Towards a Semantic Systems Biology: Biological Knowledge Management Using Semantic Web Technologies

Erick Antezana

<erick.antezana@gmail.com>

BioHackathon -Tokyo, 8 Sept. 2010

Contents

1. Introduction
2. The Cell Cycle Ontology
3. BioGateway
4. Concluding remarks
5. Future prospects

Contents

1. Introduction
   - Motivation
   - Background
2. The Cell Cycle Ontology
3. BioGateway
4. Concluding remarks
5. Future prospects

Motivation

- Amount of data generated in the biological experiments continues to grow exponentially
- Shortage of proper approaches or tools for analysing this data has created a gap between raw data and knowledge
- 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 the data

Question

What is the potential of the Semantic Web technologies for biological knowledge management in the context of a Systems Biology approach?

Strategy

- Steps:
  - Problem definition: test bed case (cell cycle)
  - Data scaffold elements: standards, terminologies and ontologies
  - Development of tools
  - Data integration and exploitation
  - Beyond cell cycle: all processes in the Gene Ontology

Background

- Data *vs.* Knowledge
  - Information?

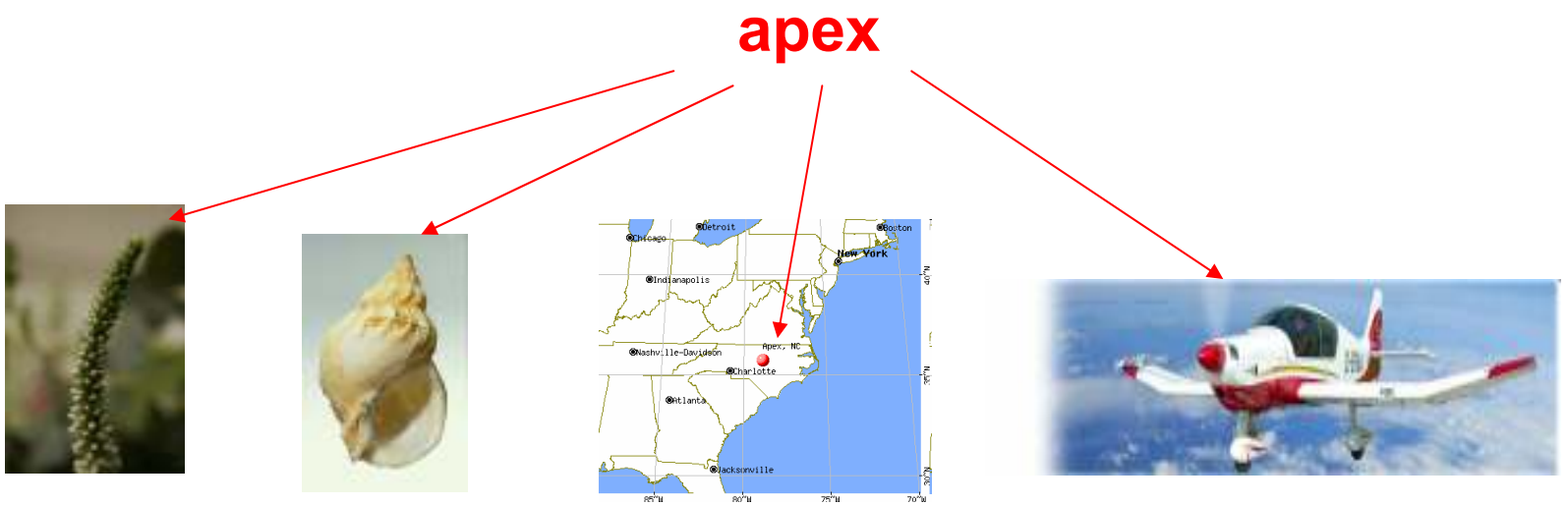
- Knowledge Management
  - Capturing, structuring, retaining and reusing
  - *Data integration (e.g. identity crisis)*
    - *Warehouse*
    - *Data federation*

Knowledge Representation (KR)

- A formalism should
  - represent real world entities (in/tangible)
  - enable efficient organisation and processing of information
  - enable shareability
- Components
  - language
  - modelling principles
- Interoperability
  - syntax (symbols + rules)
  - semantics (meaning)

Same term, different concepts

- “apex”
  - The apical meristem or its remnant on a flower
  - Tip of the spire of the shell of a gastropod
  - A town in North Carolina
  - A company building airplanes
  - …



Ontology

- What is it? (too many definitions)
  - Most cited definition: *"A formal specification of a conceptualisation"* (Gruber, 1995)
- Computer scientist
  - *A* specific ***artefact*** designed with the purpose of expressing the ***intended meaning*** of a (shared) ***vocabulary***
  - Bio-ontologist: *"A controlled vocabulary of biological terms and their relations"* (e.g. GO, RO, PO).
- Why do we need them?
  - Share and reuse information (common terminology)
  - Data integration
  - Other applications (e.g. analysis, annotation)
- Multidisciplinary teams: philosophers, computer scientists, domain experts (biologists), …

Semantic Web

- *"Next generation of the current web"*
- Goal: machine understandable content
- Keyword search will get obsolete
  - Complex query formulation
- 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, SPARL	http://neurocommons.org/
RDFScape	systems biology, cytoscape, reasoning	RDF, SPARL	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, SPARL	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/
Linked Life Data	pathways, interactions	OWL	http://www.linkedlifedata.com

Contents

1. Introduction
2. The Cell Cycle Ontology
3. BioGateway
4. Concluding remarks
5. Future prospects

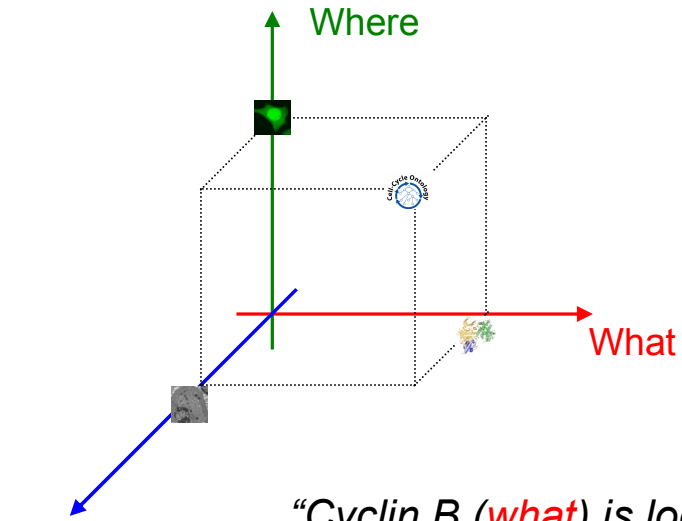
Contents

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
3. BioGateway
4. Concluding remarks
5. Future prospects

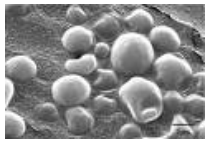
The Cell Cycle Ontology in a nutshell

- 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
- Issues: *synonymy*, *polysemy*



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

ORGANISMS:

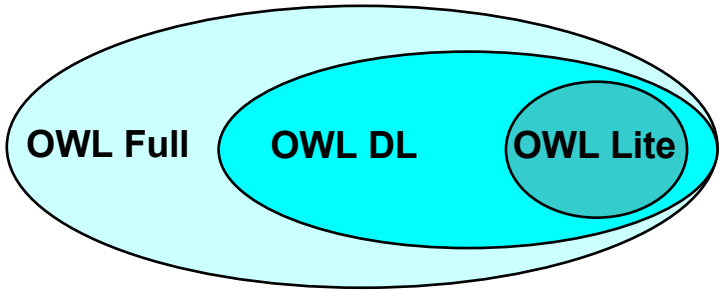
Users:

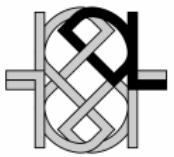
- Molecular biologist
- Bioinformatician /Computational Systems Biologist
- General audience

Antezana E. et al. Lect. Notes Bioinformatics, 2006

Knowledge representation in CCO

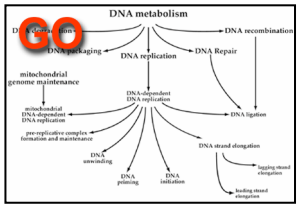
- Why OBO?
  - "Human readable"
  - Standard
  - Tools (e.g. OBOEdit)
  - http://obo.sourceforge.net
- Why OWL?
  - Web Ontology Language
  - "Computer readable"
  - Reasoning capabilities vs. computational cost ratio
  - Formal foundation (Description Logics)
  - Tools (e.g. Protégé)
- OBO2OWL mapping
  - ONTO-PERL (*Antezana E. et al. Bioinformatics 2008*)





CCO sources

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

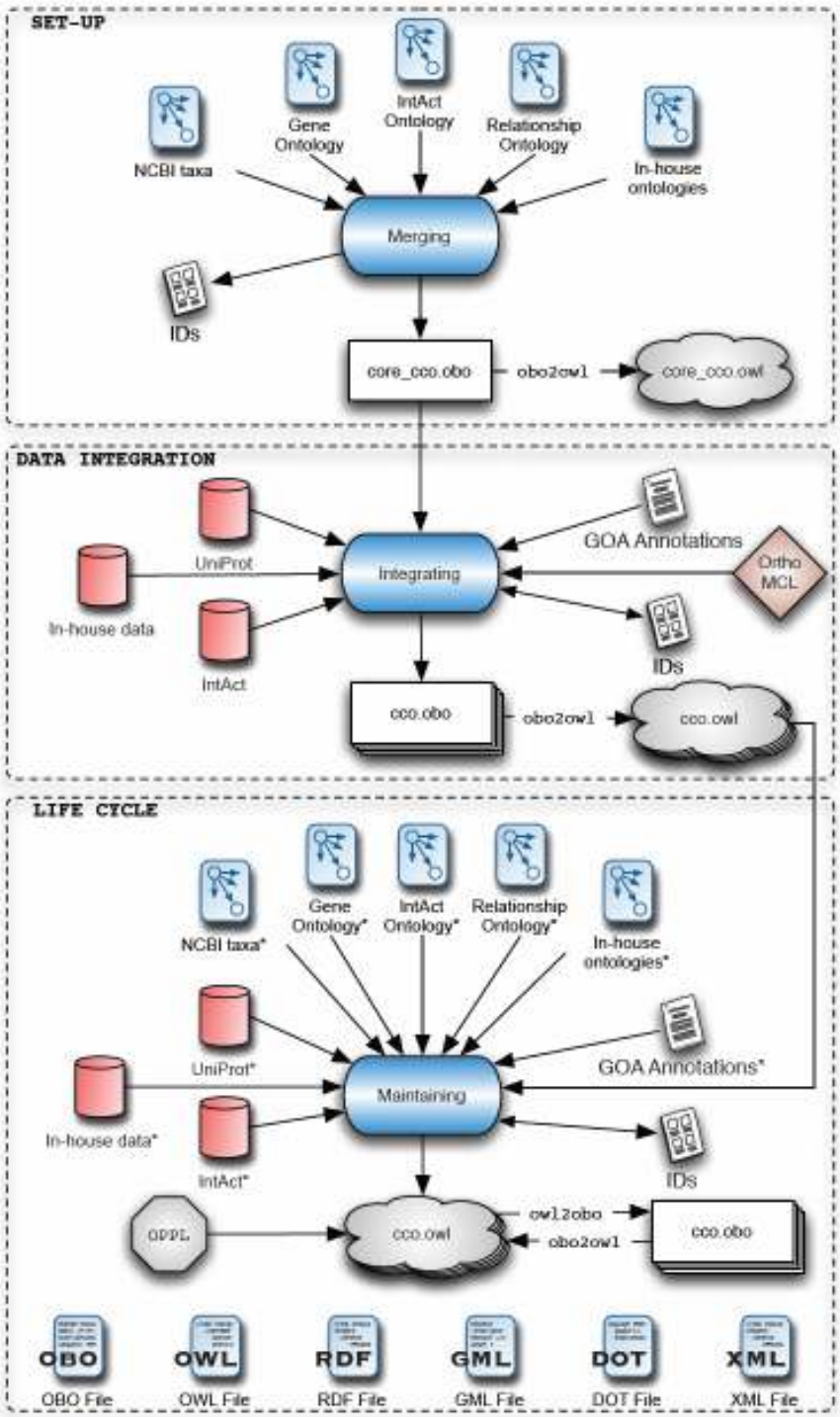


The Open Biomedical Ontologies

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



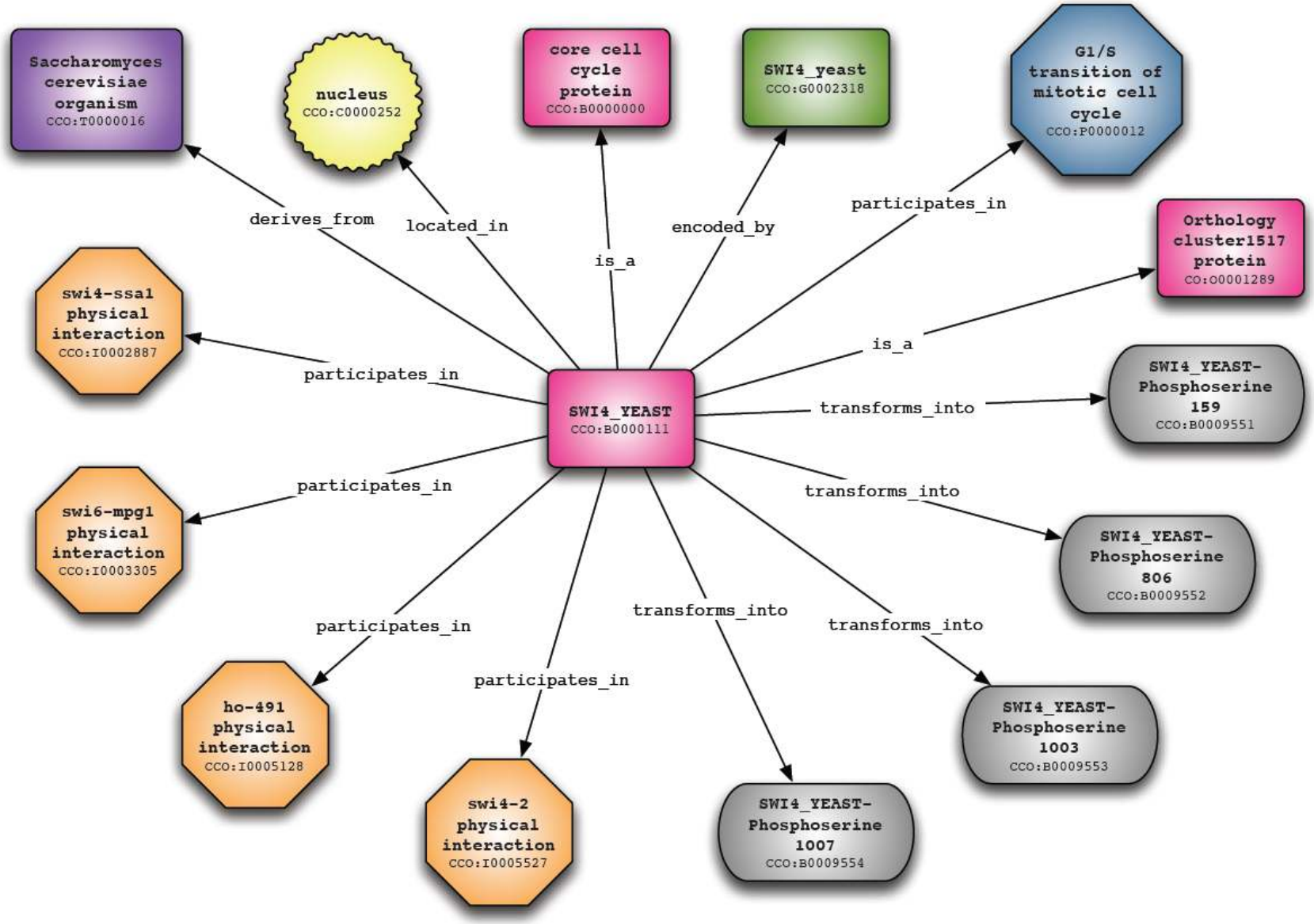
CCO Pipeline

- ontology integration (ONTO-PERL)
- format mapping

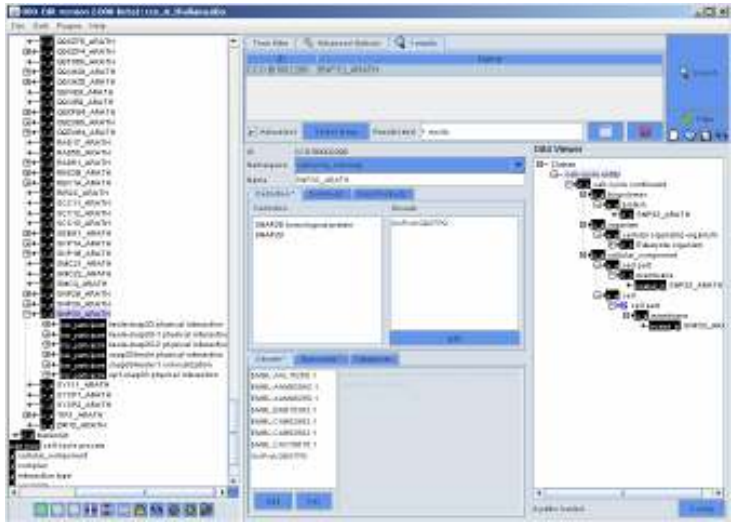
- data integration
- data annotation
- consistency checking

