BioGateway: an RDF store for supporting Systems Biology

Erick Antezana

Dept. of Plant Systems Biology

VIB/University of Ghent

erick.antezana@psb.ugent.be

Contents

1. Systems Biology
2. Data integration and exploitation
3. BioGateway
4. Concluding remarks
5. Next steps

The four steps of Systems Biology

1. Define all of the components of the system, build model, simulate and predict

2. Systematically perturb and monitor components of the system

3. Reconcile the experimentally observed responses with those predicted by the model

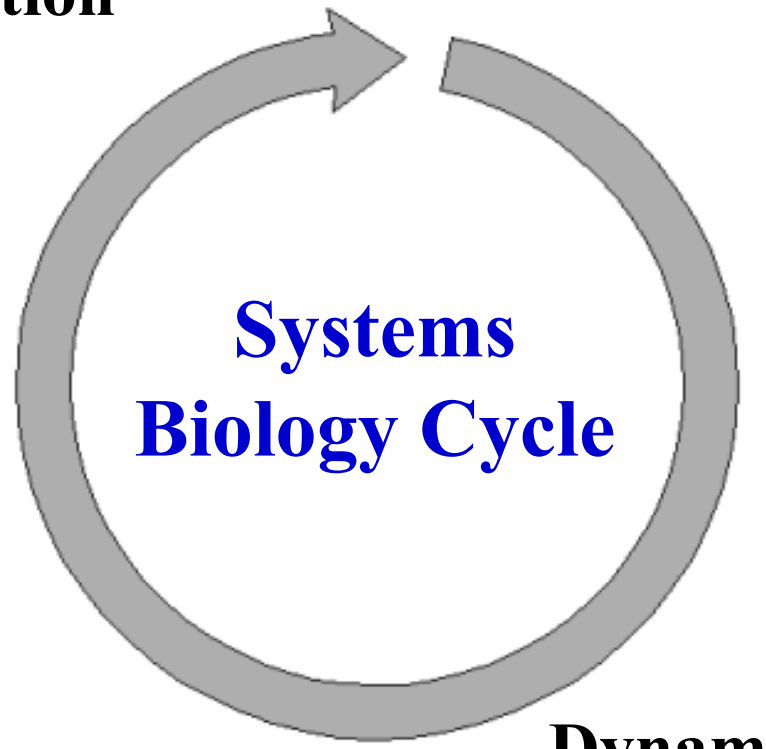
4. Design and perform new perturbation experiments to distinguish between multiple or competing model hypotheses

