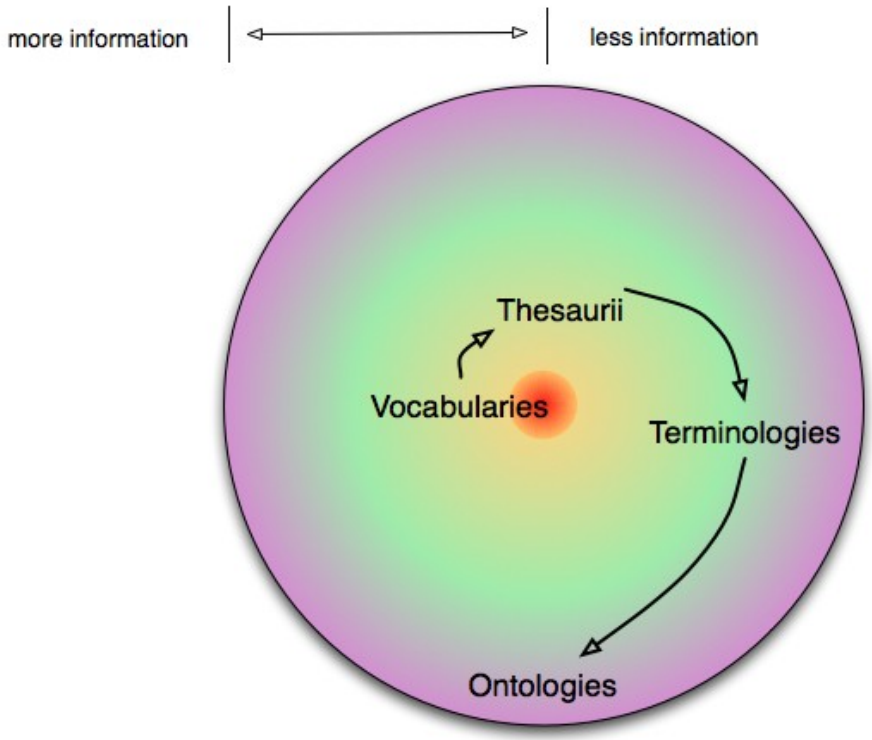
What is an ontology?

- **(Too)** many definitions:
  - "The science of categorizing beings (or their existence)" (Aristotle ~350BC)
  - "A formal specification of a conceptualization" (Gruber, 1993) -- most cited definition

  "A formal representation of knowledge domains" (Bard and Rhee, 2004)
- Computer scientist
  - *"A* specific ***artefact*** designed with the purpose of expressing the ***intended meaning*** of a (shared) ***vocabulary"***

- Life Sciences / Bio-ontologist
  - *"A controlled vocabulary of biological terms and their relations"*

Graphical overview *



* image: T. Clark

Extended definition

- “An ontology is a computer-interpretable specification that is used by an agent, application, or other information resource to declare what terms it uses, and what the terms mean.”
- Ontologies support the semantic integration of software systems through a shared understanding of the terminology in their respective ontologies.

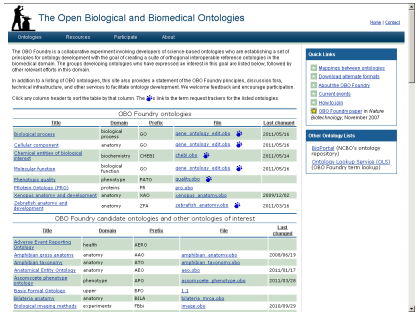
Why do we need them?

- Share and reuse information (common terminology)
- Data integration
- Other applications (*e.g.* analysis, annotations)

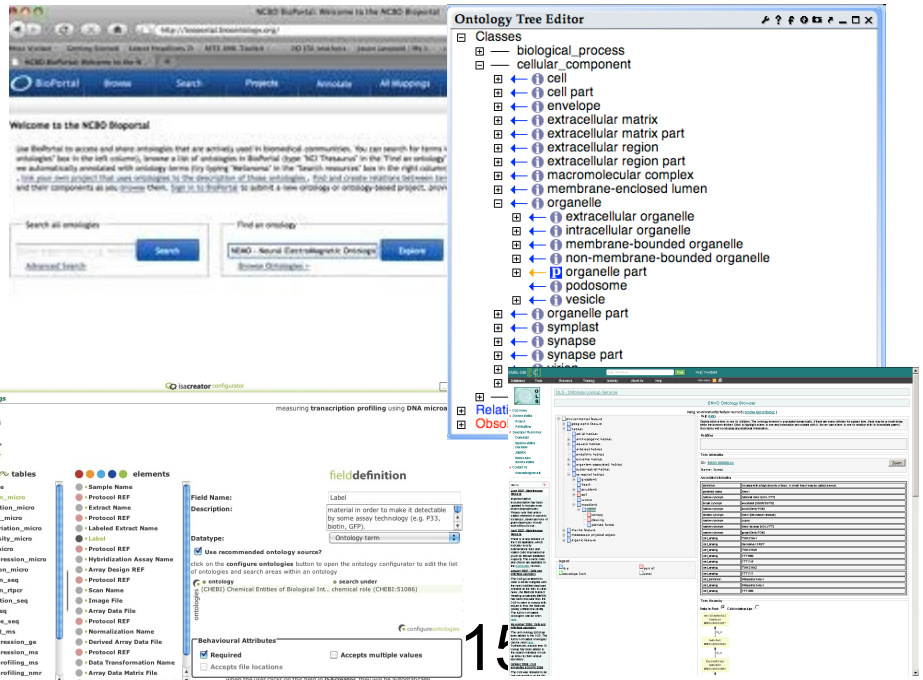
Multidisciplinary teams: philosophers, computer scientists, domain experts (*e.g.* forester), legal / IP, …

Good news

- Many efforts:
  - OBO foundry
  - RDF foundry
  - HCLS – SIG
  - BioDBCore
  - iPlant
  - …
- Tools
  - BioPortal
  - OLS
  - ISA-Tools
  - OBO-Edit, Protégé
  - …



W3C Semantic Web



Take-home messages (so far…)

1. An ontology is <u>more</u> that just a collection of ‘standard’ terms
2. Ontology building is a multidisciplinary field (*e.g.* computer scientists, foresters, molecular biologists, lawyers, …); therefore, we need each other…
3. An ontology is a “living entity”: it is constantly changing/evolving…
4. Moreover, we cannot live isolated: community effort; therefore, we *should* share our terms as well as continue learning from other ontologies
5. Therefore, the success of a data integration project will depend on most (if not all) of those “components”!

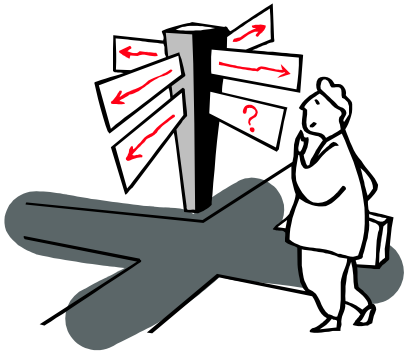
Semantic Web

- *"Next generation of the current web"*
- Goal: machine understandable content
- Keyword search will get obsolete
  - I get too many (irrelevant) hits
  - Complex query formulation (desired)
- Still a vision (technology under development)
- Life scientists are very interested
  - Health Care and Life Sciences (HCLS IG - W3C)
  - Several meetings, consortia, investments, etc.

Project	Keywords	Technologies	Website
LinkHub	document ranking, text categorisation, query corpus	RDF	http://hub.gersteinlab.org/
Lipid bibliosphere	lipids, metabolites, reasoning	OWL	
Neurocommons	uniform access, package-based distribution	RDF, SPARQL	http://neurocommons.org/
RDFScape	systems biology, cytoscape, reasoning	RDF, SPARQL	http://www.bioinformatics.org/rdfscape
S3DB	lung cancer, omics	RDF	http://www.s3db.org/
SWAN - AlzPharm	neuromedicine, alzheimer, neurodegenerative disorders	RDF, OWL	http://swan.mindinformatics.org
SEMMAS	web services, intelligent agents	OWL	http://semmas.inf.um.es/prototypes/bioinformatics.html
SOMWeb	distributed medical communities	RDF, OWL	http://www.cs.chalmers.se/proj/medview/somweb
Thea-online	protein interactions, annotations, pathways	RDF, SPARQL	http://bioinfo.unice.fr:8080/thea-online/
yOWL	yeast, phenotypes, interactions	OWL	http://ontology.dumontierlab.com/yowl-hcls

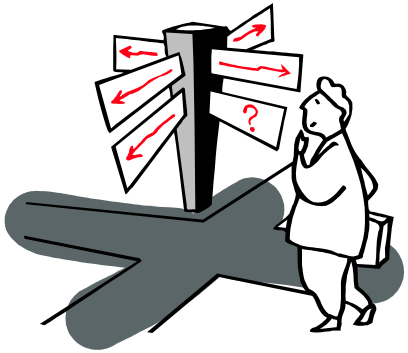
Project	Keywords	Technologies	Website
Bio2RDF	mashup, linked data, global warehouse, complex queries	RDF, SPARQL	http://bio2rdf.org/
BioDash	disease, compounds, therapeutic model, pathway	RDF, OWL	http://www.w3.org/2005/04/swls/BioDash/Demo/
BioGateway	semantic systems biology, hypothesis generation	RDF, SPARQL	http://www.semantic-systems-biology.org/biogateway/
CardioSHARE	collaborative, distributed knowledgebase, reasoning, web services	RDF, SPARQL	http://cardioshare.icapture.ubc.ca/
Cell Cycle Ontology (CCO)	cell cycle, protein-protein interactions, reasoning, ontology patterns	RDF, OWL, SPARQL	http://www.cellcycleontology.org/
CViT	cancer, tumor, gene-protein interaction networks	RDF	https://www.cvit.org/
FungalWeb	fungal species, enzyme substrates, enzyme modifications, enzyme retail	OWL	
GenoQuery	genomic warehouse, mixed query, tuberculosis	RDF, SPARQL	http://www.lri.fr/~lemoine/GenoQuery/
HCLS W3C	knowledge base, life sciences, prototype	RDF, OWL, SPARQL	http://www.w3.org/TR/hcls-kb/
Kno.e.sis	nicotine dependence, biological pathway	RDF, SPARQL, OWL	http://knoesis.wright.edu/research/semsci/application_domain/sem_life_sci/bio/research/