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

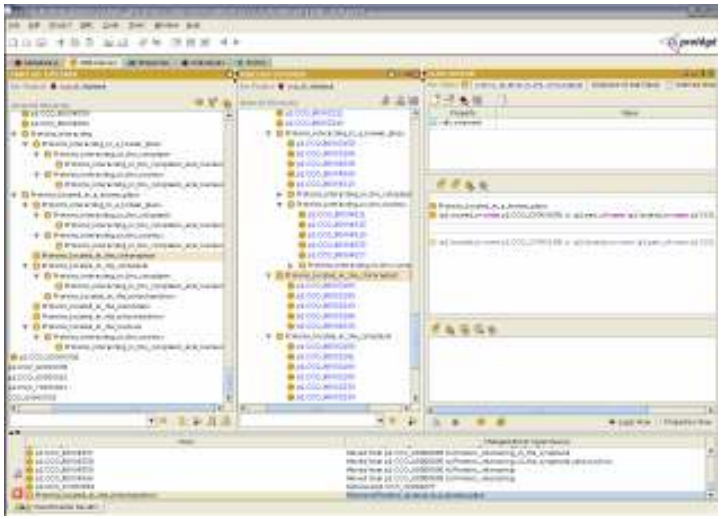
Sample knowledge in CCO



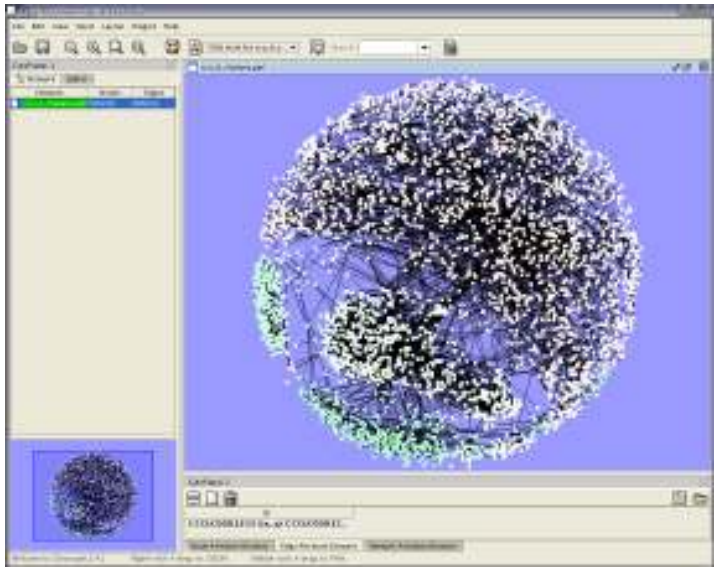
Exploring CCO (1/2)



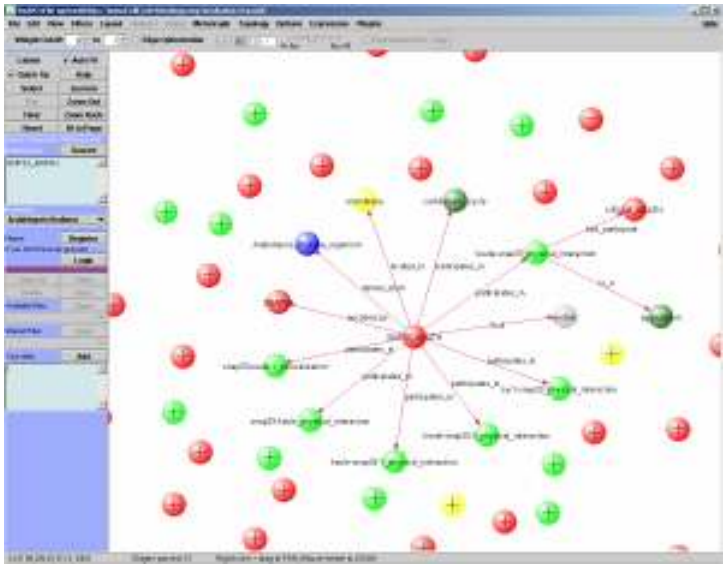
OBO-Edit



Protégé



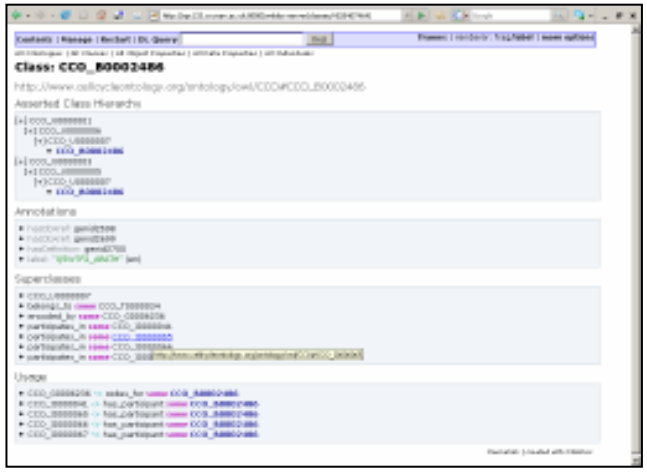
Cytoscape



visANT

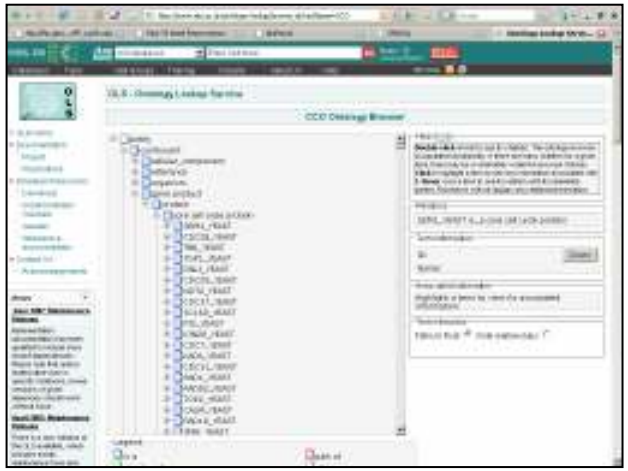
Exploring CCO (2/2)

CCO website (SPARQL)



OWLDoc server

BioPortal



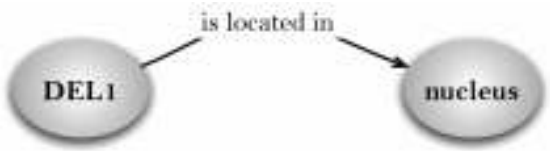
Ontology Look up Service

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) :

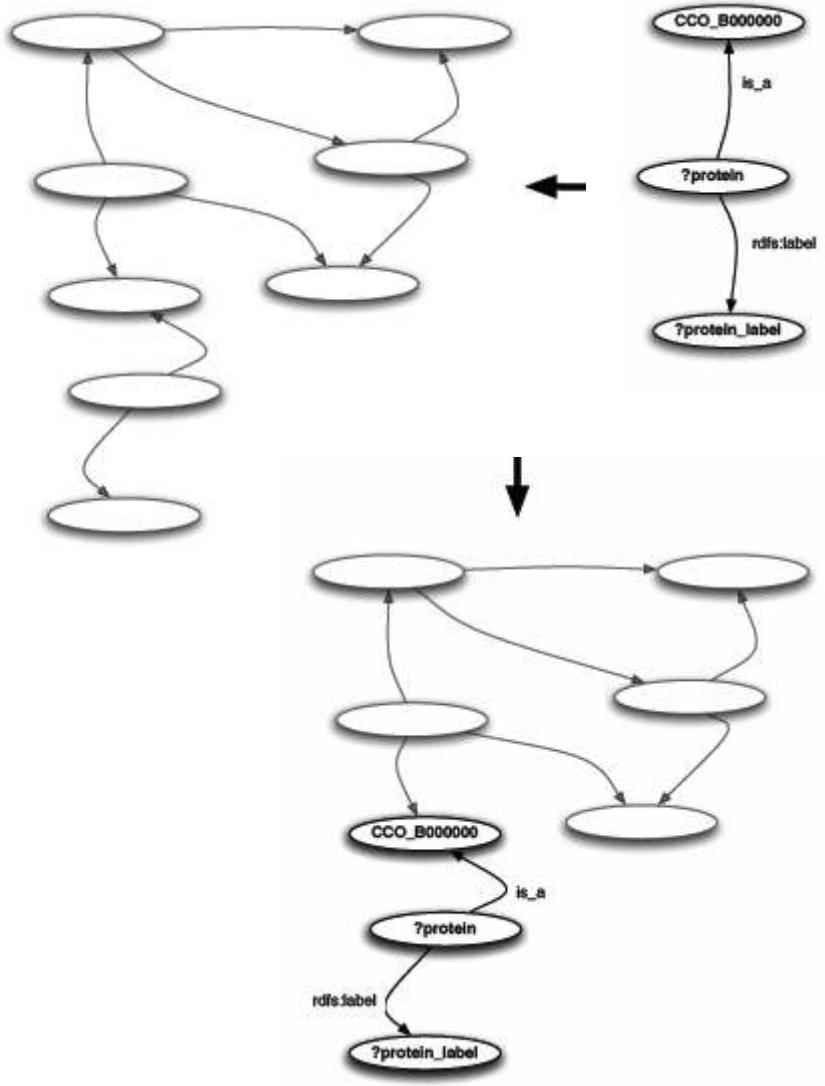
- Subject corresponds to the main entity that needs to be described.
- Predicate denotes a quality or aspect of the relation between the Subject and Object.

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



SPARQL*

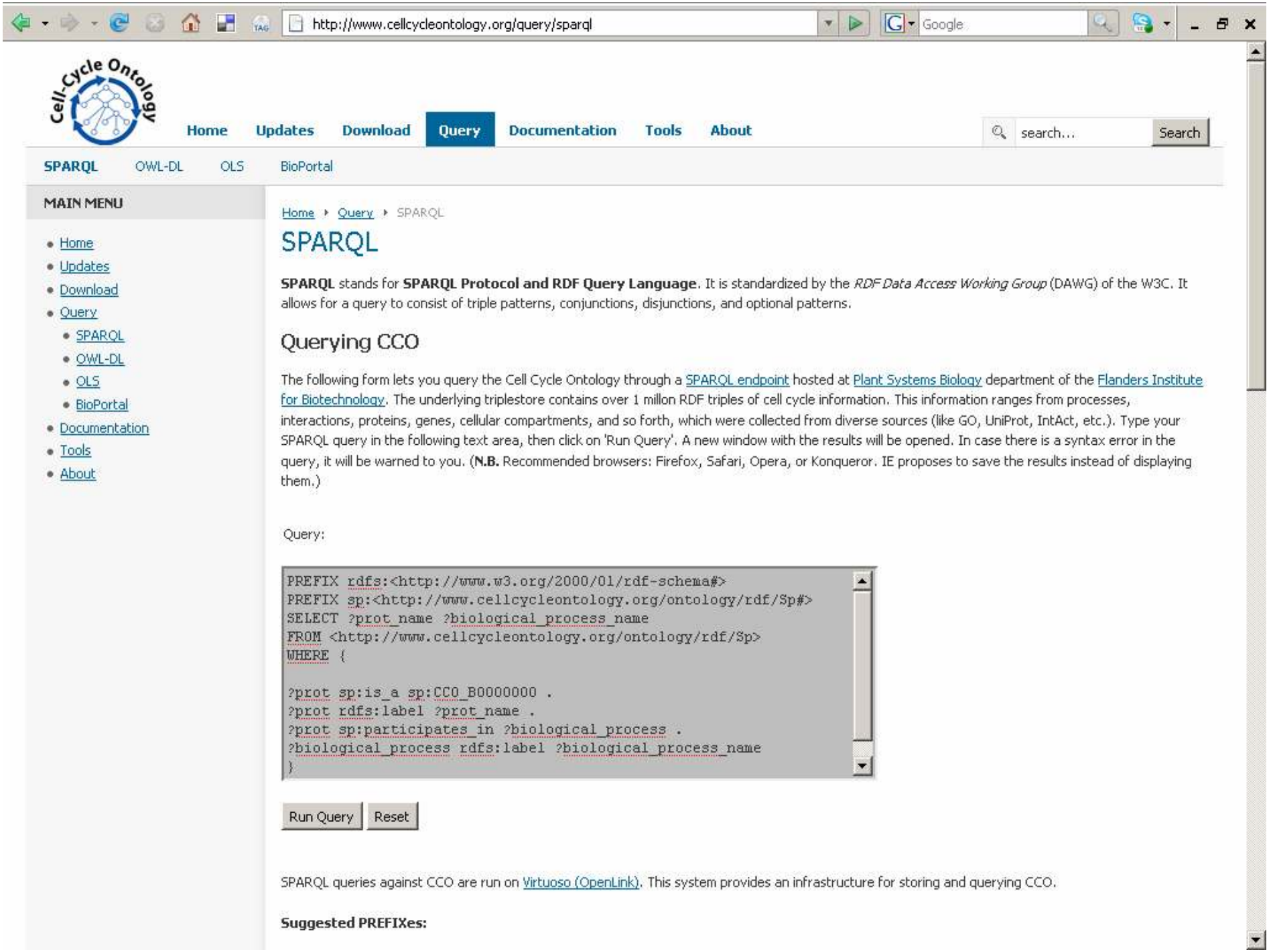
- Query RDF models (graphs)
- Powerful, flexible
- Its syntax is similar to the one of SQL.
- Virtuoso Open Server
- Example (matching two triples):



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

?protein **rdfs:label** ?protein_label

* http://www.w3.org/TR/rdf-sparql-query/


http://www.cellcycleontology.org/query/sparql
Google
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 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
Google
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
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
d browsers: Firefo
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:

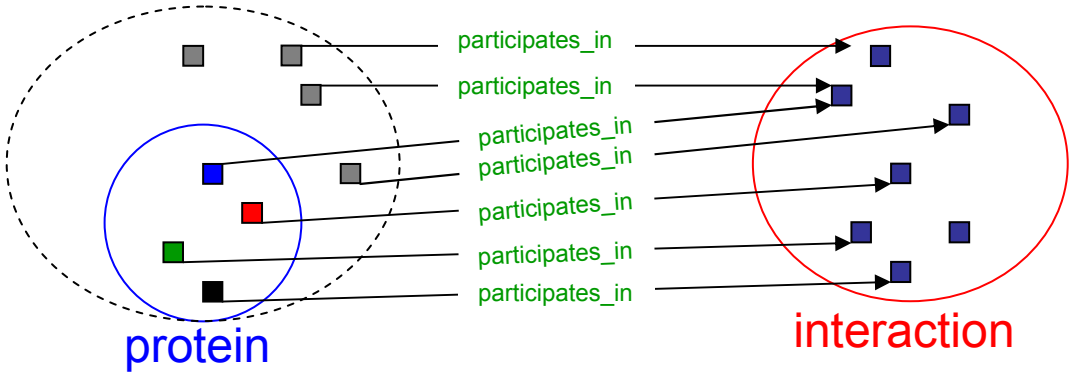
prot_name	biological_process_name
UBC11_SCHPO	G2%2FM transition of mitotic cell cycle
UBC11_SCHPO	cell cycle
UBC11_SCHPO	mitosis
UBC11_SCHPO	mitotic metaphase%2Fanaphase transition
UBC11_SCHPO	regulation of mitotic cell cycle
UBC11_SCHPO	cyclin catabolic process
SRW1_SCHPO	cell cycle
SRW1_SCHPO	cyclin catabolic process
SRW1_SCHPO	activation of anaphase-promoting complex during mitotic cell cycle
SRW1_SCHPO	cell cycle arrest in response to nitrogen starvation
SRW1_SCHPO	negative regulation of cyclin-dependent protein kinase activity
DYHC_SCHPO	dhc1-peg1-1 physical interaction
DYHC_SCHPO	synapsis
DYHC_SCHPO	meiotic recombination
DYHC_SCHPO	horsetail nuclear movement
ORB6_SCHPO	cell morphogenesis checkpoint
ORB6_SCHPO	regulation of cell cycle
DED1_SCHPO	G2%2FM transition of mitotic cell cycle

Reasoning over CCO

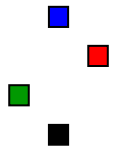
- OWL-DL: balance tractability with expressivity
- Consistency checking: no contradictory facts
- Classification: implicit2explicit knowledge
- Tools: Protégé, Reasoners (e.g. RACER, Pellet)

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

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

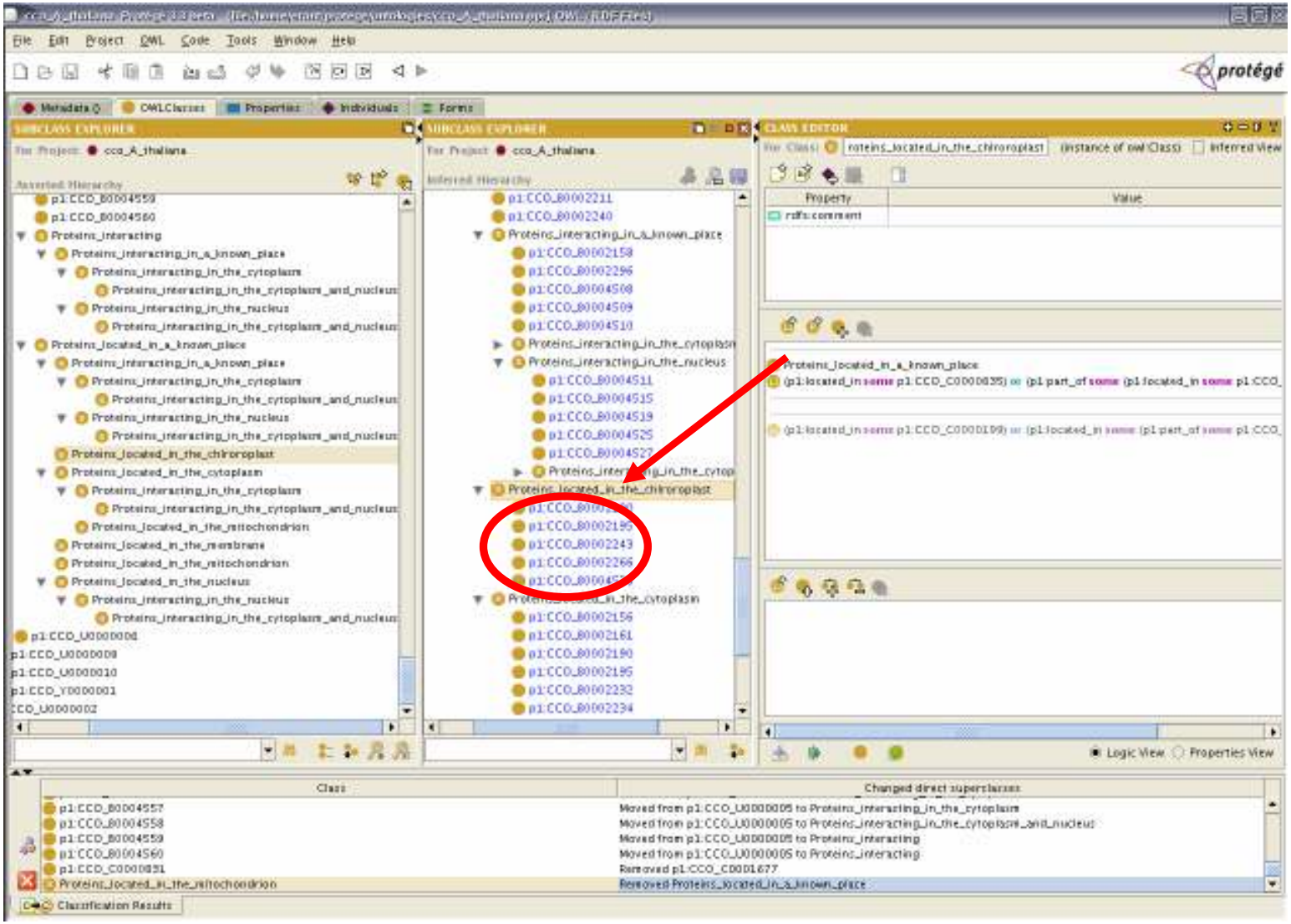


Answer:



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 were detected)
  - simple biological hypothesis generation
- Data integration based on trade-offs (e.g. multiple inheritance)
- Performance issues (technology limitations)

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
4. Concluding remarks
5. 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 GOA

- Support a Semantic Systems Biology approach

Antezana et al. BMC Bioinformatics, 2009

Systems Biology

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

Systems Biology (cont)

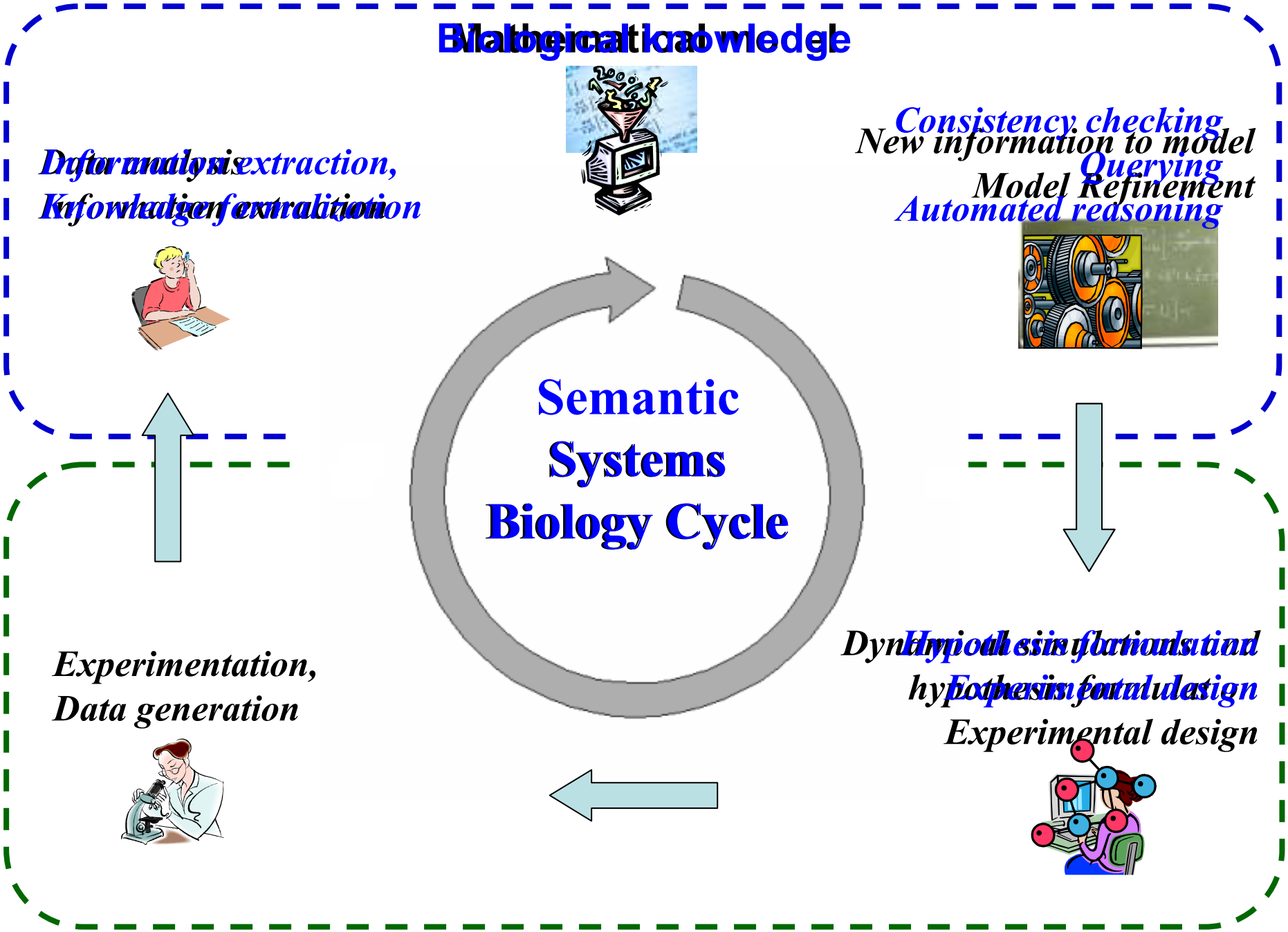
- “A system (and its properties) cannot be described in terms of their terms in isolation; its comprehension emerges when studied globally”
- Systems Biology = Approach to study biological systems.
- Arbitrary borders
- A system within a system

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, Briefings 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

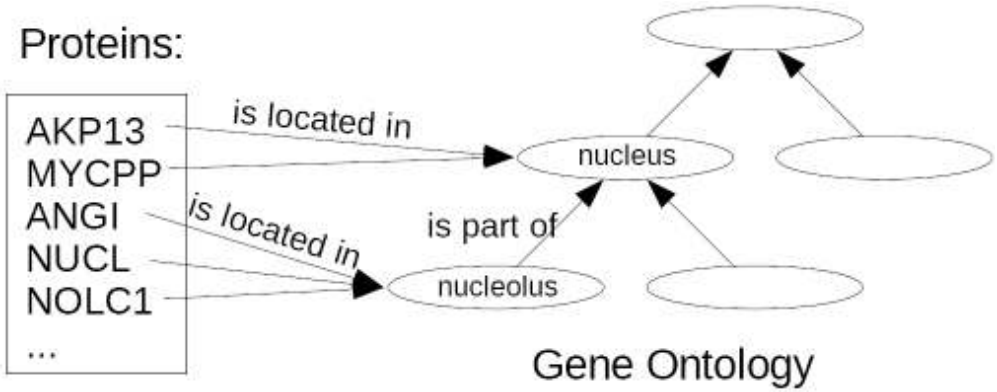

Biological knowledge
Data analysis
Information extraction,
Knowledge formalization
Consistency checking
New information to model
Querying
Model Refinement
Automated reasoning
Semantic
Systems
Biology Cycle
Experimentation,
Data generation
Hypothesis formulation
Experimental design

BioGateway: a tool to support Semantic Systems Biology

- Automatic data integration pipeline (~8 months)
- **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)*

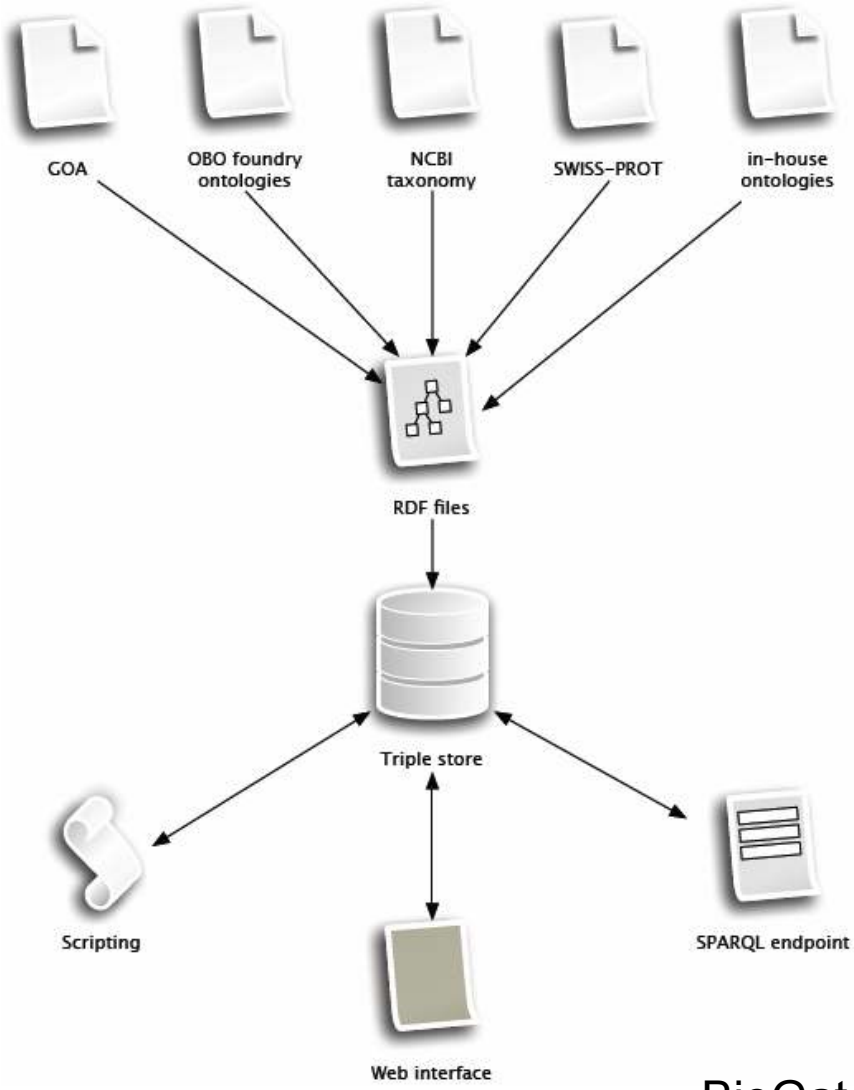
Transitive closure graphs

- If ***A part_of B***, and ***B part_of C***, then ***A part_of C*** is also added to the graph.
- 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., Antezana E. et al. ICBO, 2009

BioGateway pipeline

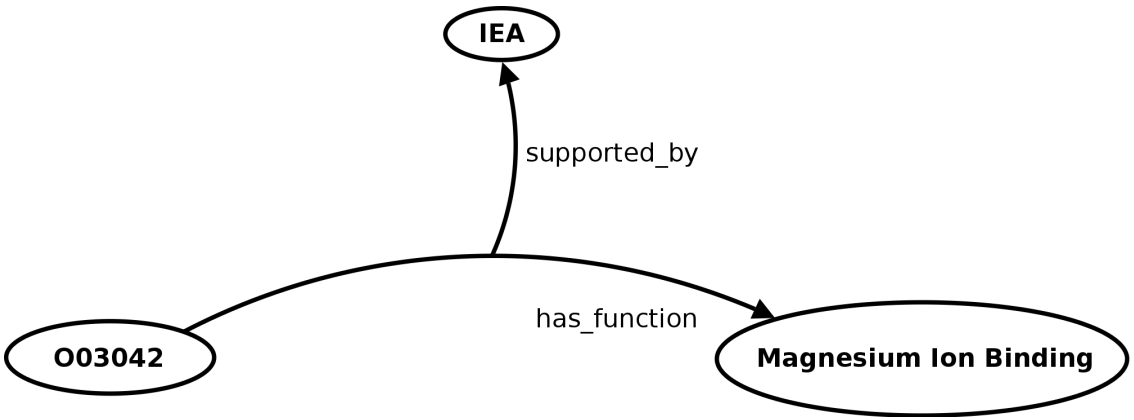


- **1 Swiss-Prot** file, the section of UniProt KB of proteins
- **1 NCBI** file with the taxonomy of organisms
- **1 Metaonto** file with information about OBO Foundry ontologies
- **2 Metarel** files with relation type properties
- **5 CCO** files with integrated information about cell cycle proteins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

BioGateway holds ~175 million RDF triples!!!