Kitano, Science, 2002

Mathematical model

Data analysis
Information extraction

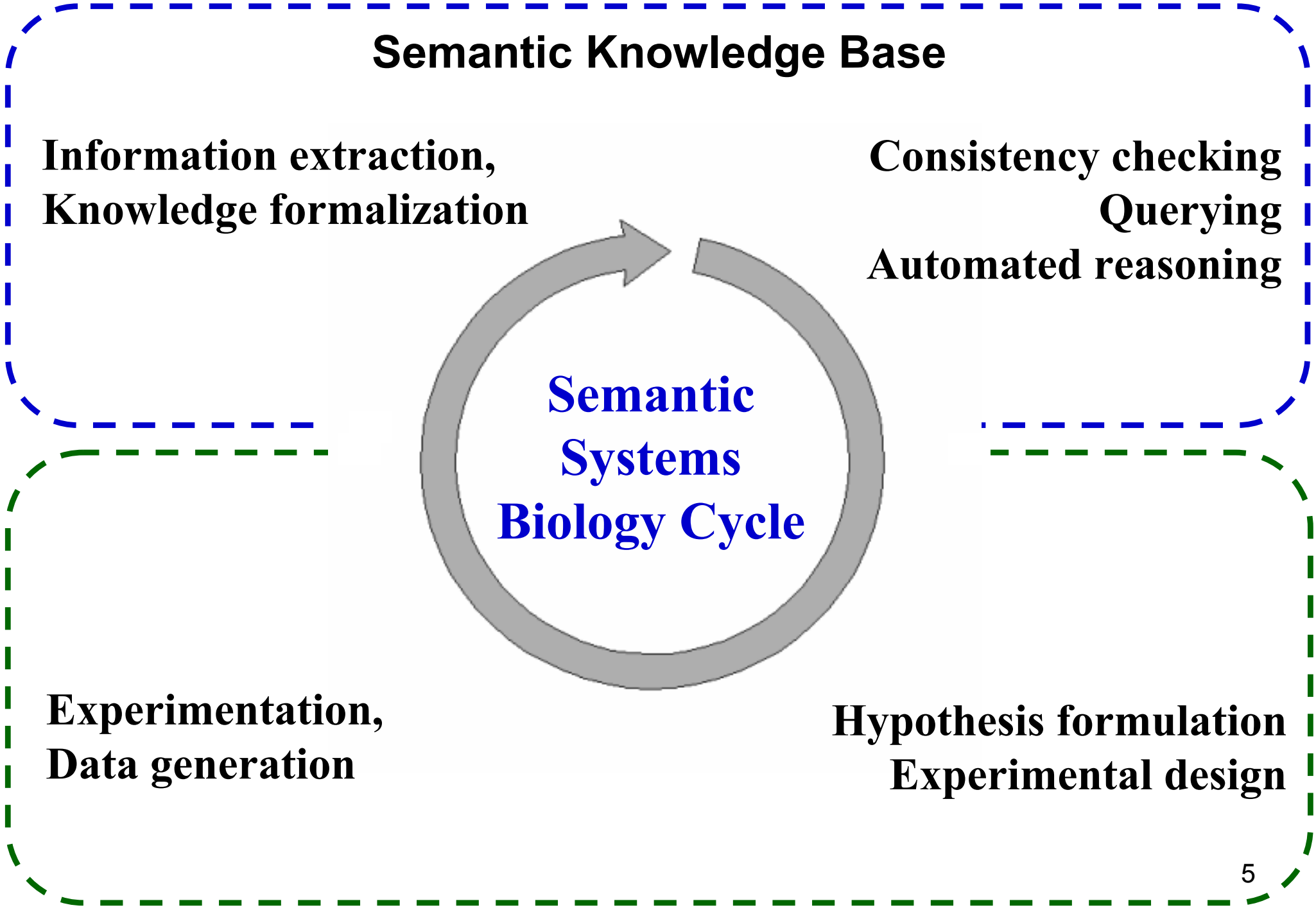
New information to model
Model Refinement