Contents



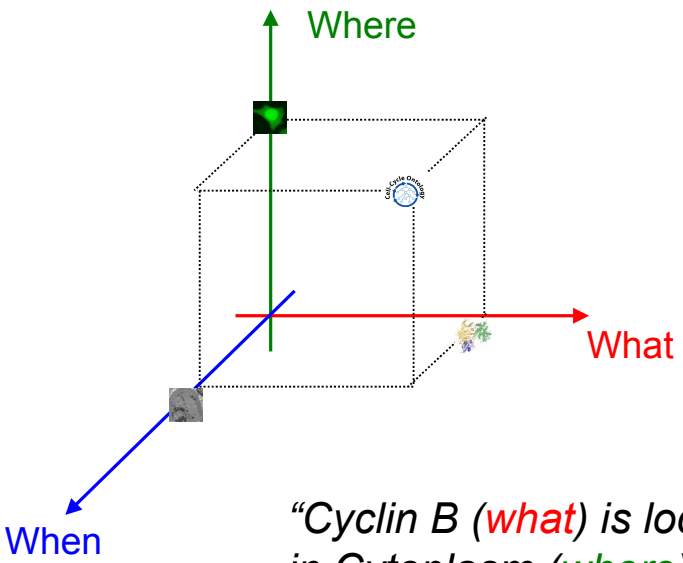
Contents

1. Introduction
2. The Cell Cycle Ontology
   - A knowledge base for cell cycle elucidation
     Antezana E. et al. Genome Biology, 2009
   - http://www.cellcycleontology.org
1. BioGateway
2. Concluding remarks
3. Future prospects



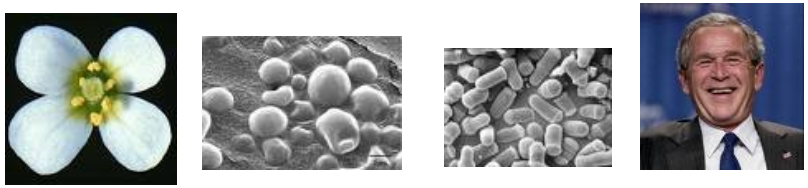
The Cell Cycle Ontology in a nutshell

- Capture knowledge of the Cell Cycle process
- “Dynamic” aspects of terms and their interrelations
- Promote sharing, reuse and enable better computational integration with existing resources



“Cyclin B (what) is located in Cytoplasm (where) during Interphase (when)”

ORGANISMS:



Users:

- Molecular biologist
- Bioinformatician / Computational Systems Biologist
- General audience

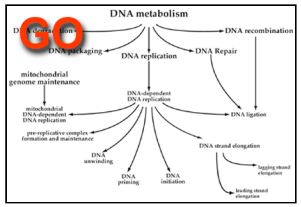
Antezana E. et al. Lect. Notes Bioinformatics, 2006
Antezana E. et al. Genome Biology, 2009

Knowledge representation in CCO

- Why OBO?
  - "Human readable"
  - Standard
  - Tools (e.g. OBOEdit)
  - http://obo.sourceforge.net
- Why OWL?
  - Web Ontology Language
  - "Computer readable"
  - Reasoning capabilities vs. computational cost ratio
  - Formal foundation (Description Logics)
  - Tools (e.g. Protégé)
- Ontology manipulation:
  - ONTO-PERL (*Antezana E. et al. Bioinformatics 2008*)
  - ONTO-Toolkit (*Antezana E. et al. BMC Bioinformatics 2010*)

CCO sources

- Ontologies
  - Gene Ontology (GO)
  - Relationships Ontology (RO)
  - Molecular Interactions (MI)
  - Upper level ontology (ULO/BFO)
- Data sources
  - SWISS-PROT
  - GAF
  - PPI: IntAct
  - Orthology (Decypher)



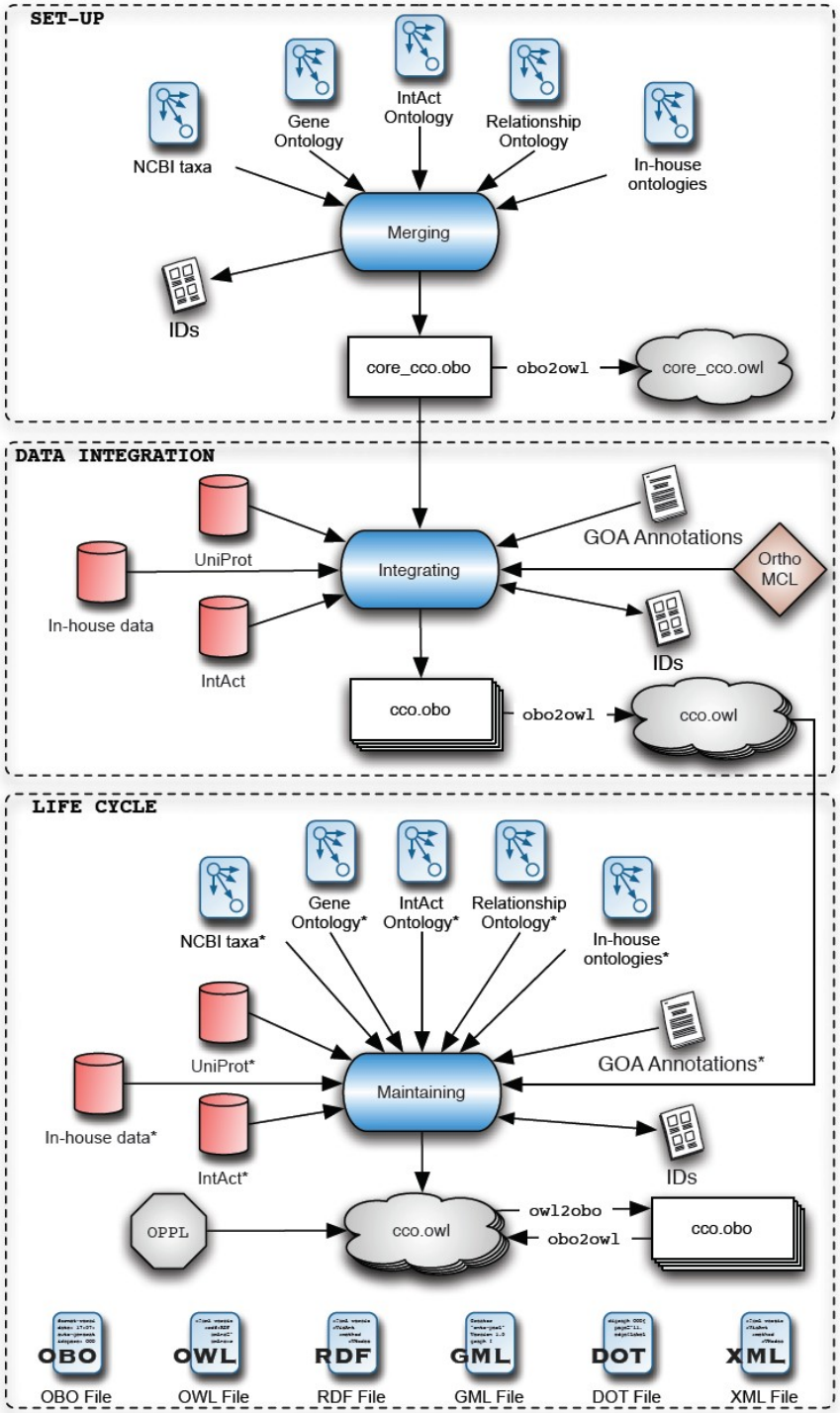
The Open Biomedical Ontologies

Type of proteins	Ontology				
	At	Hs	Sc	Sp	CCO
Core cell cycle	162	870	602	749	2383
Added from IntAct	70	1067	2542	109	3788
Modified proteins added from UniProt	27	4577	8291	486	17985
TOTAL	259	6514	11435	1344	24156

Entity	Ontology				
	At	Hs	Sc	Sp	CCO
Proteins	2958	15742	1996	12914	33610
Genes	2100	3919	3474	1246	10739
Orthology types	—	—	—	—	1653

CCO is the composite ontology = At + Hs + Sc + Sp + orthology ; **33610** proteins in CCO