Sample RDF-ication: GOA

UniProtKB O03042 O03042 GO:0000287 GOA:spkw|GO_REF:0000004 IEA

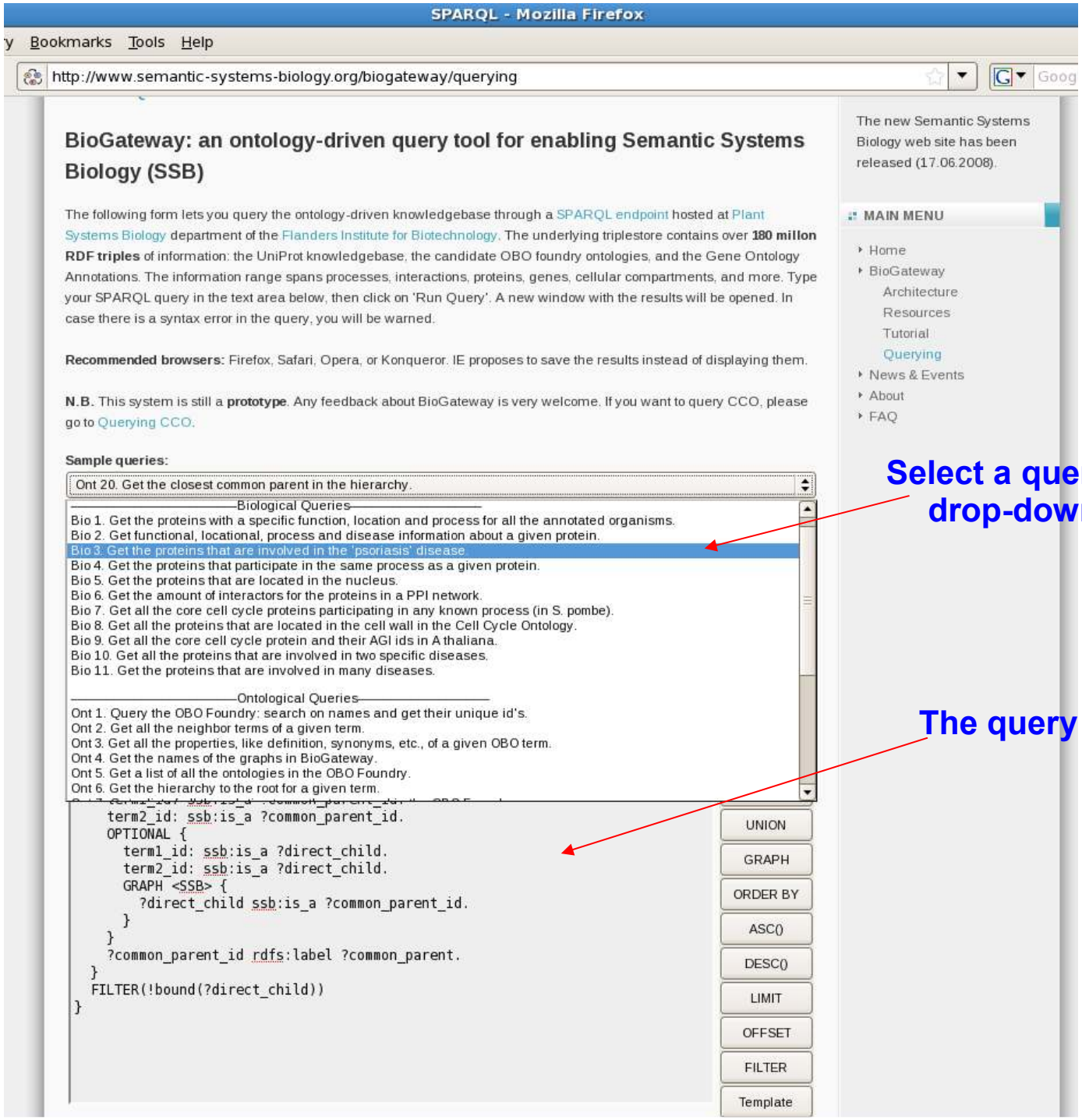


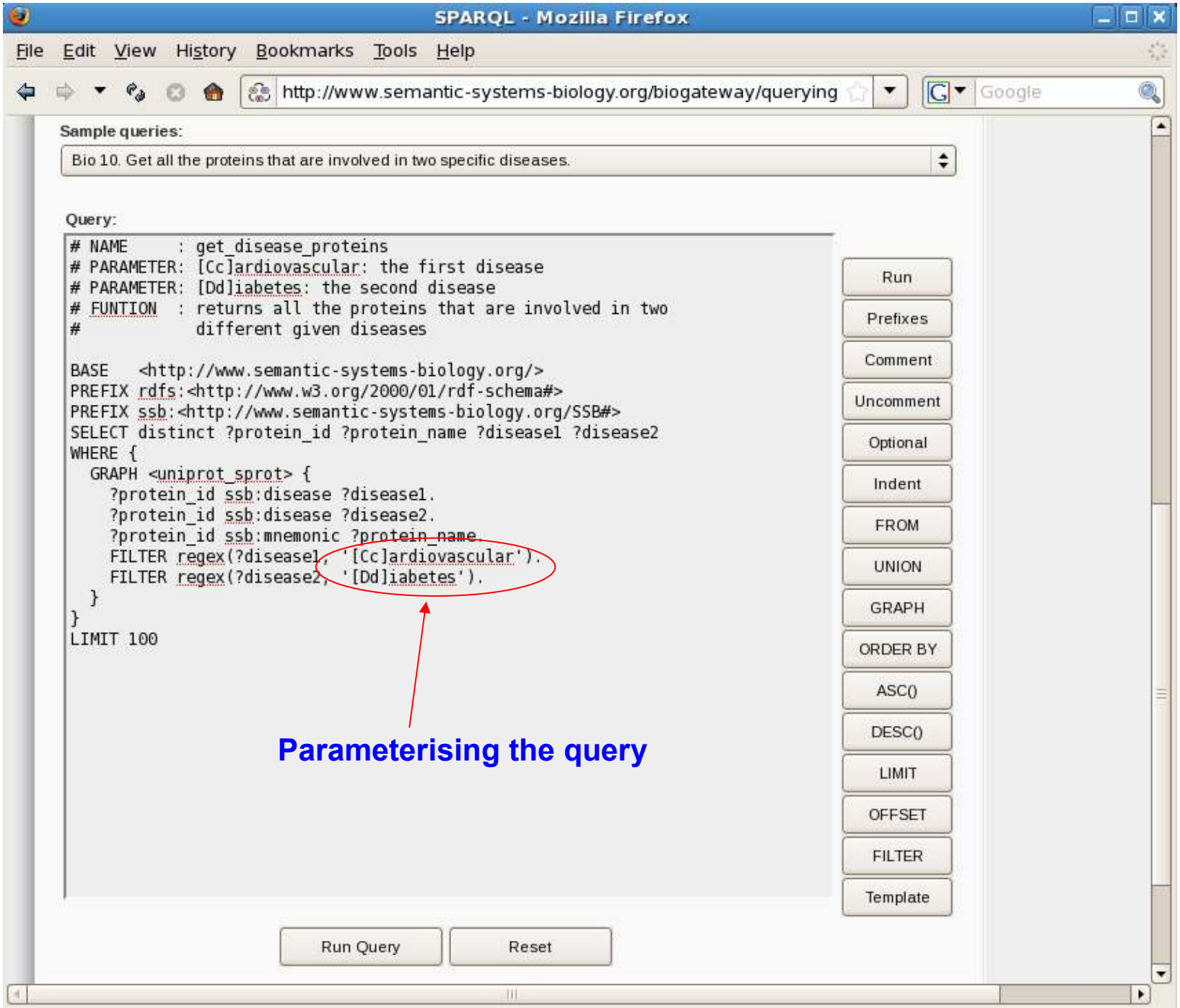
- **Protein:** *Ribulose bisphosphate carboxylase large chain* (O03042)
- **GO term (MF):** Magnesium Ion Binding (GO:0000287)
- Therefore, O03042 has the molecular function of binding magnesium ion.
- This fact is supported by **IEA**, that is, **I**nferred from **E**lectronic **A**nnotation."

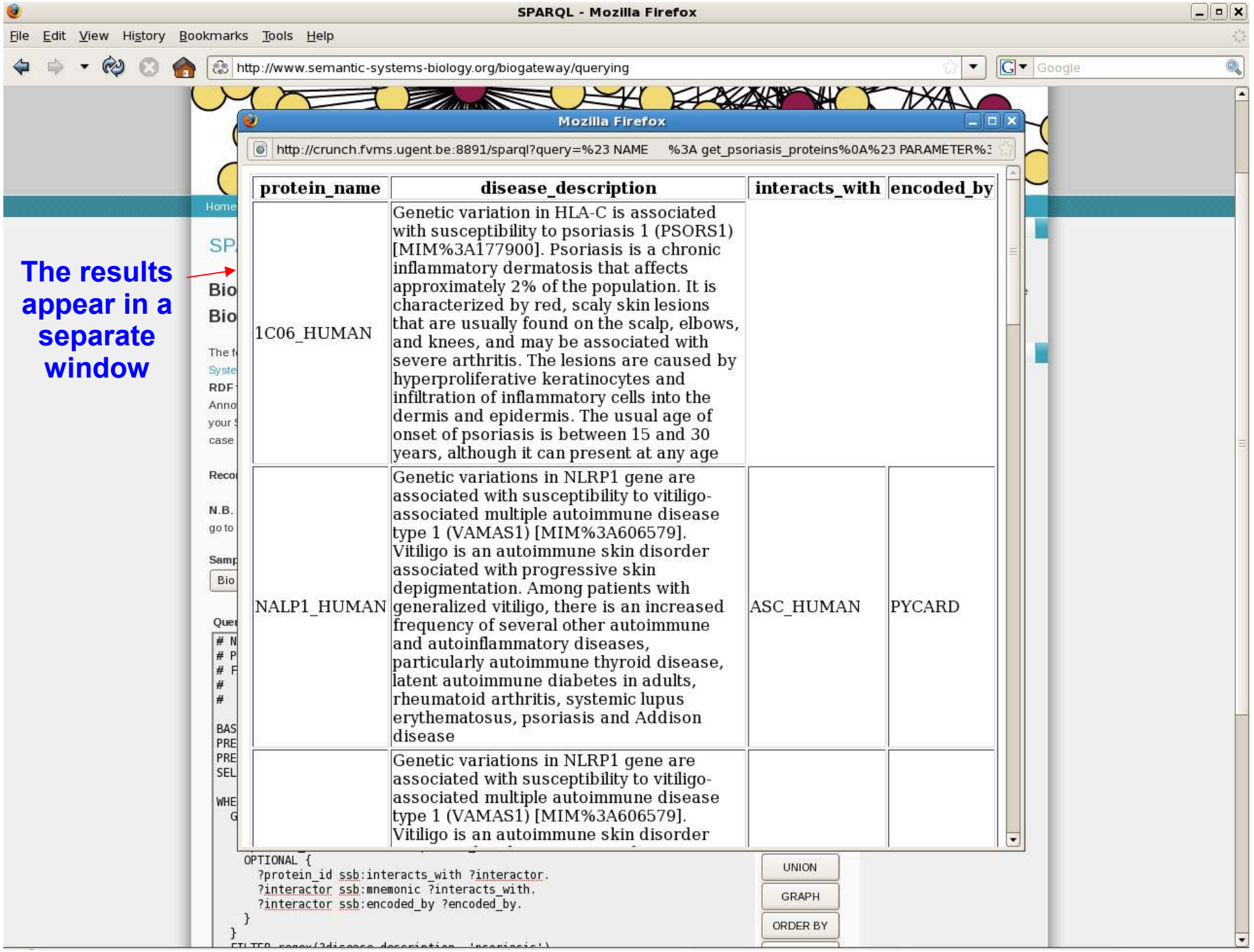
A library of queries*

- The drop-down box contains 35 queries:
  - 14 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 neighborhood, 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,...
    - …

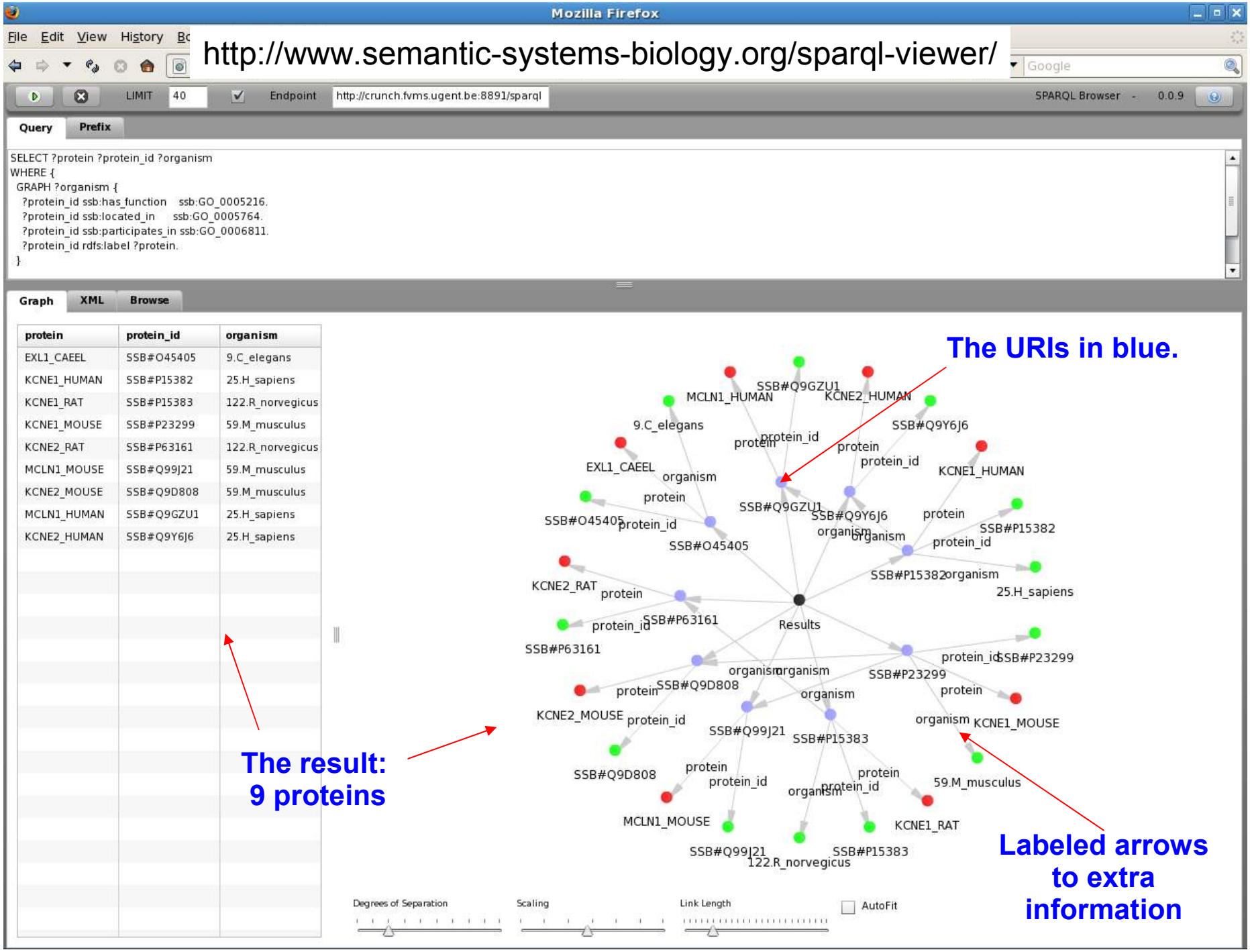
* http://www.semantic-systems-biology.org/biogateway/querying