Systems Biology Cycle

Experimentation,
Data generation

Dynamical simulations and hypothesis formulation
Experimental design


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

BioGateway

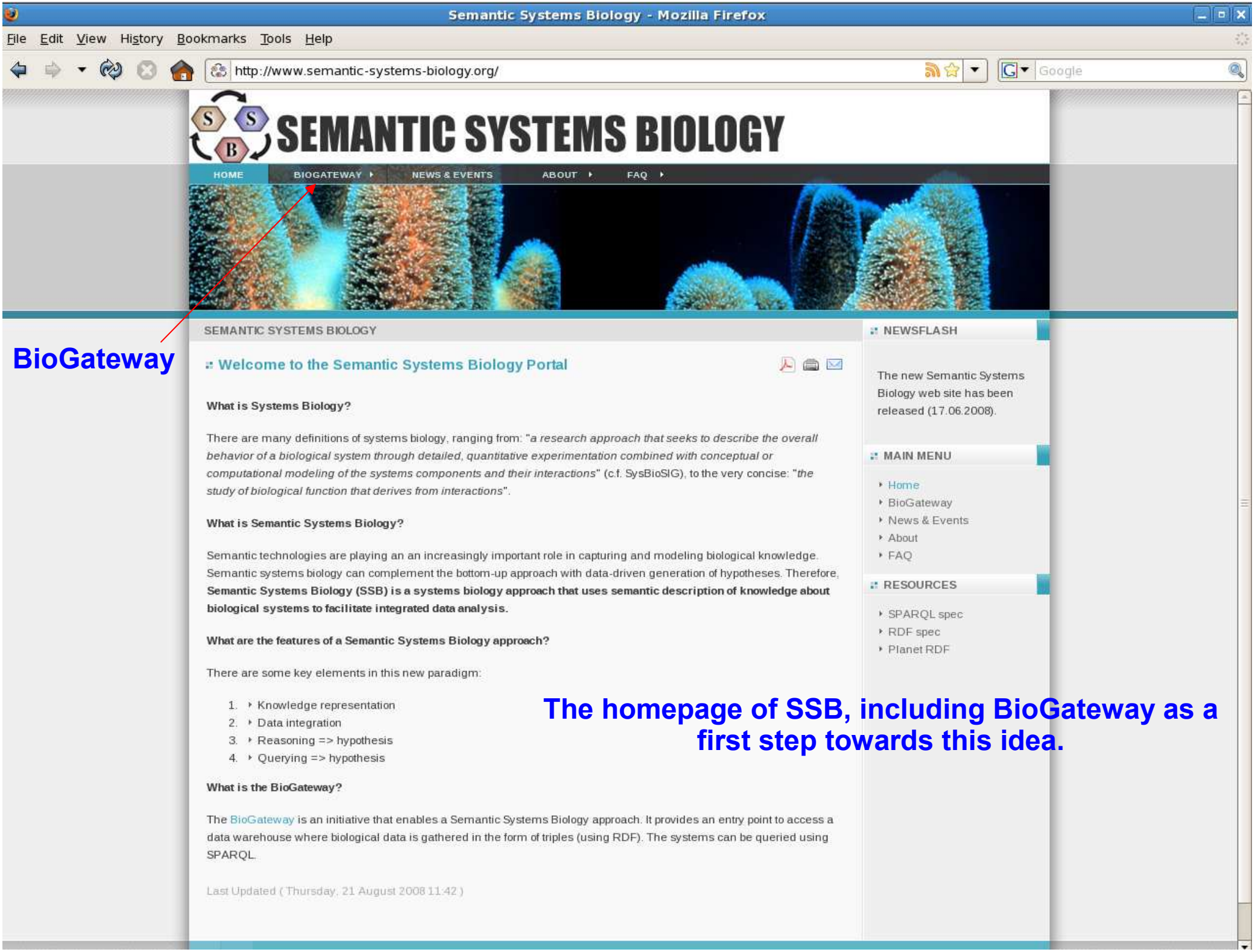
- Uses Virtuoso Open Server
  - Open Source software that can host a triple store
  - Can build this from RDF files
  - Has a DB backend
- Supports SPARQL* language which allows querying RDF data (graphs)
- Its syntax is similar to that of SQL.