CCO Pipeline

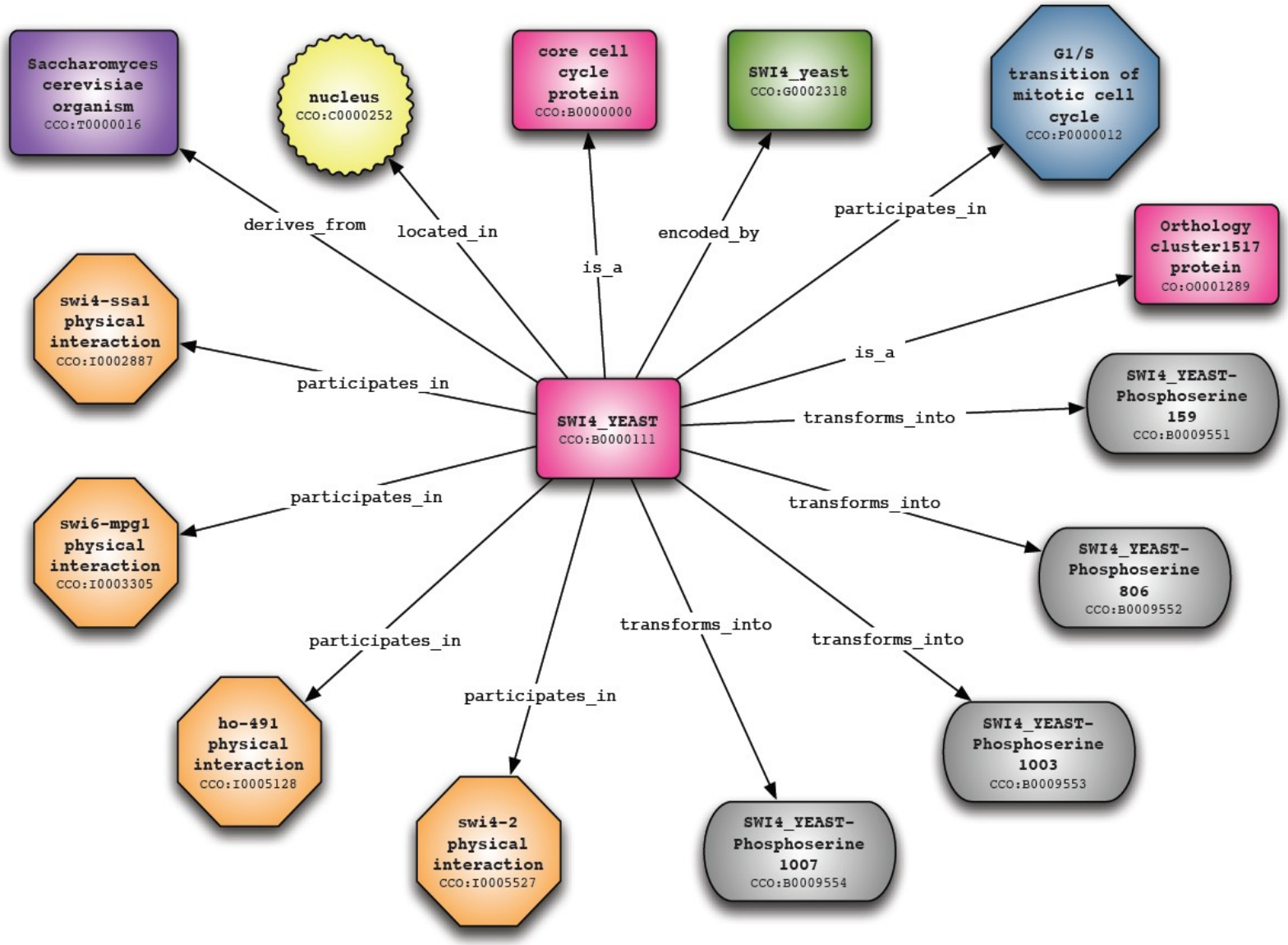


- ontology integration (ONTO-PERL)
- format mapping

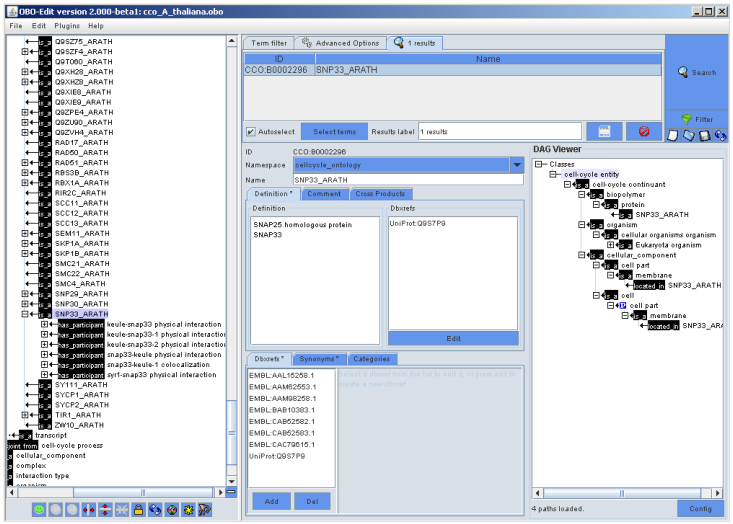
- data integration
- data annotation
- consistency checking

- maintenance
- data annotation
- semantic improvement: OPPL (*Egaña M. et al. OWL-ED, 2008*)
- ODP (*Egaña M. et al. BMC Bioinf. 2008*)

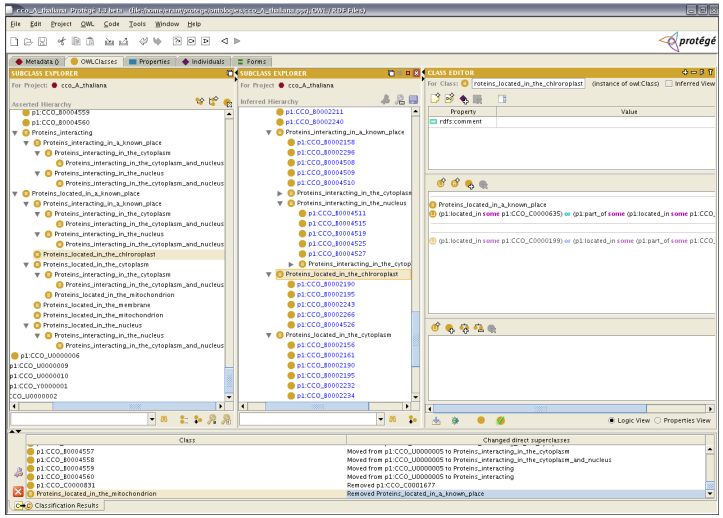
Sample piece of information in CCO



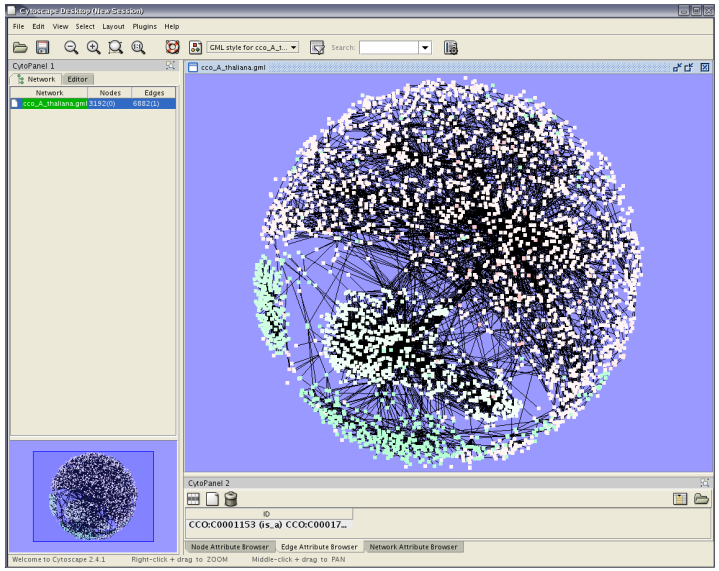
Exploring CCO (1/2)



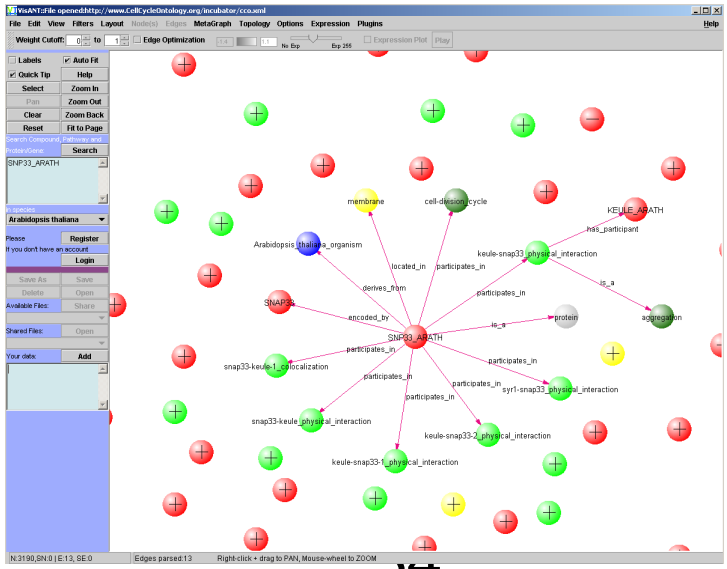
OBO-Edit



Protégé

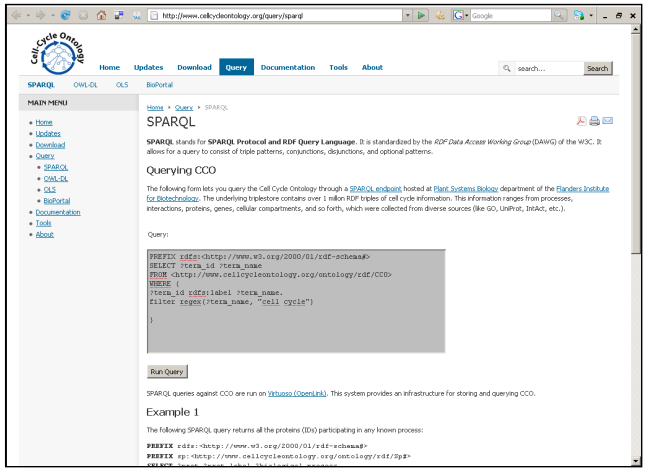


Cytoscape

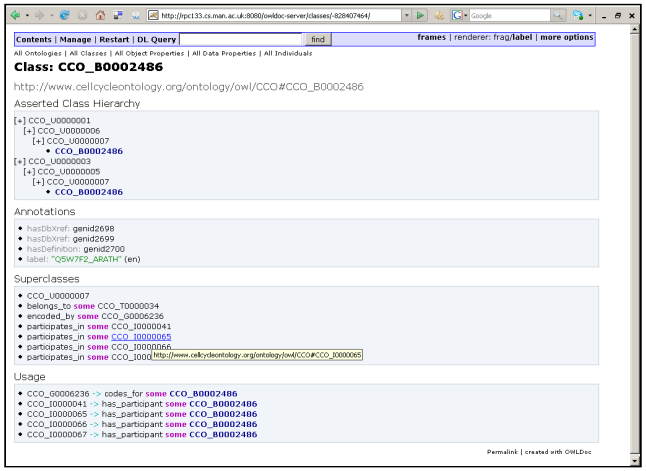


visANT

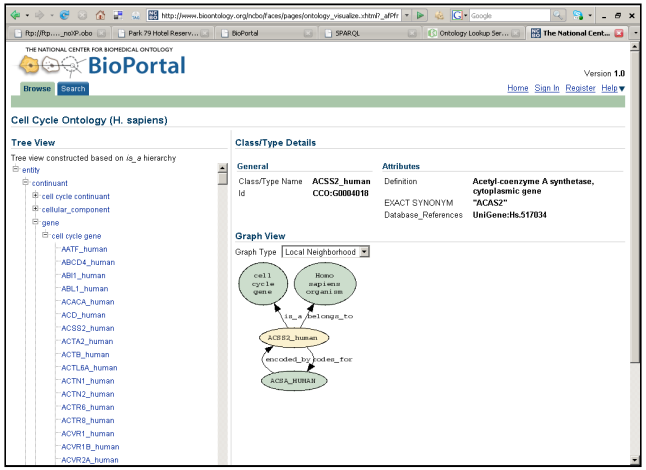
Exploring CCO (2/2)