SPARQL - Mozilla Firefox
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
Google
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
Parameterising the query



The results appear in a separate window


Mozilla Firefox
http://www.semantic-systems-biology.org/sparql-viewer/
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
Labeled arrows
to extra
information
Results
Degrees of Separation
Scaling
Link Length
AutoFit

Conclusions / 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

1. Introduction
2. The Cell Cycle Ontology
3. BioGateway
4. Concluding remarks
5. Future prospects

Conclusions and prospects

- Categories:
  - Way 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

1. Introduction
2. The Cell Cycle Ontology
3. BioGateway
4. Concluding remarks
5. Future prospects

Future prospects

- Temporal representation
- Capture **non-crisp** knowledge (e.g. protein sar1 is *usually* located in the nuclear membrane)
- Integration of multimedia (images, videos)
- “Deep down” integration of data into BioGateway (like in CCO)
- …

Acknowledgements

- Martin Kuiper (NTNU, NO)
- Vladimir Mironov (NTNU, NO)
- Mikel Egaña (U Manchester, UK / U Murcia, ES)
- Robert Stevens (U Manchester, UK)
- BioHealth Group (U Manchester, UK)
- Ward Blondé (U Ghent, BE)
- Bernard De Baets (U Ghent, BE)
- Alistair Rutherford (UK)
- Alan Ruttenberg (Science Commons, US)
- Ontology and Semantic Web community
- Users, (former/new) colleagues and friends
- …

SEMANTIC SYSTEMS BIOLOGY

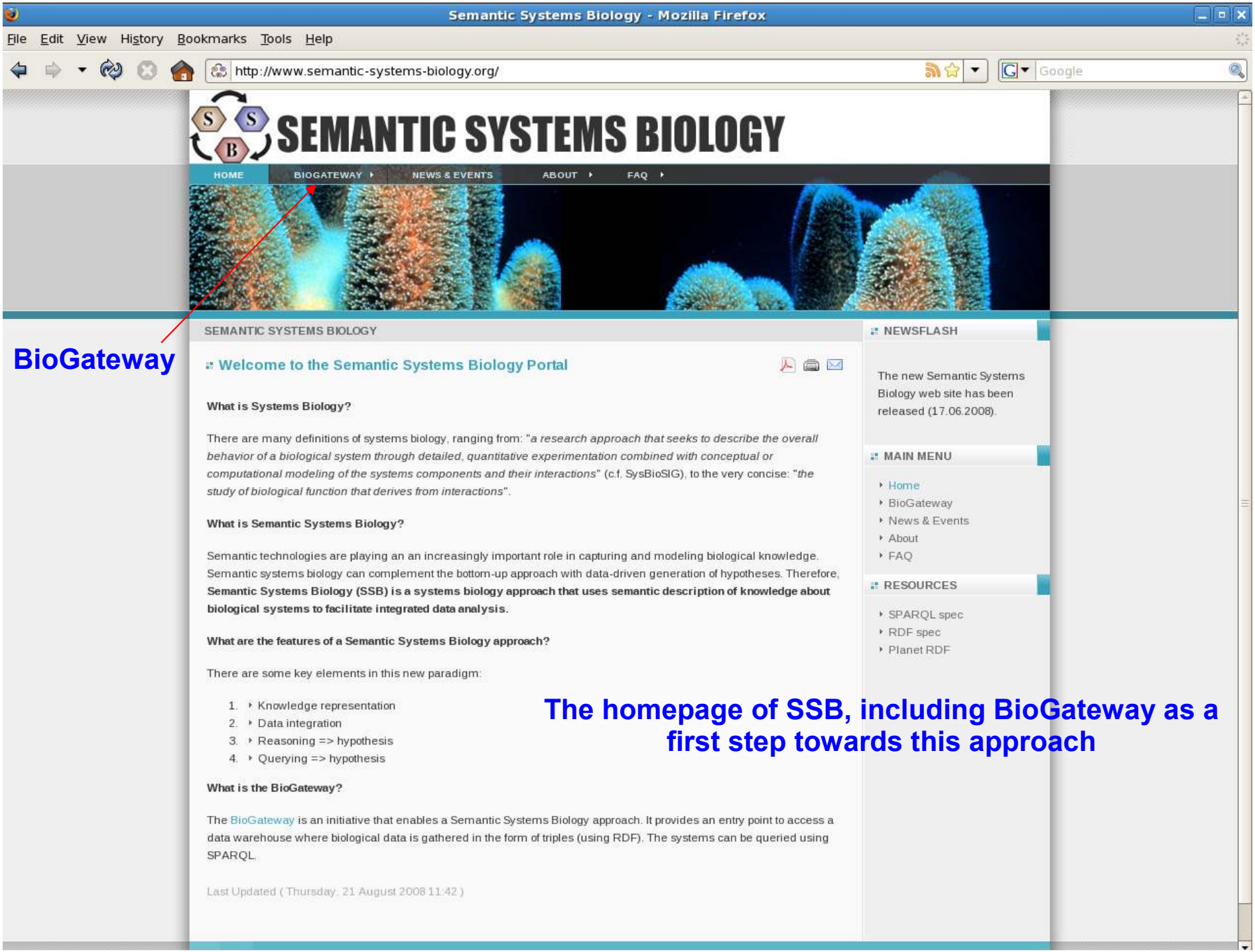
http://www.semantic-systems-biology.org

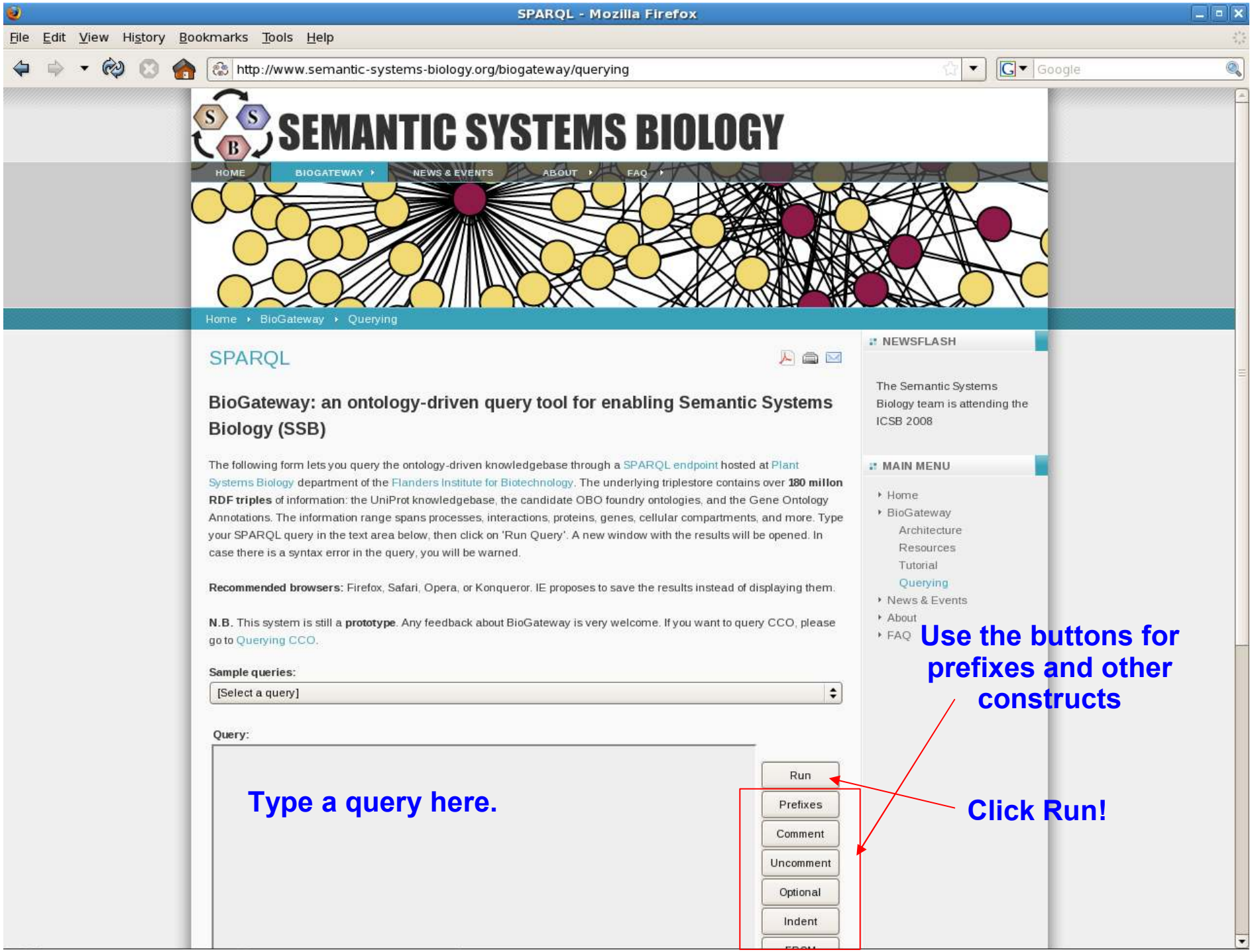
RESEARCH

Sixth Framework Programme

Bayer CropScience

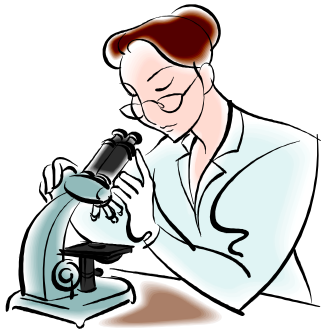
Extra slides


Semantic Systems Biology - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.semantic-systems-biology.org/
Google
SEMANTIC SYSTEMS BIOLOGY
HOME
BIOGATEWAY
NEWS & EVENTS
ABOUT
FAQ
SEMANTIC SYSTEMS BIOLOGY
Welcome to the Semantic Systems Biology Portal
What is Systems Biology?
There are many definitions of systems biology, ranging from: "a research approach that seeks to describe the overall behavior of a biological system through detailed, quantitative experimentation combined with conceptual or computational modeling of the systems components and their interactions" (c.f. SysBioSIG), to the very concise: "the study of biological function that derives from interactions".
What is Semantic Systems Biology?
Semantic technologies are playing an an increasingly important role in capturing and modeling biological knowledge. Semantic systems biology can complement the bottom-up approach with data-driven generation of hypotheses. Therefore, Semantic Systems Biology (SSB) is a systems biology approach that uses semantic description of knowledge about biological systems to facilitate integrated data analysis.
What are the features of a Semantic Systems Biology approach?
There are some key elements in this new paradigm:
1. Knowledge representation
2. Data integration
3. Reasoning => hypothesis
4. Querying => hypothesis
What is the BioGateway?
The BioGateway is an initiative that enables a Semantic Systems Biology approach. It provides an entry point to access a data warehouse where biological data is gathered in the form of triples (using RDF). The systems can be queried using SPARQL.
Last Updated (Thursday, 21 August 2008 11:42)
NEWSFLASH
The new Semantic Systems Biology web site has been released (17.06.2008).
MAIN MENU
Home
BioGateway
News & Events
About
FAQ
RESOURCES
SPARQL spec
RDF spec
Planet RDF
BioGateway
The homepage of SSB, including BioGateway as a first step towards this approach


SPARQL - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.semantic-systems-biology.org/biogateway/querying
Google
SEMANTIC SYSTEMS BIOLOGY
HOME
BIOGATEWAY
NEWS & EVENTS
ABOUT
FAQ
Home ▸ BioGateway ▸ Querying
SPARQL
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:
[Select a query]
Query:
Type a query here.
Run
Prefixes
Comment
Uncomment
Optional
Indent
NEWSFLASH
The Semantic Systems Biology team is attending the ICSB 2008
MAIN MENU
Home
BioGateway
Architecture
Resources
Tutorial
Querying
News & Events
About
FAQ
Use the buttons for prefixes and other constructs
Click Run!

Prospective users

- Molecular biologist: interacting components, events, roles that each component play. Hypothesis evaluation.



- Bioinformatician/Computational Systems Biologist: data integration, annotation, modeling and simulation.

- General audience: educational purposes.

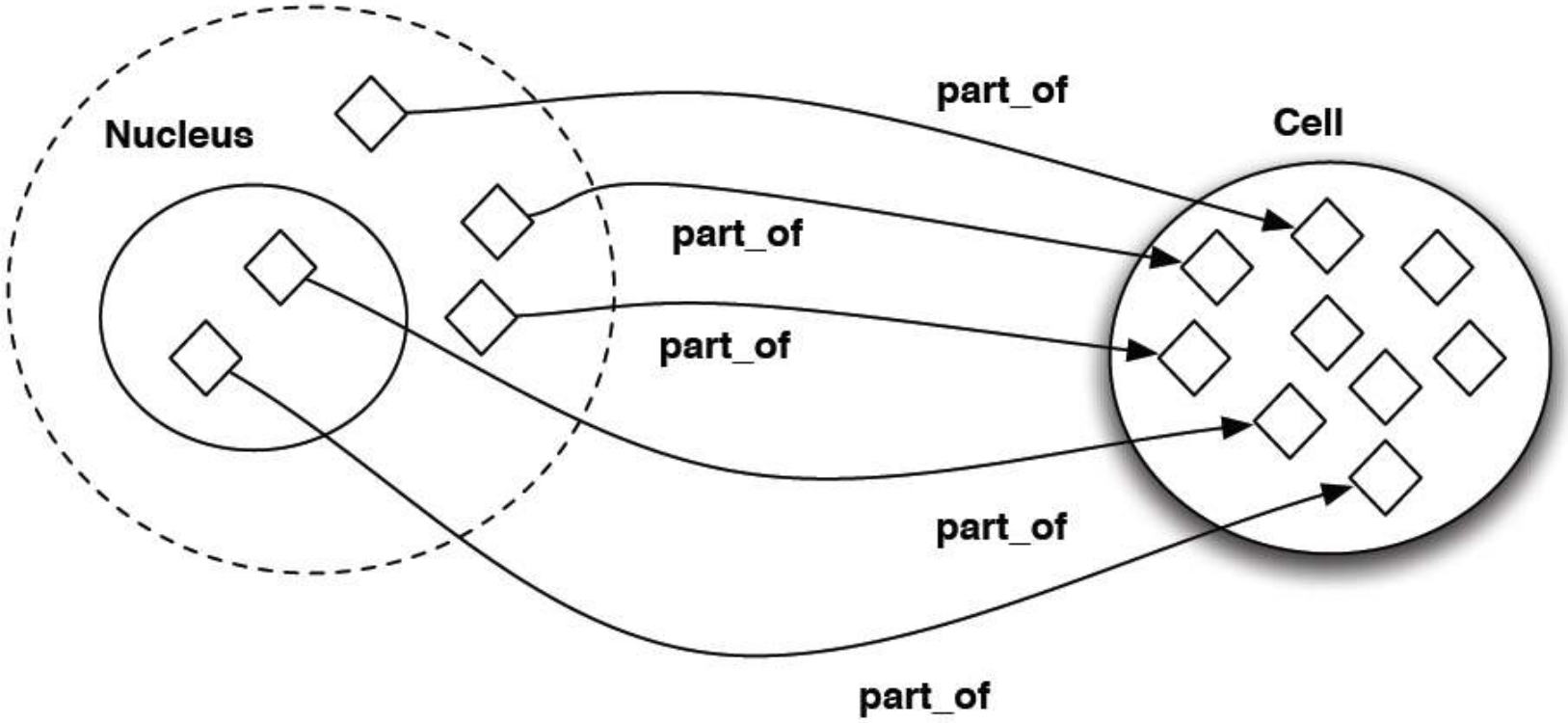
Resources

- Open Biomedical Ontologies (OBO)
  - About 60 bio-ontologies (mainly OBOF)
  - OBO Foundry
  - Multidisciplinary teams: philosophers, computer scientists, domain experts (biologists), …
- Tools (OBO-Edit, Protégé, etc.)
- Data centres (academy/industry) “migrating” towards ontology-aware resources