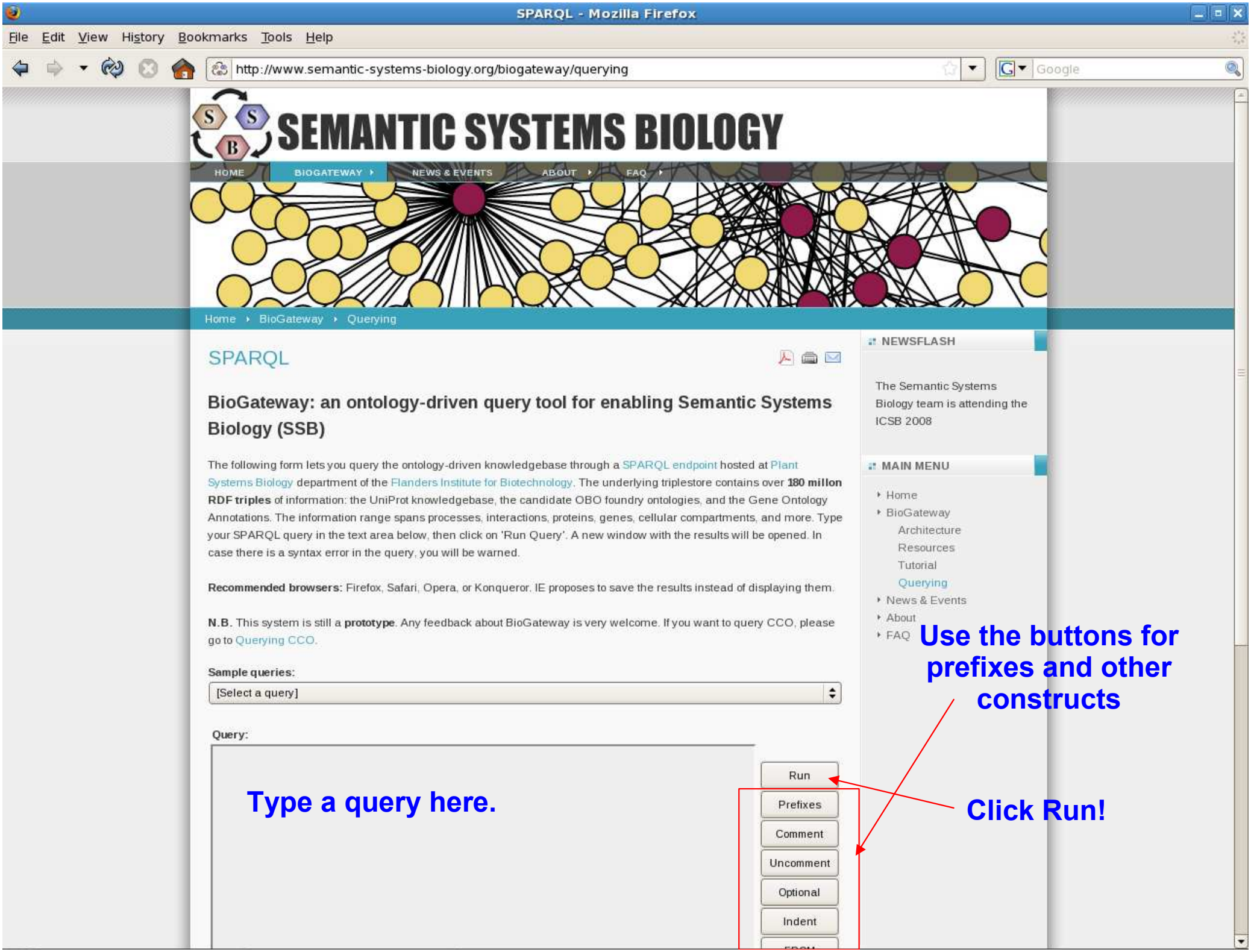
http://www.openlinksw.com/virtuoso/

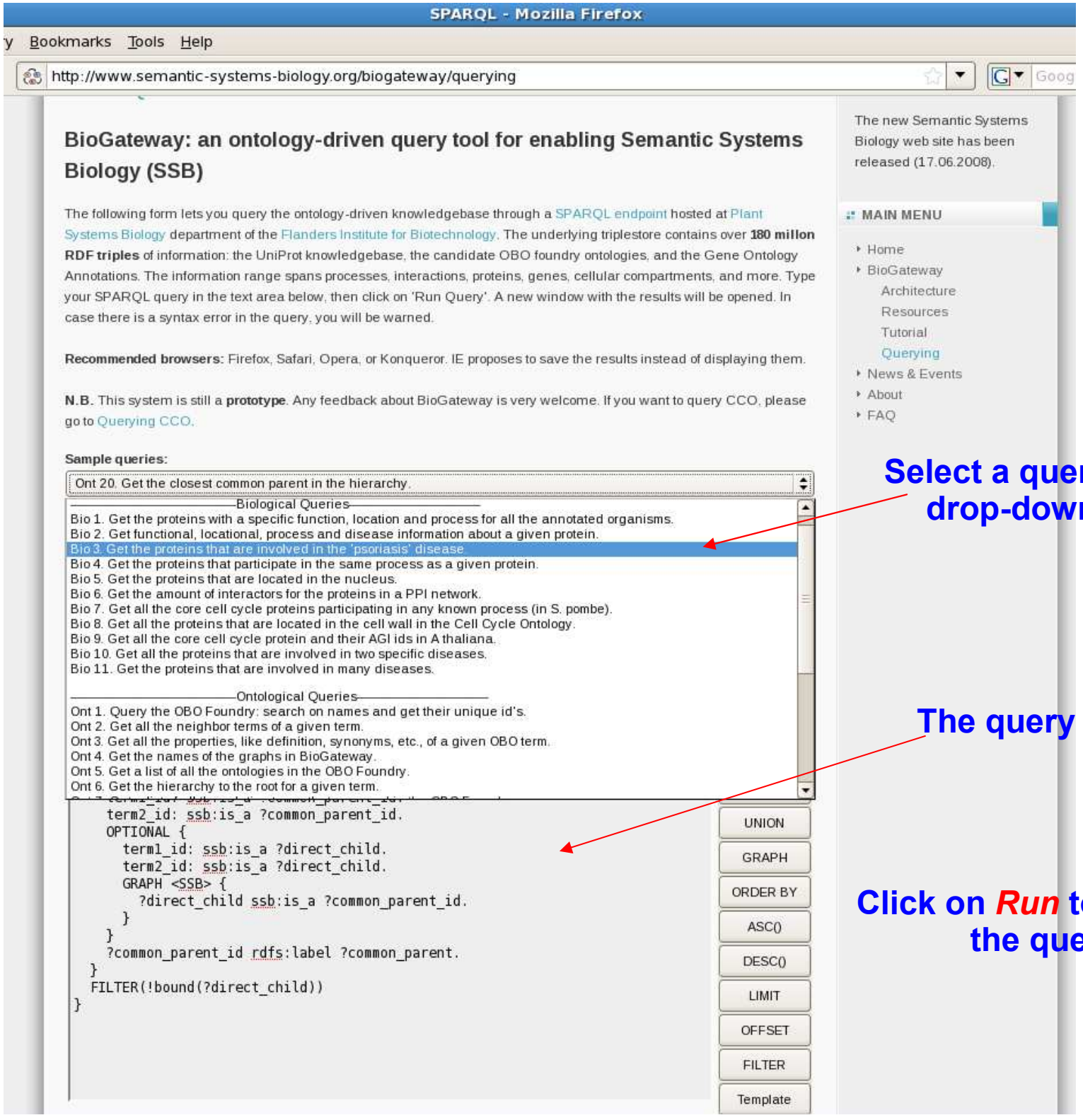
*http://www.w3.org/TR/rdf-sparql-query/

BioGateway
Some motivating questions

- **Cancer:** what candidate genes are involved in cell cycle control, S-phase to G2 transition, DNA damage response and skin cancer?
- **Gastrin:** what genes correlate with cancer and the use of anti-acids, and are involved in the gastrin response, and are associated with cell cycle control?
- **Inflammation:** give me genes that are mentioned in the context of high carbohydrate intake and play a role in (process #1 to be named) and are within x steps from a GO ontology term related to inflammation


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
BioGateway
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
The homepage of SSB, including BioGateway as a first step towards this idea.


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!