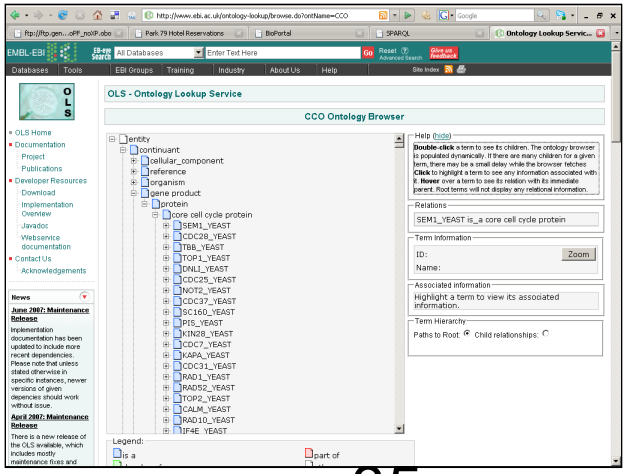
CCO website (SPARQL)



OWLDoc server



BioPortal



Ontology Look up Service

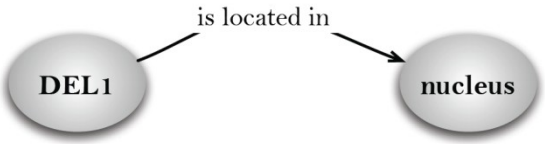
Advanced Querying

- RDF = **R**esource **D**escription **F**ramework
  - Metadata model: elements = resources
- It allows expressing knowledge about web resources in statements made of triples (basic information unit) :

Subject –- Predicate –- Object

- Subject corresponds to the main entity that needs to be described.
- Predicate denotes a quality or aspect of the relation between the Subject and Object.

- Example: "*The protein DEL1 is located in the nucleus*"
- It "means" something…

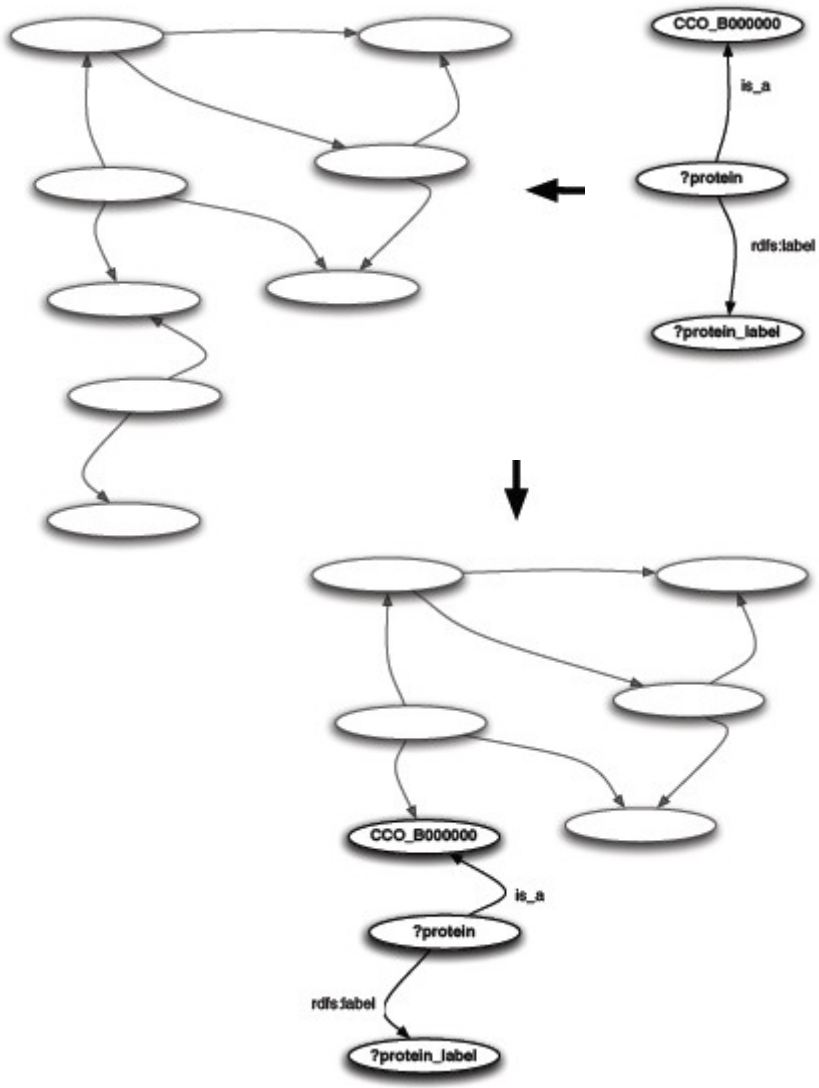


SPARQL*

- Query RDF models (graphs)
- Powerful, flexible
- Its syntax is similar to the one of SQL.
- Virtuoso Open Server**
- Benchmarking ***
- Example (matching two triples):

?protein **sp:is_a** sp:CCO_B0000000 .

?protein **rdfs:label** ?protein_label



* http://www.w3.org/TR/rdf-sparql-query/
** http://www.openlinksw.com/
*** Mironov V. et al. SWAT4LS, 2010

SPARQL

SPARQL stands for **SPARQL Protocol and RDF Query Language**. It is standardized by the *RDF Data Access Working Group* (DAWG) of the W3C. It allows for a query to consist of triple patterns, conjunctions, disjunctions, and optional patterns.

Querying CCO

The following form lets you query the Cell Cycle Ontology through a SPARQL endpoint hosted at Plant Systems Biology department of the Flanders Institute for Biotechnology. The underlying triplestore contains over 1 millon RDF triples of cell cycle information. This information ranges from processes, interactions, proteins, genes, cellular compartments, and so forth, which were collected from diverse sources (like GO, UniProt, IntAct, etc.). Type your SPARQL query in the following text area, then click on 'Run Query'. A new window with the results will be opened. In case there is a syntax error in the query, it will be warned to you. (**N.B.** Recommended browsers: Firefox, Safari, Opera, or Konqueror. IE proposes to save the results instead of displaying them.)

Query:

```
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX sp:<http://www.cellcycleontology.org/ontology/rdf/Sp#>
SELECT ?prot_name ?biological_process_name
FROM <http://www.cellcycleontology.org/ontology/rdf/Sp>
WHERE {

?prot sp:is_a sp:CCO_B0000000 .
?prot rdfs:label ?prot_name .
?prot sp:participates_in ?biological_process .
?biological_process rdfs:label ?biological_process_name
}
```

Run Query Reset

SPARQL queries against CCO are run on Virtuoso (OpenLink). This system provides an infrastructure for storing and querying CCO.

Suggested PREFIXes:

http://www.cellcycleontology.org/query/sparql

Cell-Cycle Ontology

Home | Updates | Download | Query | Documentation | Tools | About

search... Search

SPARQL | OWL-DL | OLS | BioPortal

MAIN MENU

- Home
- Updates
- Download
- Query
  - SPARQL
  - OWL-DL
  - OLS
  - BioPortal
- Documentation
- Tools
- About

Home ▸ Query ▸ SPARQL

SPARQL

SPARQL stands for **SPARQL Protocol and R** ... f the W3C. It allows for a query to consist of triple pattern

Querying CCO

The following form lets you query the Cell Cyc ... nders Institute for Biotechnology. The underlying ... sses, interactions, proteins, genes, ... , etc.). Type your SPARQL query in the following ... on 'Run ... is a syntax error in the query, it will be warned to you ... ed browsers: Firef ... ave the results instead of displaying them.)

"all the core cell cycle proteins (*S.pombe*) participating in a known process"

Query:

```
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX sp:<http://www.cellcycleontology.org/ontology/rdf/Sp#>
SELECT ?prot_name ?biological_process_name
FROM <http://www.cellcycleontology.org/ontology/rdf/Sp>
WHERE {

?prot sp:is_a sp:CCO_B0000000 .
?prot rdfs:label ?prot_name .
?prot sp:participates_in ?biological_process .
?biological_process rdfs:label ?biological_process_name
}
```

Run Query | Reset

SPARQL queries against CCO are run on Virtuoso (OpenLink). This system provides an infrastructure for storing and querying CCO.