Format mapping: OBO⇔OWL

- Mapping not totally biunivocal; however, all the data has been preserved.
- Missing properties in OWL relations:
  - reflexivity,
  - asymmetry,
  - Intransitivity, and
  - partonomic relationships.
- Existential and universal restrictions cannot be explicitly represented in OBO => Consider all as existential.
- CCO in OWL is in sync with the NCBO mapping (DL)
- Mapping efforts:
  - http://spreadsheets.google.com/ccc?key=pWN_4sBrd9l1Umn1LN8WuQQ
  - http://www.psb.ugent.be/cbd/cco/OBO2OWL%20Mappings.pdf

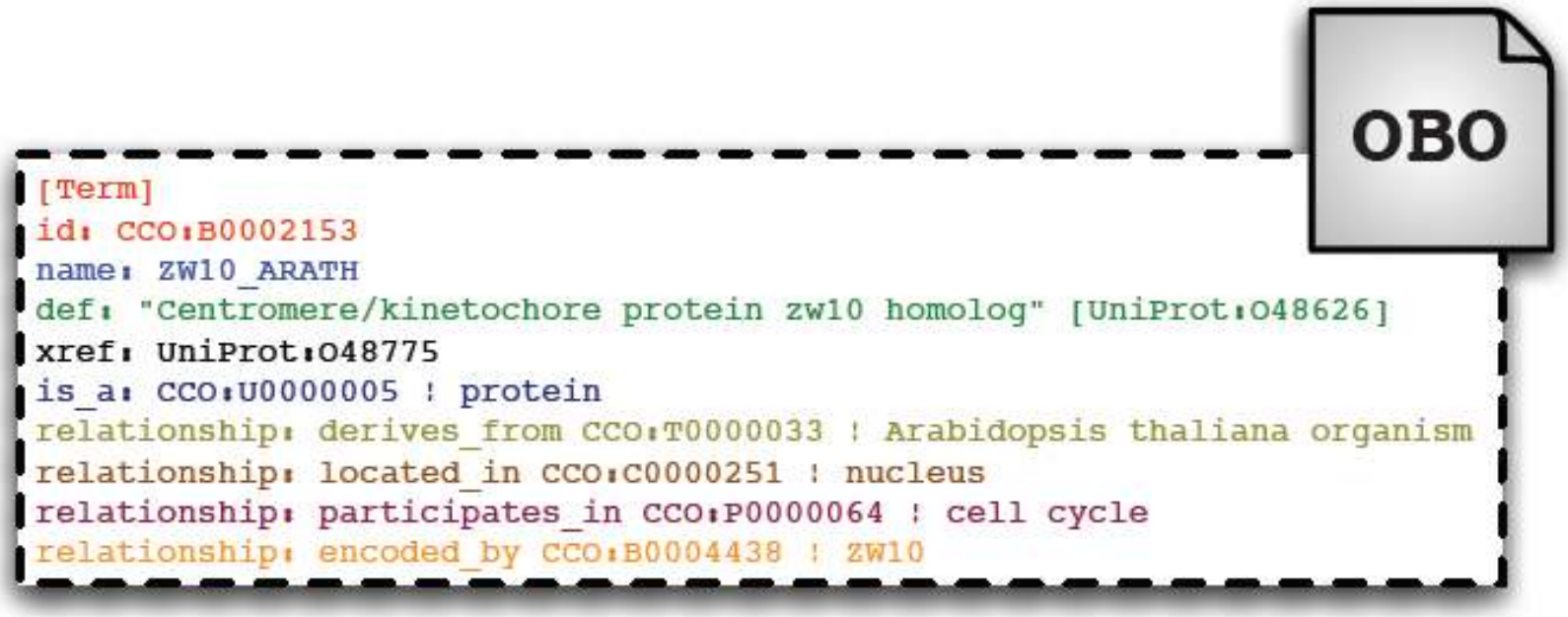
OWL restrictions



Restriction on Nucleus: some **part_of** Cell

Necessary conditions vs Necessary and sufficient conditions

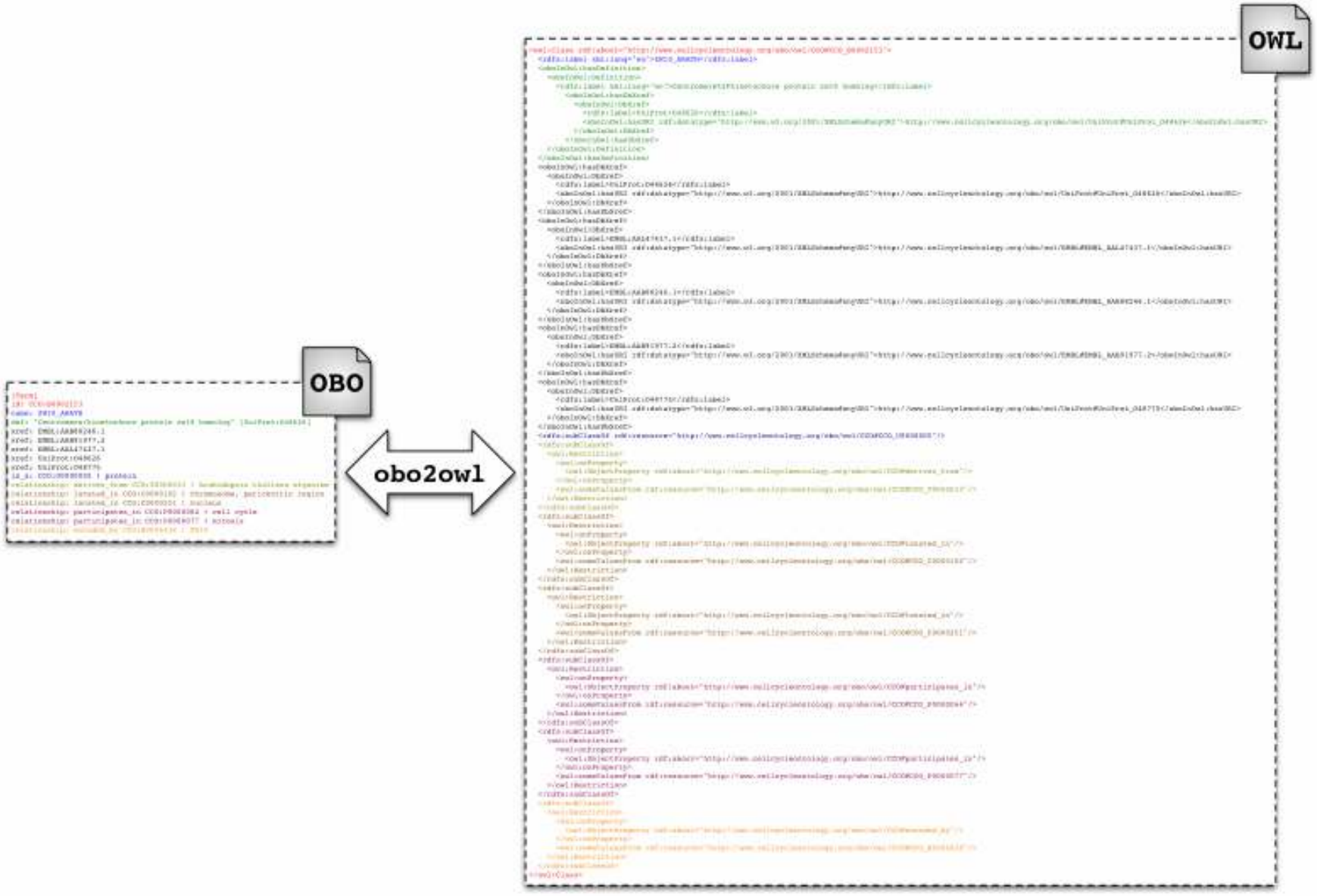
Sample entry in OBO



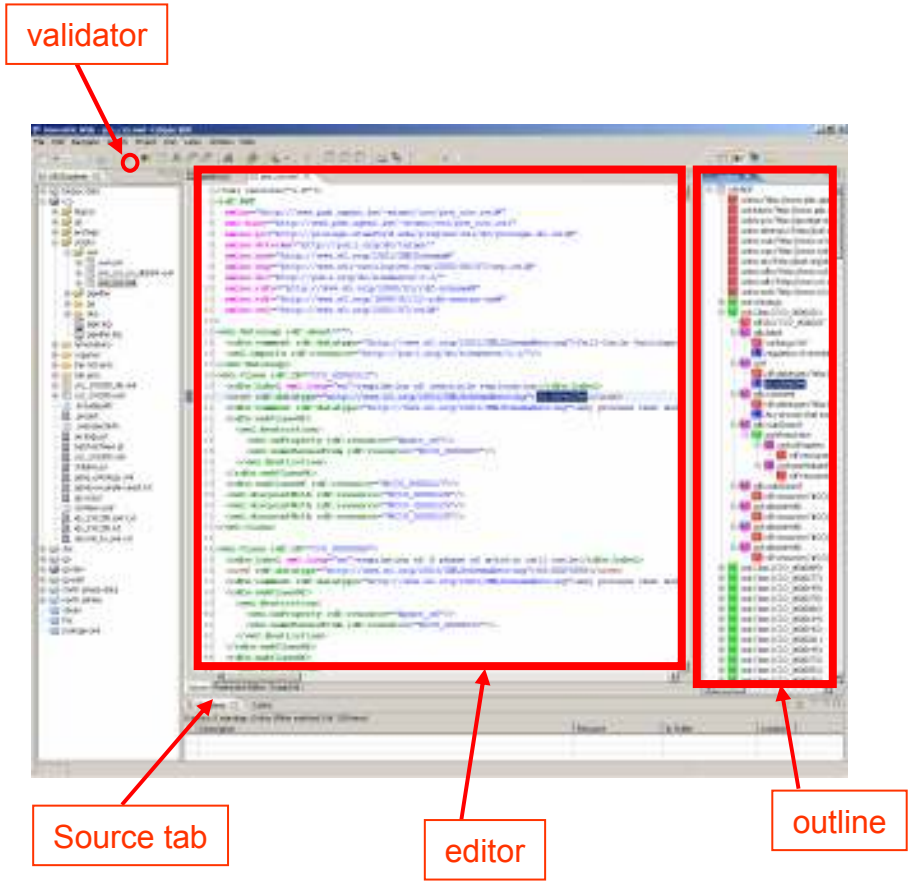
Sample entry in OWL

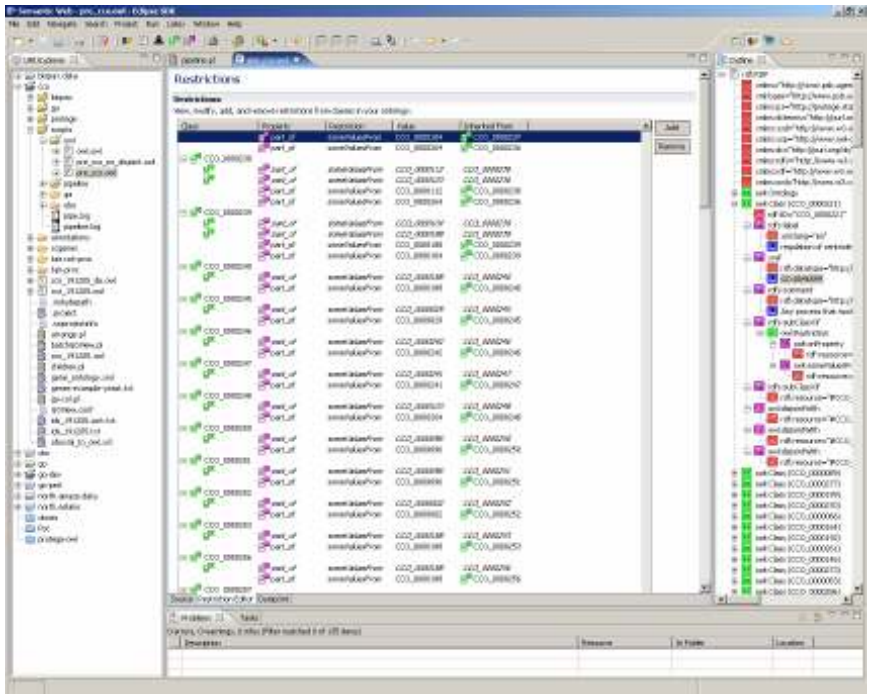
```
<owl:Class rdf:about="http://www.cellcycleontology.org/ontology/owl/CCO#CCO_B0002060">
    <rdfs:label xml:lang="en">NEB2_HUMAN</rdfs:label>
    <oboInOwl:hasDefinition>
        <oboInOwl:Definition>
            <rdfs:label xml:lang="en">Neurabin-2</rdfs:label>
            <oboInOwl:hasDbXref>
                <oboInOwl:DbXref>
                    <rdfs:label>UniProt:Q96SB3</rdfs:label>
                    <oboInOwl:hasURI rdf:datatype="http://www.w3.org/2001/XMLSchema#anyURI">
                        http://www.cellcycleontology.org/ontology/owl/UniProt#UniProt_Q96SB3
                    </oboInOwl:hasURI>
                </oboInOwl:DbXref>
            </oboInOwl:hasDbXref>
        </oboInOwl:Definition>
    </oboInOwl:hasDefinition>
    <oboInOwl:hasDbXref>
        <oboInOwl:DbXref>
            <rdfs:label>UniProt:Q8TCR9</rdfs:label>
            <oboInOwl:hasURI rdf:datatype="http://www.w3.org/2001/XMLSchema#anyURI">
                http://www.cellcycleontology.org/ontology/owl/UniProt#UniProt_Q8TCR9
            </oboInOwl:hasURI>
        </oboInOwl:DbXref>
    </oboInOwl:hasDbXref>
    <rdfs:subClassOf
rdf:resource="http://www.cellcycleontology.org/ontology/owl/CCO#CCO_B0000000"/>
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty>
                <owl:ObjectProperty rdf:about=
                "http://www.cellcycleontology.org/ontology/owl/CCO#belongs_to"/>
            </owl:onProperty>
```