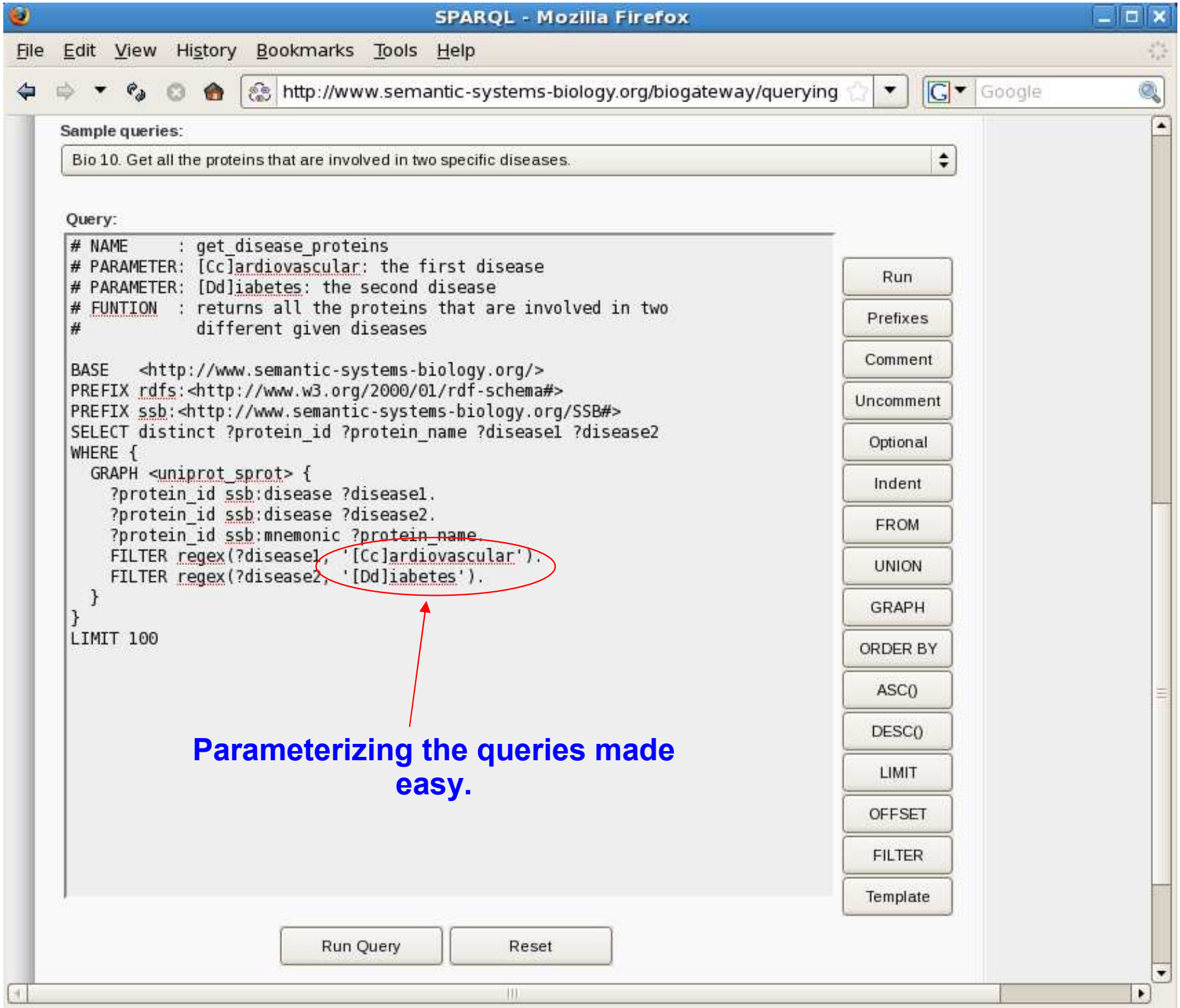
Select a query in the drop-down box

The query editor

Click on *Run* to execute the query

A library of queries

- The drop-down box contains (so far) 31 queries:
  - 11 protein-centric biological queries:
    - The role of proteins in diseases
    - Their interactions
    - Their functions
    - Their locations
  - 20 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,...


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
Parameterizing the queries made easy.