Suggested PREFIXes:

prot_name	biological_process_name
UBC11_SCHPO	G2%2FM transition of mitotic cell cycle
UBC11_SCHPO	cell cycle
UBC11_SCHPO	mitosis
UBC11_SCHPO	mitotic metaphase%2Fanaphase transition
UBC11_SCHPO	regulation of mitotic cell cycle
UBC11_SCHPO	cyclin catabolic process
SRW1_SCHPO	cell cycle
SRW1_SCHPO	cyclin catabolic process
SRW1_SCHPO	activation of anaphase-promoting complex during mitotic cell cycle

http://www.semantic-systems-biology.org/apo/queryingcco/sparql

SEMANTIC SYSTEMS BIOLOGY

HOME | BIOGATEWAY | APO | METAREL | BIOCURATION | TOOLS | EVENTS | ABOUT | FAQ | PHYLOGENETIC ONTOLOGY

Home › APO › Querying › SPARQL

Querying Application Ontologies with SPARQL

SPARQL stands for **SPARQL Protocol and RDF Query Language**. It is standardized by the *RDF Data Access Working Group* (DAWG) of the W3C. It allows for a query to consist of triple patterns, conjunctions, disjunctions, and optional patterns.

Querying Application Ontologies

The following form lets you query the Cell Cycle Ontology and the application ontologies included in the Gene Expression Knowledge Base through a SPARQL endpoint hosted by the Systems Biology group at the Norwegian University of Science and Technology (NTNU). The underlying triplestore contains over 25 million RDF triples of cell cycle information. This information ranges from processes, interactions, proteins, genes, cellular compartments, and so forth, which were collected from diverse sources (like GO, UniProt, IntAct, etc.). Type your SPARQL query in the following text area, then click on 'Run Query'. A new window with the results will be opened.

Sample queries:

[Select a query]

Query:

Run | Prefixes | Comment | Uncomment | Optional | Indent | FROM | UNION

NEWSFLASH

Doctoral Degree for Aravind Venkatesan On 13th Feb 2014, Aravind Venkatesan successfully defended his Ph.D. thesis **"Application of Semantic Web Technology to establish knowledge management and discovery in the Life Sciences"** and was awarded the title of Doctor in Sciences by NTNU.

BioGateway 2.00
Major revision
1. Local URIs replaced with external resolvable URIs
2. Data from SwissProt extended to UniProt
3. Data from IntAct added
4. Scope limited to 147 Reference Proteomes
5. Inferred triples limited to those supported explicitly by the ontology plus 'priority over is_a'

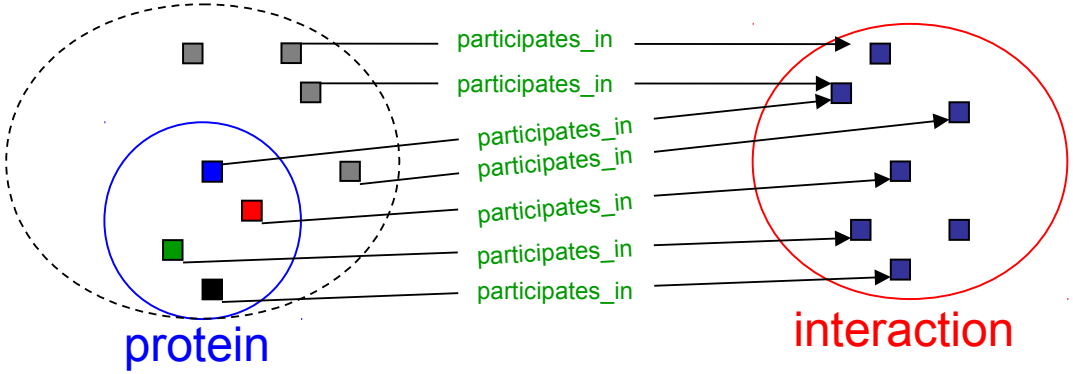
Genes2GO: **Gene Ontology matrix builder**
Genes2GO builds a binary matrix with genes and GO IDs/Terms.

Reasoning over CCO

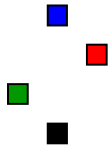
- OWL-DL: balance tractability with expressivity
- Consistency checking: no contradictory facts
- Classification: implicit2explicit knowledge
- Tools: Protégé, Reasoners (e.g. RACER, Pellet)

- Sample Query
  - *"Which cell cycle related proteins participate in a reported interaction?"*

protein **and**
participates_in **some** interaction

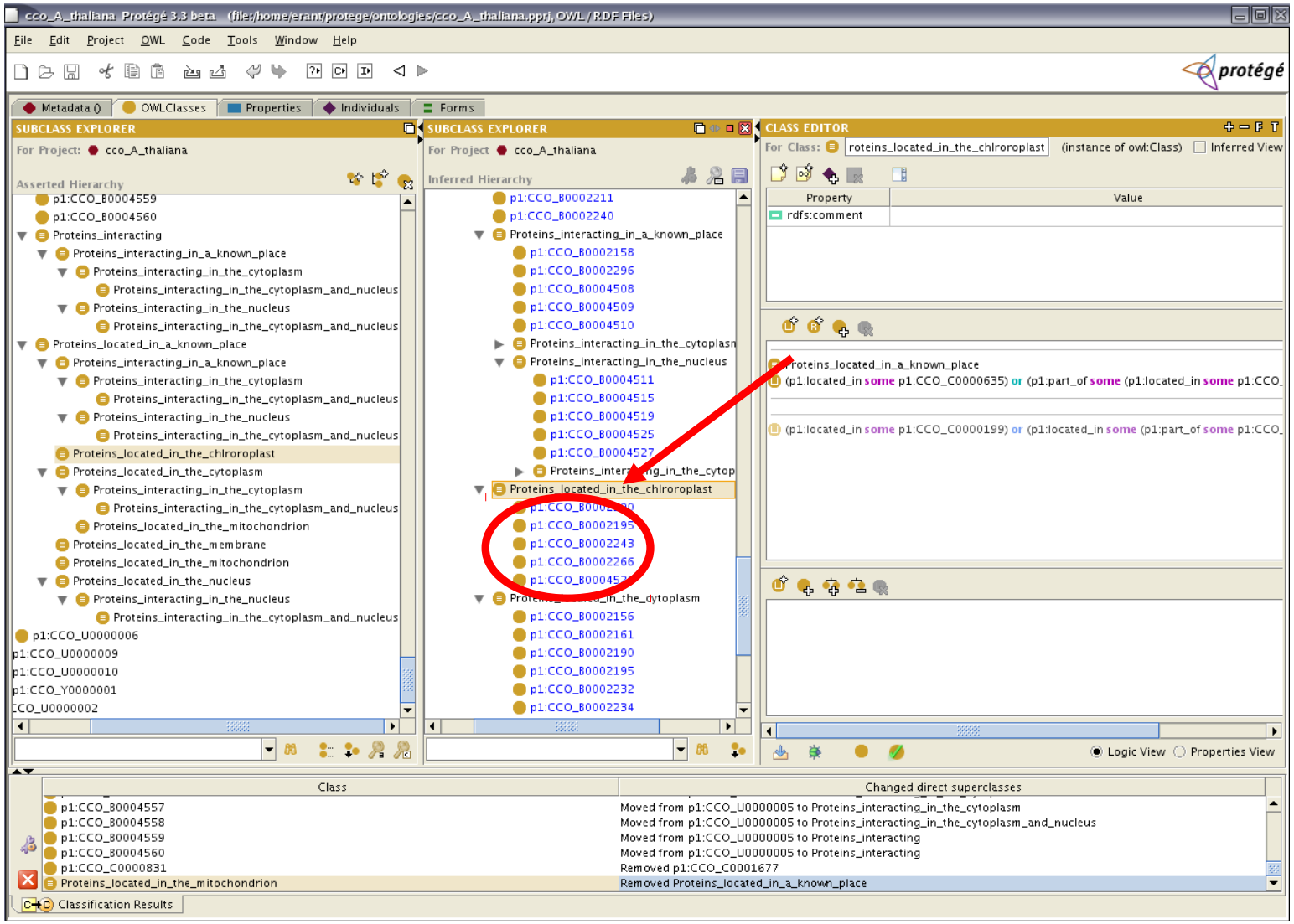


Answer:



Cellular localization checks

- Query: *"If a protein is cell cycle regulated, it must not be located in the chloroplast (IDEM: mitochondria)"* (RACER*)



* http://www.racer-systems.com

Conclusions

- Adequate knowledge representation:
  - enables automated reasoning (many inconsistencies could be detected)
  - simple biological hypothesis generation
- Data integration based on trade-offs (*e.g.* multiple inheritance)
- Performance issues (technology limitations)
- (Work in progress → GRAO)

Contents

1. Introduction
2. The Cell Cycle Ontology
3. BioGateway
   - An integrative approach for supporting Semantic Systems Biology
     Antezana et al. BMC Bioinformatics, 2009
   - http://www.SematicSystemsBiology.org
1. Concluding remarks
2. Future prospects

BioGateway

- From “cell cycle” to the entire set of processes in the Gene Ontology

- **CCO:** deep downwards (coverage)
- **BioGateway:** broad coverage
- **BioGateway’s goal:** build “complex” queries over the entire set of organisms annotated by the GAF

- Support a Semantic Systems Biology approach*

** Antezana et al. BMC Bioinformatics, 2009*

Systems Biology

- Yet another definition
- Key term: system
- What is a system?
- System =
  - set of elements,
  - dynamically interrelated,
  - having an activity,
  - to reach an objective (sub-aims),
  - **INPUT**: energy/matter/data
  - **OUTPUT**: energy/matter/information

Systems Biology (cont.)