OBO2OWL mapping sample



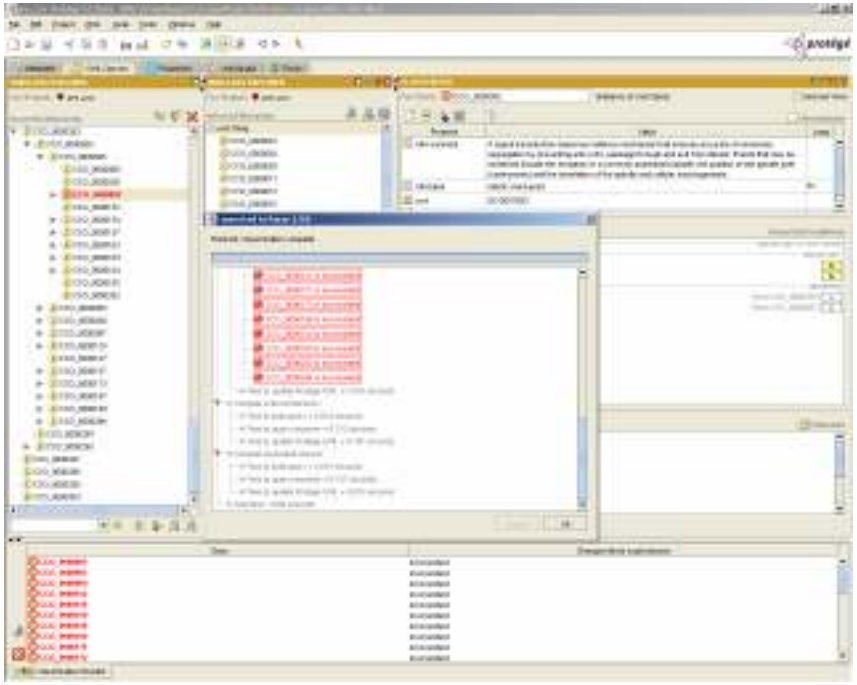
CCO checked with…

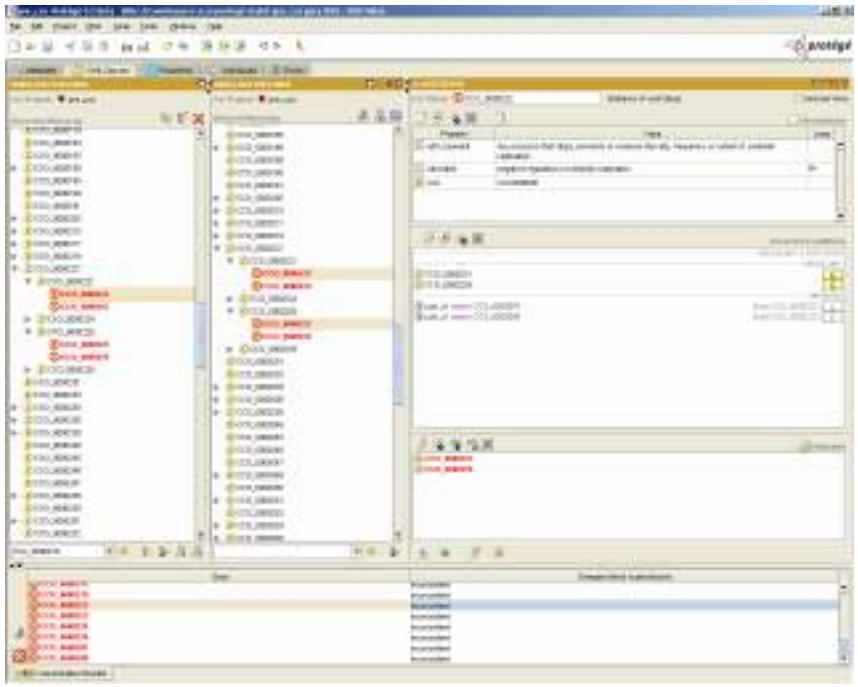




SWeDE Eclipse plug-in: http://owl-eclipse.projects.semwebcentral.org

Checked with…

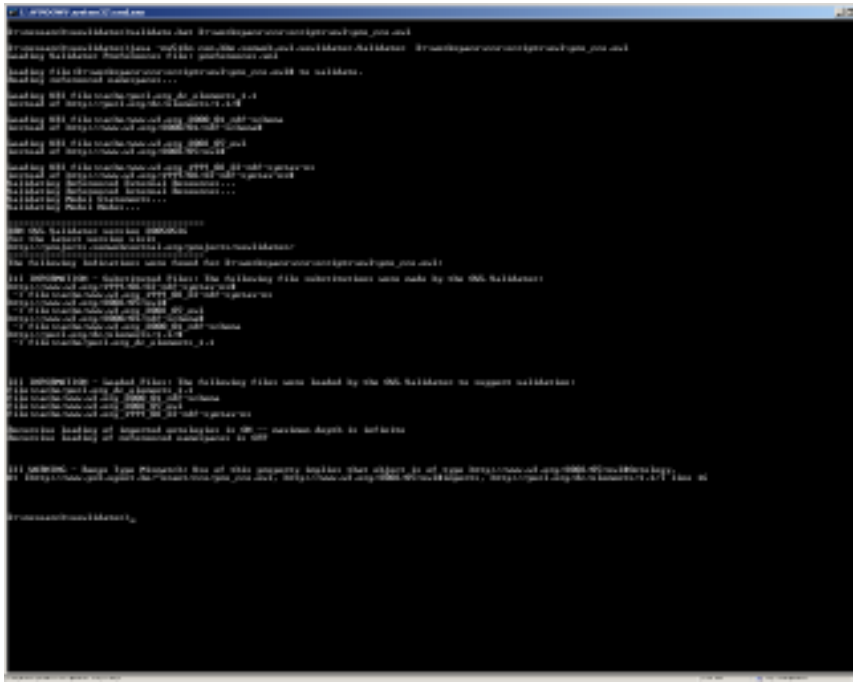




Protégé: http://protege.stanford.edu/

A reasoner (RACER) was used to identify those inconsistencies

and with…



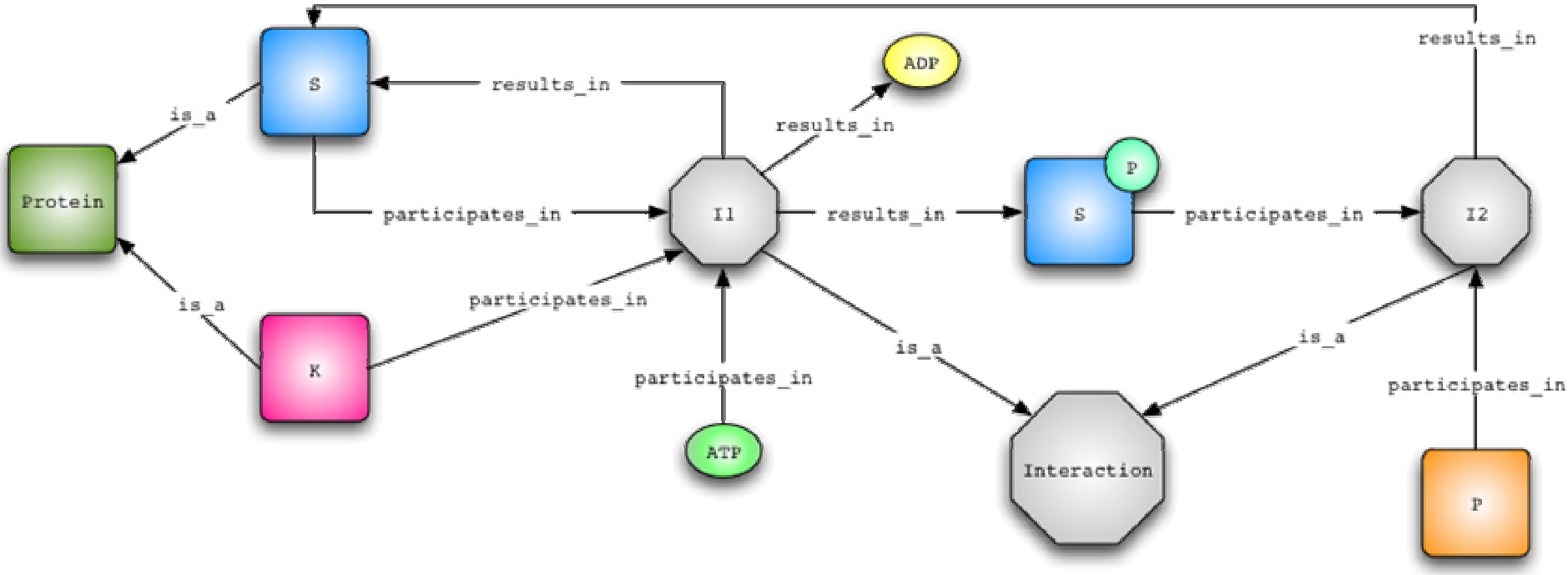
Vowlidator: http://projects.semwebcentral.org/projects/vowlidator/)

OPPL in CCO

```
# Add a class called "interaction".
# Add the following neccesary condition to the newly added "interaction" class:
# the participants are only the union of protein_1 and protein_2.
# Add the rdfs:label "interaction" to the newly added  "interaction" class.

ADD Class interaction;
ADD subClassOf has_participant only (protein_1 or protein_2);
ADD label "interaction";

# Select any class that has the following condition as a superclass:
# the participants are only the union of protein_1 and protein_2.
# Remove the rdfs:label "interaction" from any selected class.
# Add the rdfs:label "interaction of protein_1 and protein_2" to any selected class.

SELECT subClassOf has_participant only (protein_1 or protein_2);
REMOVE label "interaction";
ADD label "interaction of protein_1 and protein_2";
```


Egaña, M., Stevens, R. Antezana, OWL-ED, 2008

Sample model



Phosphorylation - Dephosphorylation modeling:
S (Substrate), K (Kinase), P (Phosphatase): Proteins
I_1, I_2: Interactions
ATP, ADP: Small molecules

http://www.CellCycleOntology.org

Other sample query in OWL

- Entities that are the location of proteins participating in the S-phase (CCO_P0000014) or any process which is part of it.?

location_of **some** (
participates_in **some** (
CCO_P0000014 **or** (
part_of **some** CCO_P0000014)))

Initial question (revisited)

"to what extent can current Semantic Web technologies support biological knowledge management for basic or complex querying, for automated reasoning and inferencing, specifically in the context of a Systems Biology approach where integrated knowledge could be utilised to address the relations between components of the cell cycle control mechanism, their involvement in (sub)modules of cell cycle control, and their potential place in overall network topology?"

Mathematical model

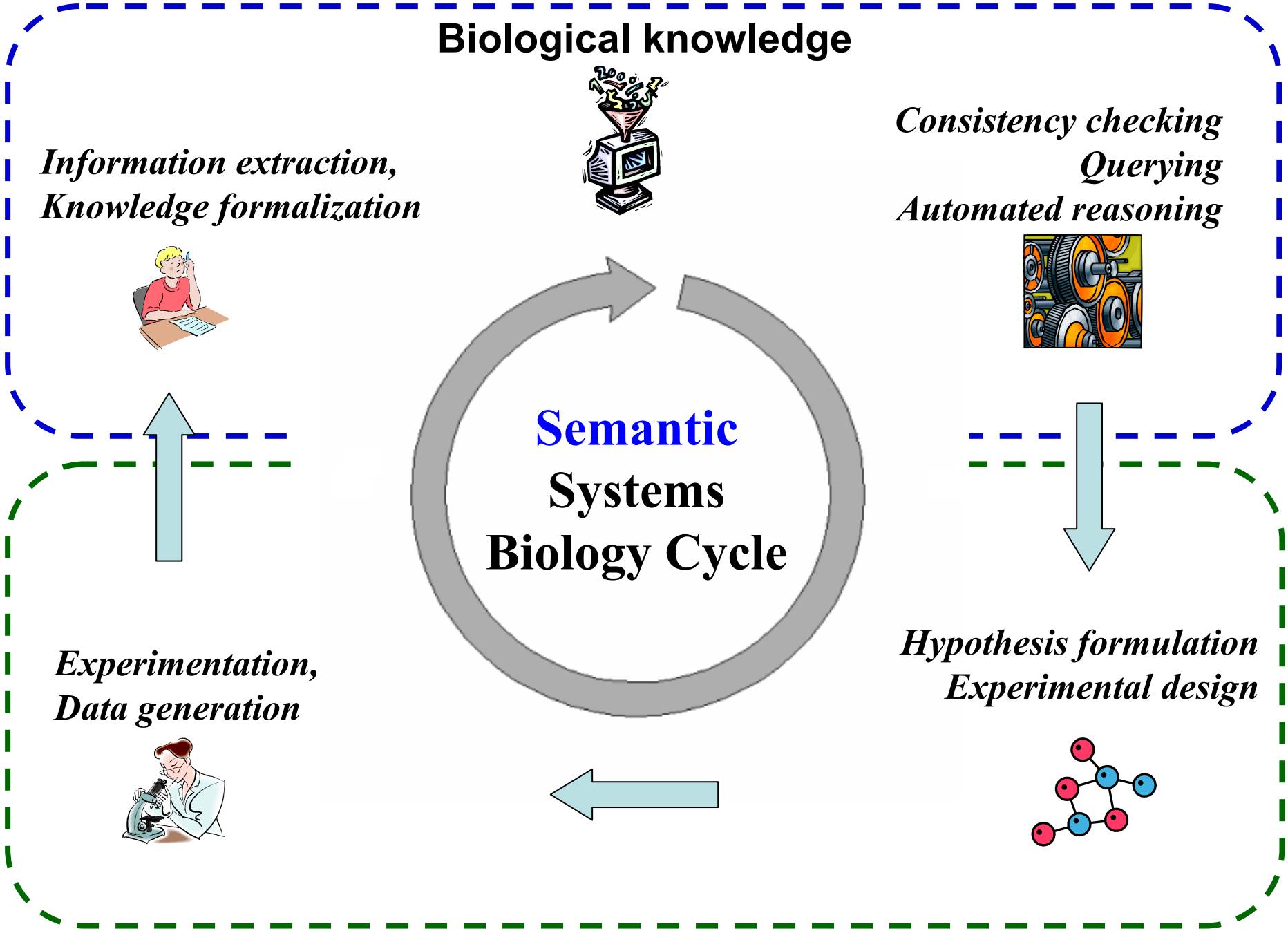
New information to model
Model Refinement

Data analysis
Information extraction

Systems Biology Cycle

Experimentation,
Data generation

Dynamical simulations and
hypothesis formulation
Experimental design


Biological knowledge
Information extraction,
Knowledge formalization
Consistency checking
Querying
Automated reasoning
Semantic
Systems
Biology Cycle
Experimentation,
Data generation
Hypothesis formulation
Experimental design

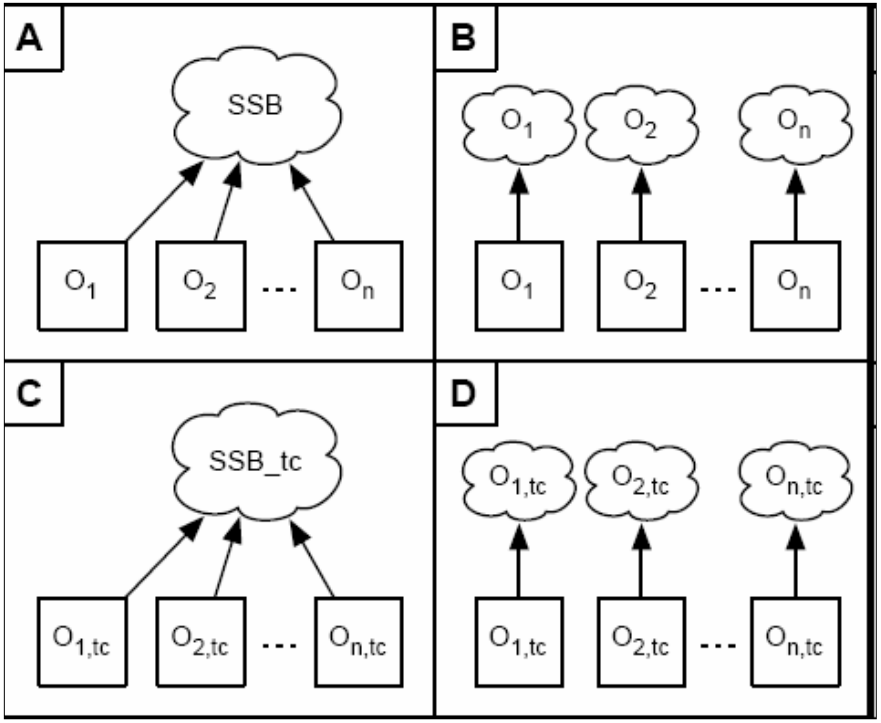
BioGateway

- Automatic pipeline
  - Run on a regular basis (~6 months)
  - Latest data available (from scratch)
- Uses Virtuoso Open Server
  - Open Source software that can host a triple store
  - Can build this from RDF files
  - Has a DB backend
- Supports SPARQL*