All the queries are explained in a tutorial

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

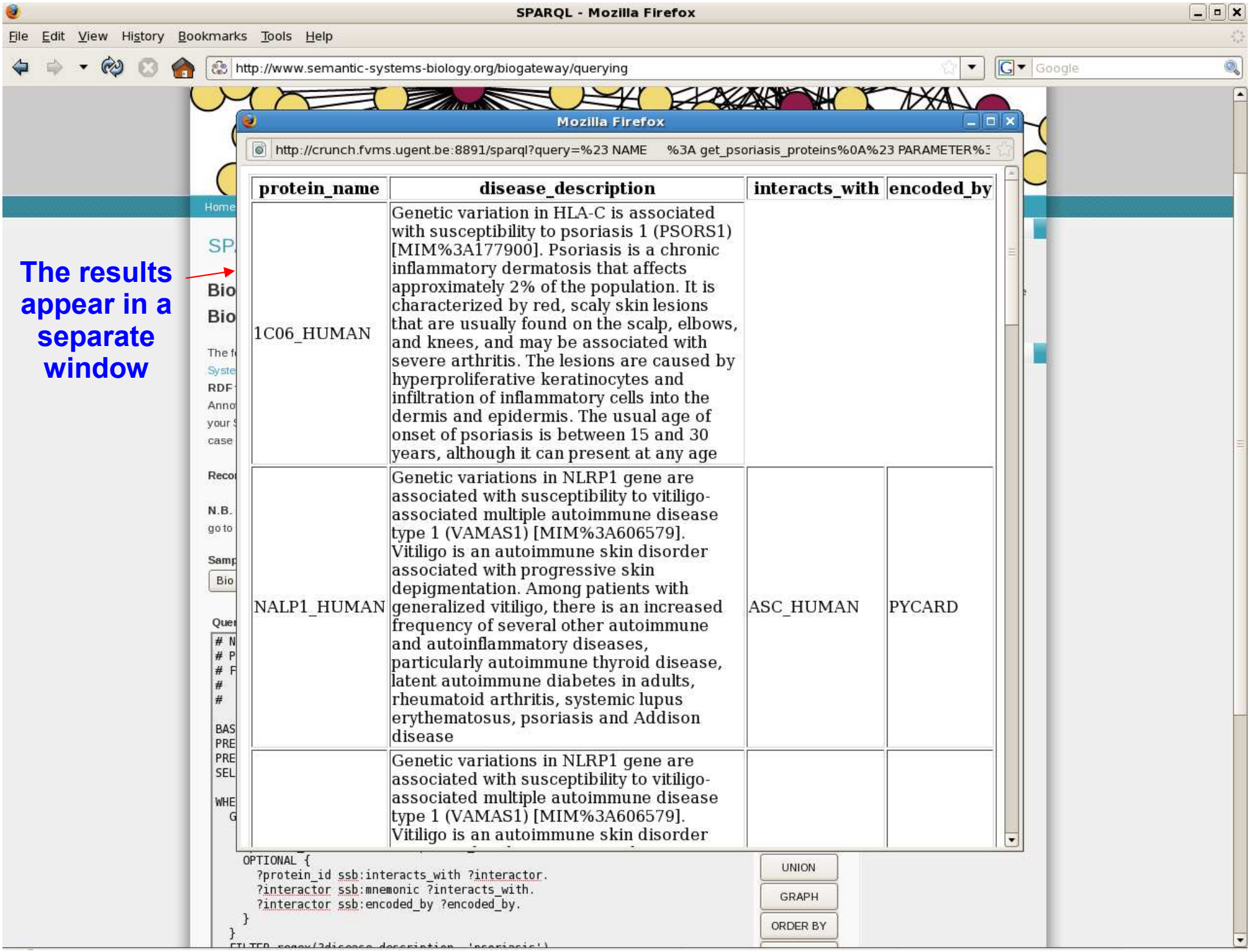
```
# NAME: get_specific_proteins
# PARAMETER: GO_0005216: ion channel activity
# PARAMETER: GO_0005764: lysosome
# PARAMETER: GO_0006811: ion transport
# FUNCTION: returns all the proteins with the same function,
# process and location and the organism in which
# they can be found

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.


SPARQL - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.semantic-systems-biology.org/biogateway/querying
Google