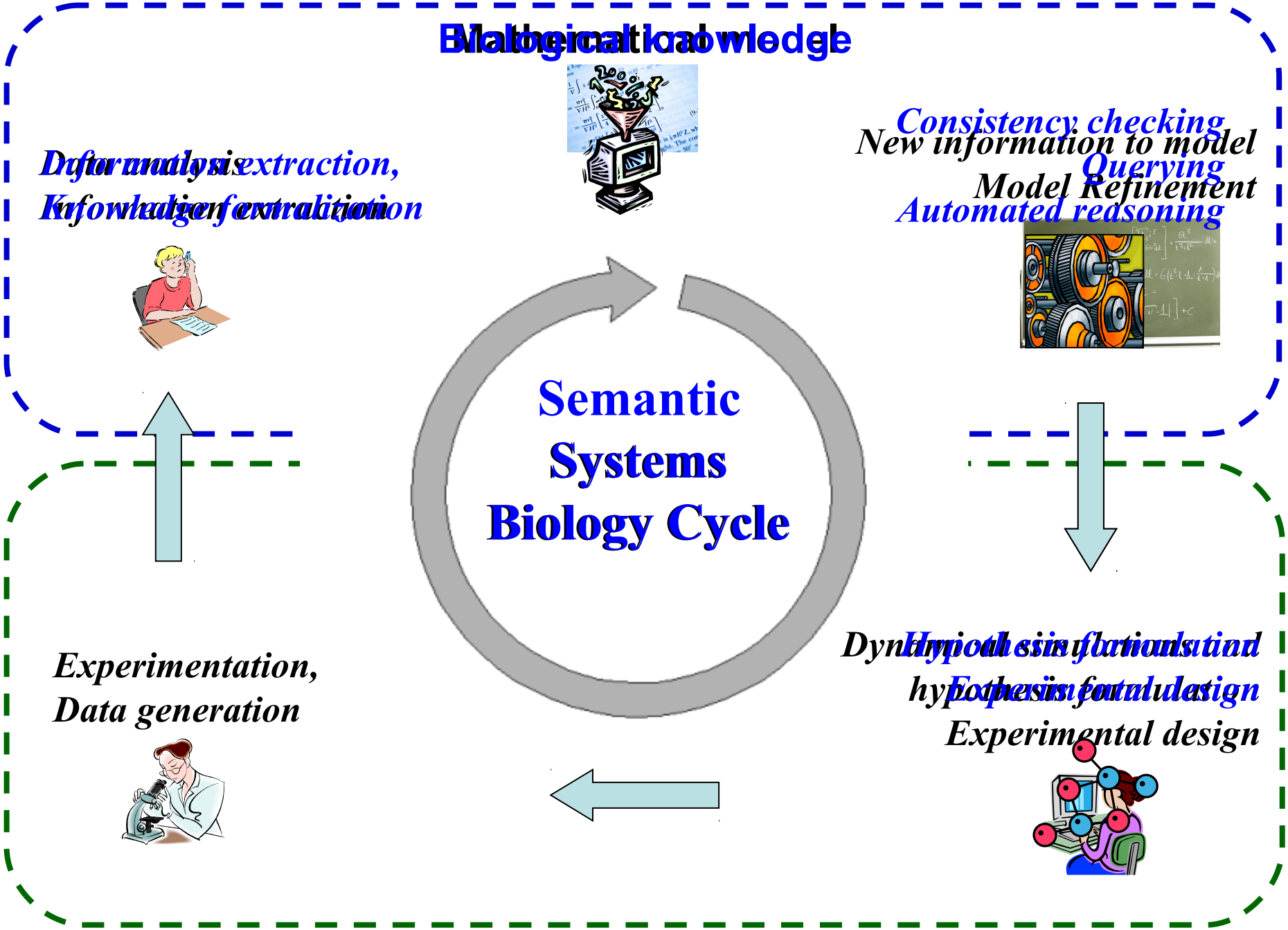
- “A system (and its properties) cannot be described in terms of their terms in isolation; its comprehension emerges when studied globally”
- Systems Biology = Approach to study biological systems.
- Arbitrary borders
- A system within a system

Systems Biology (cont.)

- Types of systems biology:
  - "Standard/Classical" Systems Biology
    (Kitano, Science *2002. Sauer et al, Science 2007)*
  - Translational Systems Biology
    (Vodovotz, PLoS Comp Biol *2008)*
  - Semantic Systems Biology
    (Antezana et al, Brief. in Bioinformatics *2009)*

Semantic Systems Biology

- Semantic?
  - New emerging technologies for analyzing data and formalizing knowledge extracted from it
- A new paradigm elements:
  - Knowledge representation
  - Reasoning ==> hypothesis
  - Querying

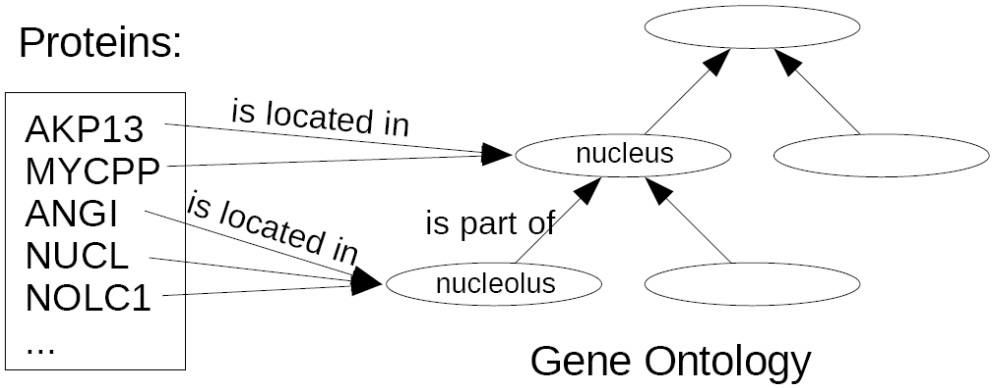

Biological knowledge
Mathematical knowledge
Data analysis
Information extraction,
Knowledge formalization
Consistency checking
New information to model
Querying
Model Refinement
Automated reasoning
Semantic
Systems
Biology Cycle
Experimentation,
Data generation
Dynamic simulations and
hypothesis formulation
Hypothesis formulation
Experimental design
Experimental design

BioGateway: a tool to support Semantic Systems Biology

- Automatic data integration pipeline (~1 x Year)
- **Quick query results:** performance, choice: "tuned" RDF (no OWL), 1 graph per resource
- **Human "readable" output:**
  - labels, no IDs or URI…
- **Good practice:**
  - Standards (RDF) => orthogonality, …
  - Representation issues (*e.g.* n-ary relations)
- **Transitive closure**:
  - *is_a (*subsumption relation), *part_of (partonomy)*

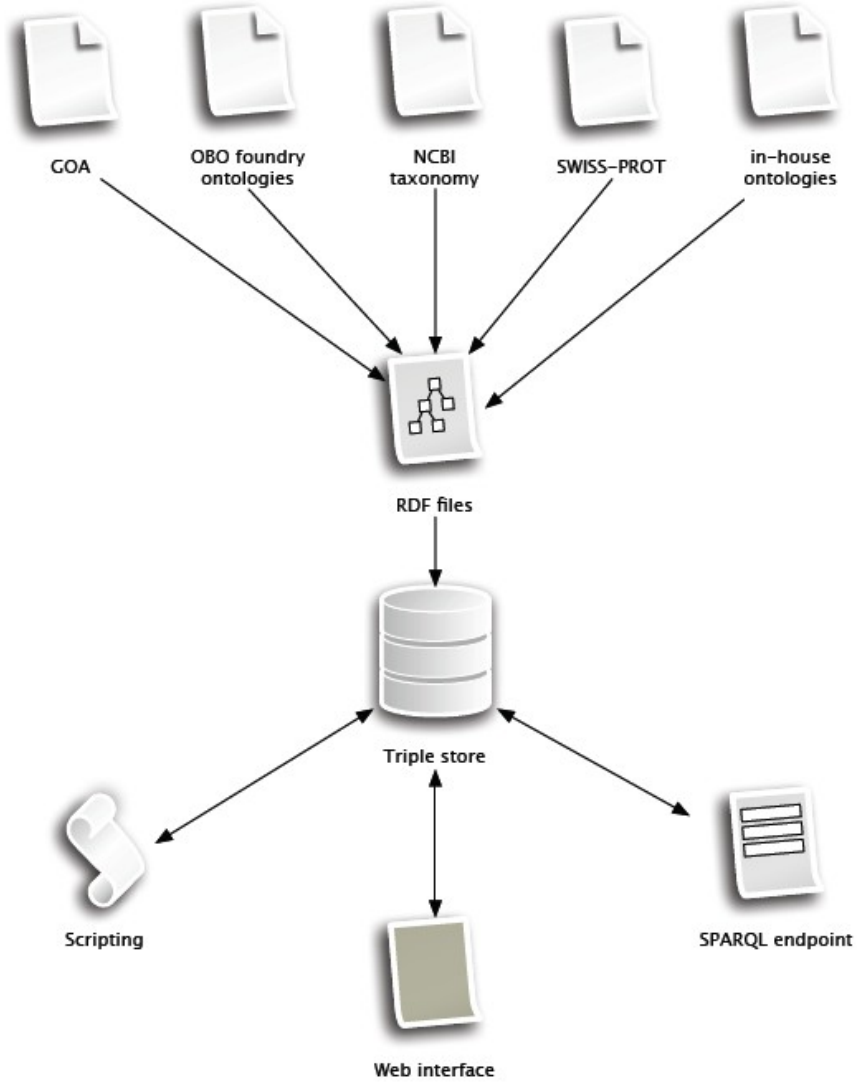
Transitive closure graphs

- If ***A part_of B***, and ***B part_of C***, then ***A part_of C*** is also added to the graph/ontology.
- Many interesting queries can be done in a performant way with it, like ***'What are the proteins that are located in the cell nucleus or any subpart thereof?'***
- The graphs **without** transitive closure are available for querying as well.



Blondé, W. et al. ICBO, 2009
Blondé, W. et al. Bioinformatics, 2011

BioGateway pipeline



- **1 Swiss-Prot** file, the section of UniProt KB of proteins
- **1 NCBI** file with the taxonomy of organisms + closures
- **1 Metaonto** file with information about OBO Foundry ontologies
- **2 Metarel** files with relation type properties
- **5 CCO** files with integrated information about cell cycle proteins
- **84 OBO Foundry** files with diverse biomedical information + Transitive Closure
- **51 Transitive Closure** files to enhance query abilities
- **1983 GOA** files with GO annotations + closures
- **1 OBI**
- **1 GALEN**

BioGateway holds 1,979,717,488 RDF triples!!!