http://www.openlinksw.com/virtuoso/

*http://www.w3.org/TR/rdf-sparql-query/

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>).
}
```

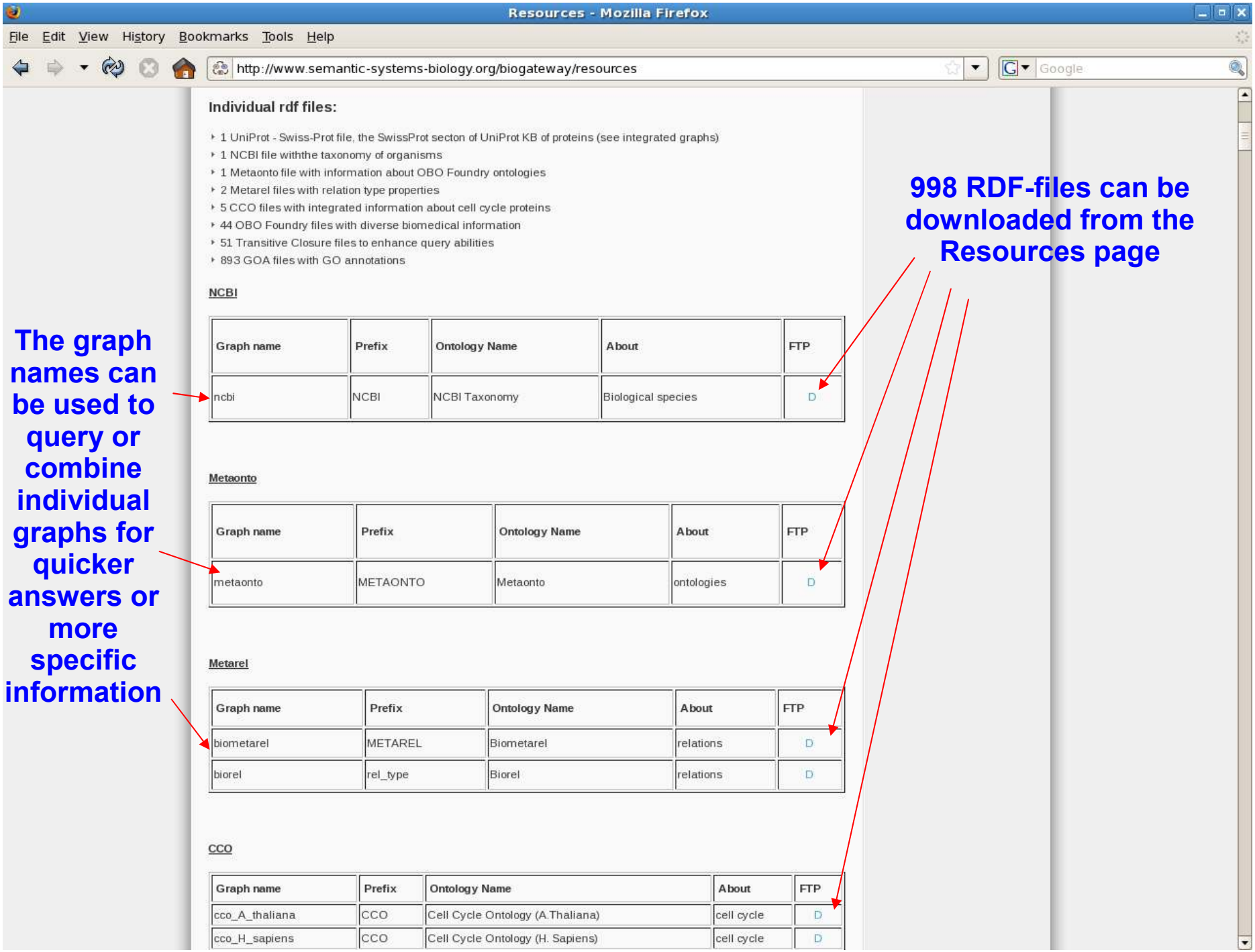
Click here to select this query in the drop-down box on the query-page and edit it

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
Google
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
The graph names can be used to query or combine individual graphs for quicker answers or more specific information
998 RDF-files can be downloaded from the Resources page

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

term_as_child	outward_arrow	head_name
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

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

Outgoing arrows

Incoming arrows

tail_name	inward_arrow	term_as_parent
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

Principles

1. Orthogonality
2. A “common language” (e.g. RDF)
3. Unique ID + resolution (e.g. purl.org)
4. Comply to: ULO (e.g. BFO), RO, …
5. Explicit semantics
6. Rich axiomatisation
7. Application-driven development (e.g. SB)
8. Peer review (community evaluation)
9. Tooling (e.g. visualisation)
10. Licensing (e.g. CC)

Metarel

- Metarel is a generic ontological hierarchy for relation types, consistent with OBOF and RDF.
- It includes meta-information like transitivity, reflexivity and composition.
- BioMetarel includes all the biological relation types that are used in BioGateway.
- We are still testing the exploitation of composition, like ***A located in B*** and ***B part of C***, gives ***A located in C***.

Blondé, W., Antezana et al. BMC Bioinformatics, 2009

The RDF export specifications

- The RDF is automatically generated with onto-perl, our own ontology API.
- Many choices for the RDF specifications were made during the testing of the queries.
- The resources are available either as part of an integrated graph or as individual graphs.
- BioMetarel, a relation ontology, provides labels for the URIs of the relations.
- OWL(XML/RDF) was avoided because it is too verbose. We preferred RDF optimized for querying.

Next steps

- More data sources (e.g. Nutrigenomics, pathways etc.)
- RDF rules (e.g. RuleML)
- A more user-friendly interface
- Reasoning
- OBO cross products
- …

Systems biology paradigm

top-down and bottom-up modeling

top-down ***bottom-up***

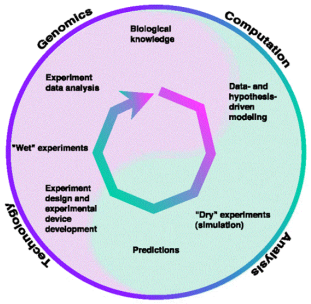
data driven ***hypothesis driven***

Biological Process

Genome-scale functional genomics data

Predictive mathematical model

Statistics Mining



Knowledge Mathematics

Knowledge Management

Discussion

- W3C standards – limitations (e.g. spatio-temporal information, microarrays experiments)
- **Biological identifiers**: URIs, LSIDs, MIRIAM URIs, etc. *They should be scalable & resolvable.*
- **Lack of semantic content**: poor axiomatisation, inadequately codified. *Use of standard languages (e.g. RDF).*
- **Adequate tools**: not adapted for real-size problems (e.g. reasoning). *Designed with a universal architecture in mind.*

SSB at a community level*

- **Semantic bio-content: encourage and facilitate**
- **Best practices for such creation (standards)**
- **Mechanism for identifying biological entities**
- **Bridge semantic technology developers and life scientists (=the users)**

WIKI: http://www.bio.ntnu.no/systemsbiology/ssbwiki/doku.php

Conclusions / Results

- Data integration pipeline: life cycle of the KB
- Existing integration obstacles due to:
  - diversity of data formats
  - lack of formalization approaches
- Reasoning services: inconsistency checks, classification => hypothesis
- Trade-offs: complex queries, representational issues

Current issues

- Temporal & spatial representation
  - OBOF not enough…
- Performance (reasoners)
  - Huge ontologies
- Weighted knowledge (*often, sometimes*)

CCO accession number

CCO:[CPFRTIBGOU]nnnnnnn

namespace

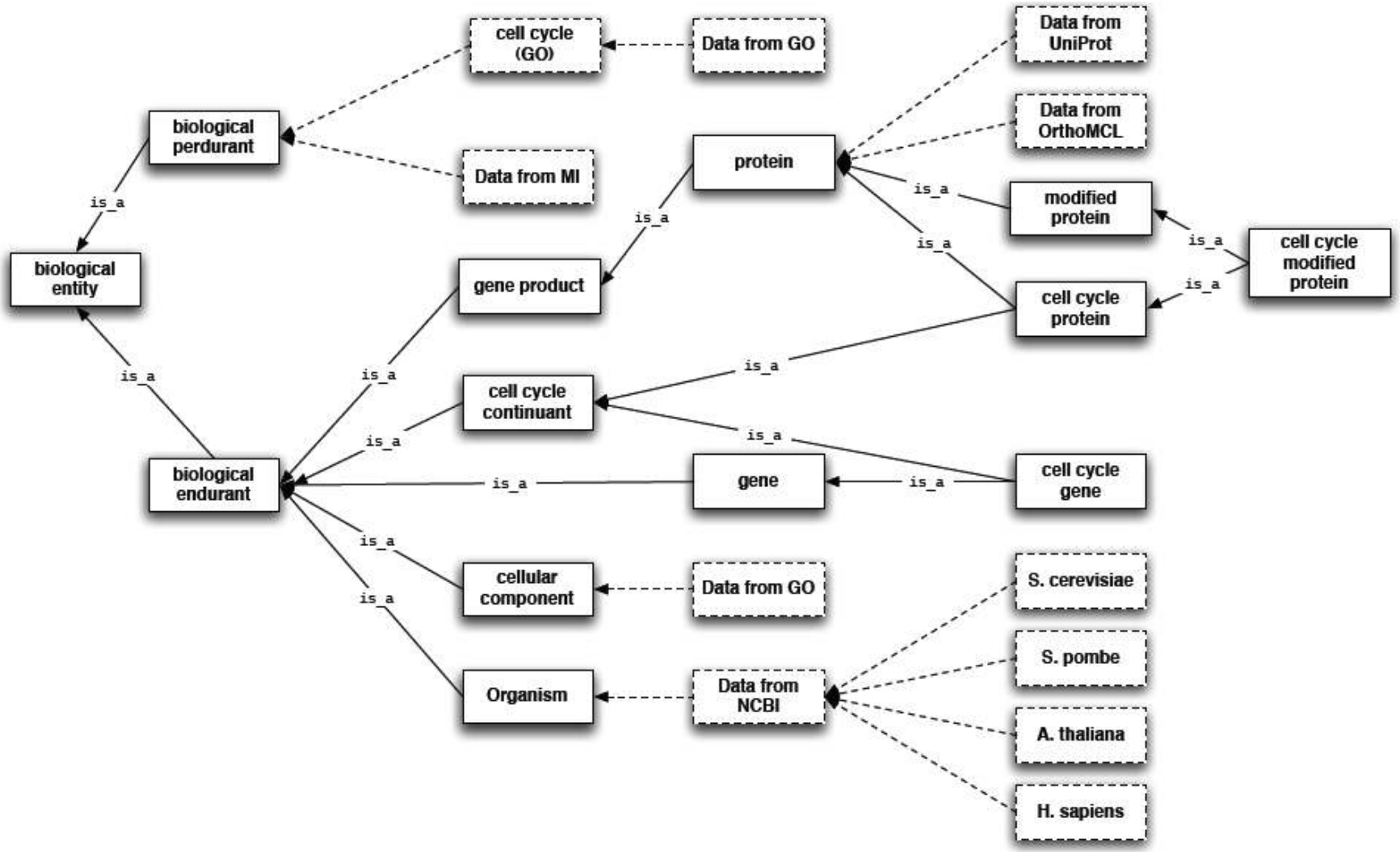
sub-namespace

7 digits

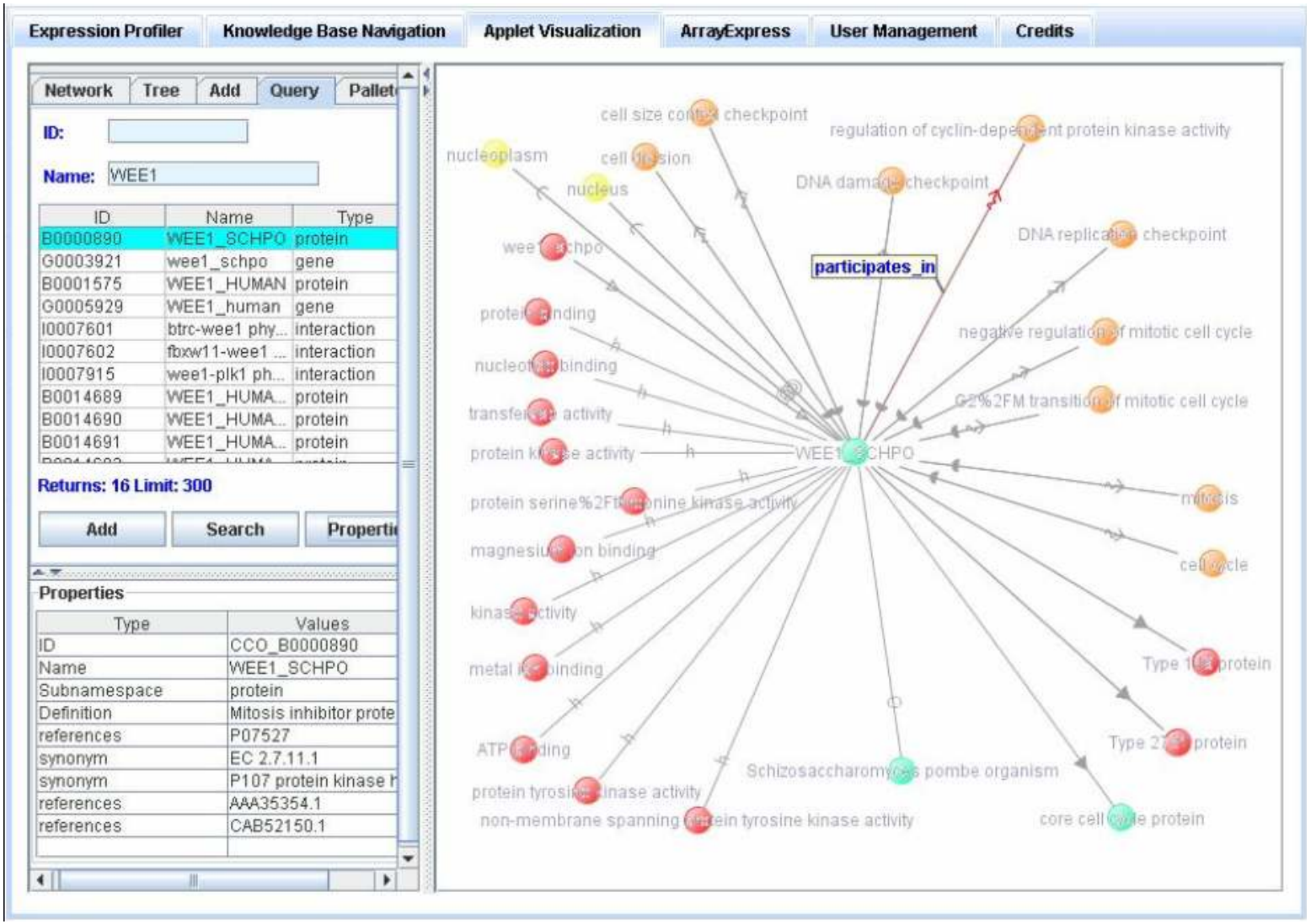
C: *cellular component*
P: *biological process*
F: *molecular function*
R: *reference*
T: *taxon*
I: *interaction*
B: *protein*
G: *gene*
O: *ortholog*
U: *upper-level term*

Example: CCO: P0000056 ⟷ "*cell cycle*"

Upper Level Ontology for Application Ontologies



DIAMONDS platform *



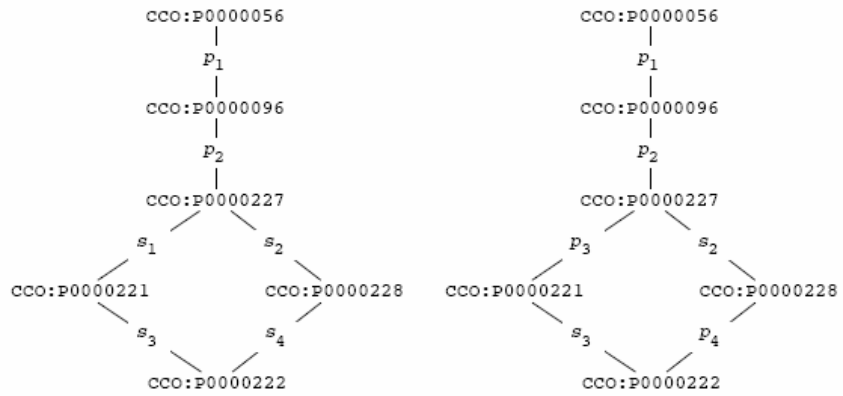
* EU project

Example: checking the single inheritance principle

- Principle: *"No class in a classification should have more than one is_a parent on the immediate higher level" (Smith B. et al.)*
- Detect the relationships which violate that rule using a reasoner (RACER*)
- Solution: disjoint among the terms at the same level of the structure
- 32 problems found:
  - 4: "part_of" instead of "is_a"
  - 18: should stay without any change (FP)
  - 10: not consistent (used terminology)

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

part_of instead of ***is_a***



The sub-ontology on the left has inconsistent relation s_4 (is_a) which has been changed into part_of (right side) .

CCO ID	Term
CCO:P0007049	cell cycle
CCO:P0000096	centrosome cycle
CCO:P0000227	regulation of centrosome cycle
CCO:P0000221	regulation of centriole replication
CCO:P0000228	negative regulation of centrosome cycle
CCO:P0000222	negative regulation of centriole replication