A library of queries*

- The drop-down box contains 35 queries:
  - 23 protein-centric biological queries:
    - The role of proteins in diseases
    - Their interactions
    - Their functions
    - Their locations
    - …
  - 21 ontological queries:
    - Browsing abilities in RDF like getting the neighbourhood, the path to the root, the children,...
    - Meta-information about the ontologies, graphs, relations
    - Queries to show the possibilities of SPARQL on BioGateway, like counting, filtering, combining graphs,...
    - …

* http://www.semantic-systems-biology.org/biogateway/querying

SPARQL - Mozilla Firefox

y Bookmarks Tools Help

http://www.semantic-systems-biology.org/biogateway/querying

BioGateway: an ontology-driven query tool for enabling Semantic Systems Biology (SSB)

The following form lets you query the ontology-driven knowledgebase through a SPARQL endpoint hosted at Plant Systems Biology department of the Flanders Institute for Biotechnology. The underlying triplestore contains over **180 millon RDF triples** of information: the UniProt knowledgebase, the candidate OBO foundry ontologies, and the Gene Ontology Annotations. The information range spans processes, interactions, proteins, genes, cellular compartments, and more. Type your SPARQL query in the text area below, then click on 'Run Query'. A new window with the results will be opened. In case there is a syntax error in the query, you will be warned.

Recommended browsers: Firefox, Safari, Opera, or Konqueror. IE proposes to save the results instead of displaying them.

N.B. This system is still a **prototype**. Any feedback about BioGateway is very welcome. If you want to query CCO, please go to Querying CCO.

Sample queries:

Ont 20. Get the closest common parent in the hierarchy.

——————————Biological Queries——————————
Bio 1. Get the proteins with a specific function, location and process for all the annotated organisms.
Bio 2. Get functional, locational, process and disease information about a given protein.
Bio 3. Get the proteins that are involved in the 'psoriasis' disease.
Bio 4. Get the proteins that participate in the same process as a given protein.
Bio 5. Get the proteins that are located in the nucleus.
Bio 6. Get the amount of interactors for the proteins in a PPI network.
Bio 7. Get all the core cell cycle proteins participating in any known process (in S. pombe).
Bio 8. Get all the proteins that are located in the cell wall in the Cell Cycle Ontology.
Bio 9. Get all the core cell cycle protein and their AGI ids in A thaliana.
Bio 10. Get all the proteins that are involved in two specific diseases.
Bio 11. Get the proteins that are involved in many diseases.

——————————Ontological Queries——————————
Ont 1. Query the OBO Foundry: search on names and get their unique id's.
Ont 2. Get all the neighbor terms of a given term.
Ont 3. Get all the properties, like definition, synonyms, etc., of a given OBO term.
Ont 4. Get the names of the graphs in BioGateway.
Ont 5. Get a list of all the ontologies in the OBO Foundry.
Ont 6. Get the hierarchy to the root for a given term.

```
    term2_id: ssb:is_a ?common_parent_id.
    OPTIONAL {
      term1_id: ssb:is_a ?direct_child.
      term2_id: ssb:is_a ?direct_child.
      GRAPH <SSB> {
        ?direct_child ssb:is_a ?common_parent_id.
      }
    }
    ?common_parent_id rdfs:label ?common_parent.
  }
  FILTER(!bound(?direct_child))
}
```

UNION
GRAPH
ORDER BY
ASC()
DESC()
LIMIT
OFFSET
FILTER
Template

The new Semantic Systems Biology web site has been released (17.06.2008).

MAIN MENU

- Home
- BioGateway
  - Architecture
  - Resources
  - Tutorial
  - Querying
- News & Events
- About
- FAQ

Select a query in the drop-down box

The query editor

SPARQL - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.semantic-systems-biology.org/biogateway/querying

Sample queries:

Bio 10. Get all the proteins that are involved in two specific diseases.

Query:

```
# NAME      : get_disease_proteins
# PARAMETER: [Cc]ardiovascular: the first disease
# PARAMETER: [Dd]iabetes: the second disease
# FUNTION  : returns all the proteins that are involved in two
#            different given diseases

BASE   <http://www.semantic-systems-biology.org/>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX ssb:<http://www.semantic-systems-biology.org/SSB#>
SELECT distinct ?protein_id ?protein_name ?disease1 ?disease2
WHERE {
  GRAPH <uniprot_sprot> {
    ?protein_id ssb:disease ?disease1.
    ?protein_id ssb:disease ?disease2.
    ?protein_id ssb:mnemonic ?protein_name.
    FILTER regex(?disease1, '[Cc]ardiovascular').
    FILTER regex(?disease2, '[Dd]iabetes').
  }
}
LIMIT 100
```

Run
Prefixes
Comment
Uncomment
Optional
Indent
FROM
UNION
GRAPH
ORDER BY
ASC()
DESC()
LIMIT
OFFSET
FILTER
Template

Run Query
Reset

Placeholders to adapt the query

SPARQL - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.semantic-systems-biology.org/biogateway/querying

Mozilla Firefox

http://crunch.fvms.ugent.be:8891/sparql?query=%23 NAME %3A get_psoriasis_proteins%0A%23 PARAMETER%3

protein_name	disease_description	interacts_with	encoded_by
1C06_HUMAN	Genetic variation in HLA-C is associated with susceptibility to psoriasis 1 (PSORS1) [MIM%3A177900]. Psoriasis is a chronic inflammatory dermatosis that affects approximately 2% of the population. It is characterized by red, scaly skin lesions that are usually found on the scalp, elbows, and knees, and may be associated with severe arthritis. The lesions are caused by hyperproliferative keratinocytes and infiltration of inflammatory cells into the dermis and epidermis. The usual age of onset of psoriasis is between 15 and 30 years, although it can present at any age		
NALP1_HUMAN	Genetic variations in NLRP1 gene are associated with susceptibility to vitiligo-associated multiple autoimmune disease type 1 (VAMAS1) [MIM%3A606579]. Vitiligo is an autoimmune skin disorder associated with progressive skin depigmentation. Among patients with generalized vitiligo, there is an increased frequency of several other autoimmune and autoinflammatory diseases, particularly autoimmune thyroid disease, latent autoimmune diabetes in adults, rheumatoid arthritis, systemic lupus erythematosus, psoriasis and Addison disease	ASC_HUMAN	PYCARD
	Genetic variations in NLRP1 gene are associated with susceptibility to vitiligo-associated multiple autoimmune disease type 1 (VAMAS1) [MIM%3A606579]. Vitiligo is an autoimmune skin disorder		

The results appear in a separate window

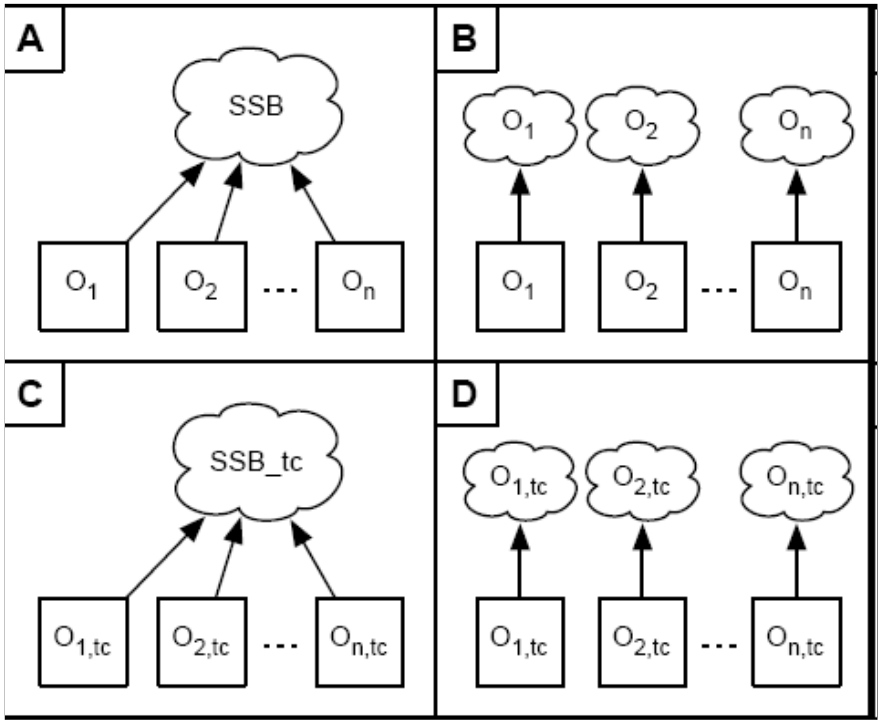
```
OPTIONAL {
  ?protein_id ssb:interactors_with ?interactor.
  ?interactor ssb:mnemonic ?interacts_with.
  ?interactor ssb:encoded_by ?encoded_by.
}
}
```

UNION

GRAPH

ORDER BY

BioGateway graphs



Each RDF-resource in BioGateway has a **URI** of this form:
http://www.semantic-systems-biology.org/SSB#resource_id

Each RDF-graph in BioGateway has a **URI** of this form:
http://www.semantic-systems-biology.org/graph_name

All the queries are explained in a tutorial*

1. Get the proteins with a specific function, location and process for all the annotated organisms.

```
# NAME: get_specific_proteins
# PARAMETER: GO_0005216: ion channel activity
# PARAMETER: GO_0005764: lysosome
# PARAMETER: GO_0006811: ion transport
# FUNCTION: returns all the proteins with the same function,
# process and location and the organism in which
# they can be found

BASE <http://www.semantic-systems-biology.org/>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX ssb:<http://www.semantic-systems-biology.org/SSB#>
SELECT ?organism ?protein ?protein_id
WHERE {
  GRAPH ?organism {
    ?protein_id ssb:has_function ssb:GO_0005216.
    ?protein_id ssb:located_in ssb:GO_0005764.
    ?protein_id ssb:participates_in ssb:GO_0006811.
    ?protein_id rdfs:label ?protein.
  }
  FILTER(?organism != <SSB> && ?organism != <GOA>).
}
```

Click here to select this query in the drop-down box on the query-page and edit it

Click here to see the results

For every query the name, the parameters and the function are indicated at the top.

The parameters are indicated in red.

* http://www.semantic-systems-biology.org/biogateway/tutorial

Resources - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.semantic-systems-biology.org/biogateway/resources

Individual rdf files:

- 1 UniProt - Swiss-Prot file, the SwissProt secton of UniProt KB of proteins (see integrated graphs)
- 1 NCBI file withthe taxonomy of organisms
- 1 Metaonto file with information about OBO Foundry ontologies
- 2 Metarel files with relation type properties
- 5 CCO files with integrated information about cell cycle proteins
- 44 OBO Foundry files with diverse biomedical information
- 51 Transitive Closure files to enhance query abilities
- 893 GOA files with GO annotations

NCBI

Graph name	Prefix	Ontology Name	About	FTP
ncbi	NCBI	NCBI Taxonomy	Biological species	D

Metaonto

Graph name	Prefix	Ontology Name	About	FTP
metaonto	METAONTO	Metaonto	ontologies	D

Metarel

Graph name	Prefix	Ontology Name	About	FTP
biometarel	METAREL	Biometarel	relations	D
biorel	rel_type	Biorel	relations	D

CCO

Graph name	Prefix	Ontology Name	About	FTP
cco_A_thaliana	CCO	Cell Cycle Ontology (A.Thaliana)	cell cycle	D
cco_H_sapiens	CCO	Cell Cycle Ontology (H. Sapiens)	cell cycle	D

998 RDF-files can be downloaded from the Resources page

The graph names can be used to query or combine individual graphs for quicker answers or more specific information

The neighbourhood of the human protein 1443F in the RDF-graph

term_as_child	outward_arrow	head_name	tail_name	inward_arrow	term_as_parent
1433F_HUMAN	participates in	intracellular protein transport			
1433F_HUMAN	participates in	glucocorticoid catabolic process			
1433F_HUMAN	participates in	positive regulation of transcription			
1433F_HUMAN	participates in	regulation of synaptic plasticity			
1433F_HUMAN	participates in	glucocorticoid receptor signaling pathway			
1433F_HUMAN	participates in	regulation of neuron differentiation			
1433F_HUMAN	participates in	negative regulation of dendrite morphogenesis			
1433F_HUMAN	is located in	cytoplasm			
1433F_HUMAN	has function	protein binding			
1433F_HUMAN	has function	transcription activator activity			
1433F_HUMAN	has function	actin binding			
1433F_HUMAN	has function	insulin-like growth factor receptor binding			
1433F_HUMAN	has function	protein domain specific binding			
1433F_HUMAN	has function	glucocorticoid receptor binding			
1433F_HUMAN	has source	Homo sapiens			
1433F_HUMAN	interacts with	PARD3_HUMAN			
1433F_HUMAN	interacts with	PFTK1_HUMAN			
1433F_HUMAN	interacts with	RAF1_HUMAN			
1433F_HUMAN	interacts with	GREM1_HUMAN			
1433F_HUMAN	interacts with	MARK4_HUMAN			
1433F_HUMAN	interacts with	PAR6A_HUMAN			
1433F_HUMAN	interacts with	PAR6B_HUMAN			
1433F_HUMAN	interacts with	KPCI_HUMAN			
			PARD3_HUMAN	interacts with	1433F_HUMAN
			PFTK1_HUMAN	interacts with	1433F_HUMAN
			RAF1_HUMAN	interacts with	1433F_HUMAN
			GREM1_HUMAN	interacts with	1433F_HUMAN
			PAR6A_HUMAN	interacts with	1433F_HUMAN
			PAR6B_HUMAN	interacts with	1433F_HUMAN
			KPCI_HUMAN	interacts with	1433F_HUMAN
			ADA22_HUMAN	interacts with	1433F_HUMAN
			HNRPD_HUMAN	interacts with	1433F_HUMAN

The resulting triples (arrows) are represented as a small grammatical sentence: subject, predicate, object.

Outgoing arrows

Incoming arrows

Mozilla Firefox

File Edit View History Bo

http://www.semantic-systems-biology.org/sparql-viewer/

Google

LIMIT 40 Endpoint http://crunch.fvms.ugent.be:8891/sparql SPARQL Browser - 0.0.9

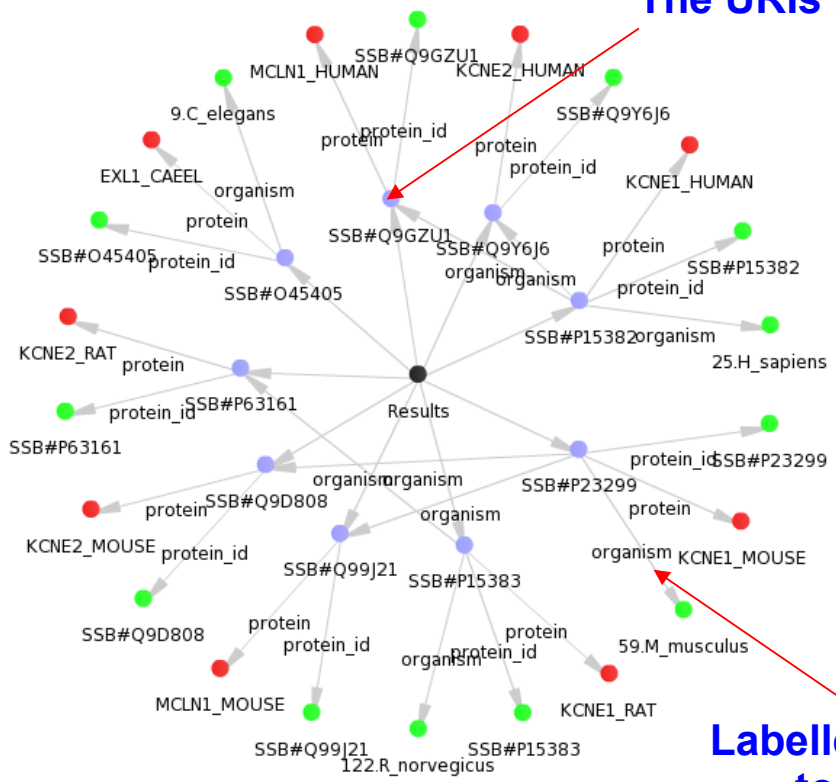
Query Prefix

```
SELECT ?protein ?protein_id ?organism
WHERE {
 GRAPH ?organism {
  ?protein_id ssb:has_function    ssb:GO_0005216.
  ?protein_id ssb:located_in      ssb:GO_0005764.
  ?protein_id ssb:participates_in ssb:GO_0006811.
  ?protein_id rdfs:label ?protein.
 }
```

Graph XML Browse

protein	protein_id	organism
EXL1_CAEEL	SSB#O45405	9.C_elegans
KCNE1_HUMAN	SSB#P15382	25.H_sapiens
KCNE1_RAT	SSB#P15383	122.R_norvegicus
KCNE1_MOUSE	SSB#P23299	59.M_musculus
KCNE2_RAT	SSB#P63161	122.R_norvegicus
MCLN1_MOUSE	SSB#Q99J21	59.M_musculus
KCNE2_MOUSE	SSB#Q9D808	59.M_musculus
MCLN1_HUMAN	SSB#Q9GZU1	25.H_sapiens
KCNE2_HUMAN	SSB#Q9Y6J6	25.H_sapiens

The URIs in blue.



The result: 9 proteins

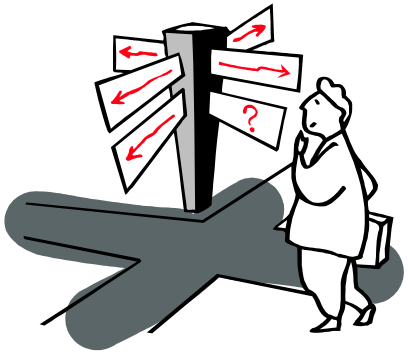
Labelled arrows to extra information

Degrees of Separation Scaling Link Length AutoFit

Results

- BioGateway: RDF store for Biosciences (prototype!)
- Data integration pipeline: BioGateway
- Queries and knowledge sources and system design go hand-in-hand (user interaction)
- Enables building relatively “complex” questions
- Existing integration obstacles due to:
  - diversity of data formats
  - lack of formalization approaches
- Semantic Web technologies add a new dimension of knowledge integration to Systems Biology

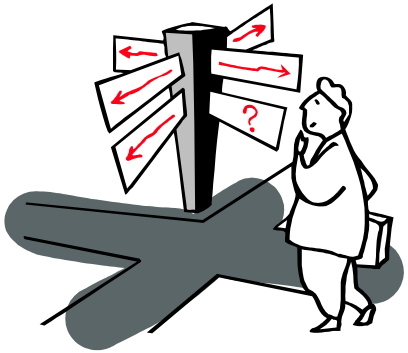
Contents



Conclusions

- Categories:
  - Importance of computationally representing biological knowledge
  - Exploitation of such knowledge
- Both gave rise to a new (complementary) form of Systems Biology: Semantic Systems Biology approach
  - Data integration
  - Holistic (systemic) approach
  - Data exploitation (*e.g.* querying, reasoning)
  - Ultimately, create new hypothesis
- Semantic Web technologies *do* have the potential to provide a sound framework for biological data integration

Contents



Future prospects

- **Linked Data**

SPARQLGraph

Visual Query Builder for Biological RDF Databases

http://sparqlgraph.i-med.ac.at