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:interacts_with ?interactor.
?interactor ssb:mnemonic ?interacts_with.
?interactor ssb:encoded_by ?encoded_by.
}
}
UNION
GRAPH
ORDER BY

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

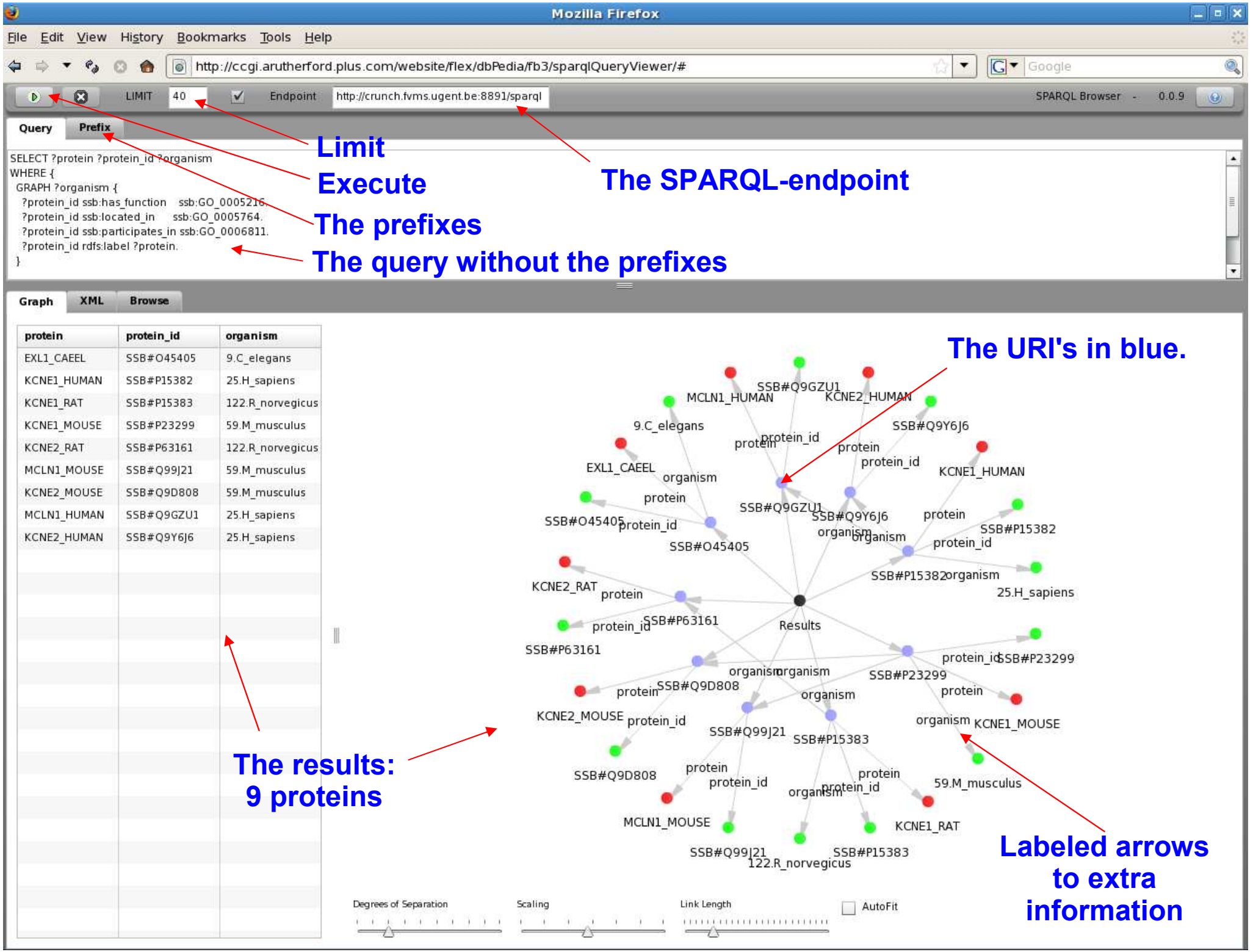
term_as_child	outward_arrow	head_name
1433F_HUMAN	participates in	intracellular protein transport
1433F_HUMAN	participates in	glucocorticoid catabolic process
1433F_HUMAN	participates in	positive regulation of transcription
1433F_HUMAN	participates in	regulation of synaptic plasticity
1433F_HUMAN	participates 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

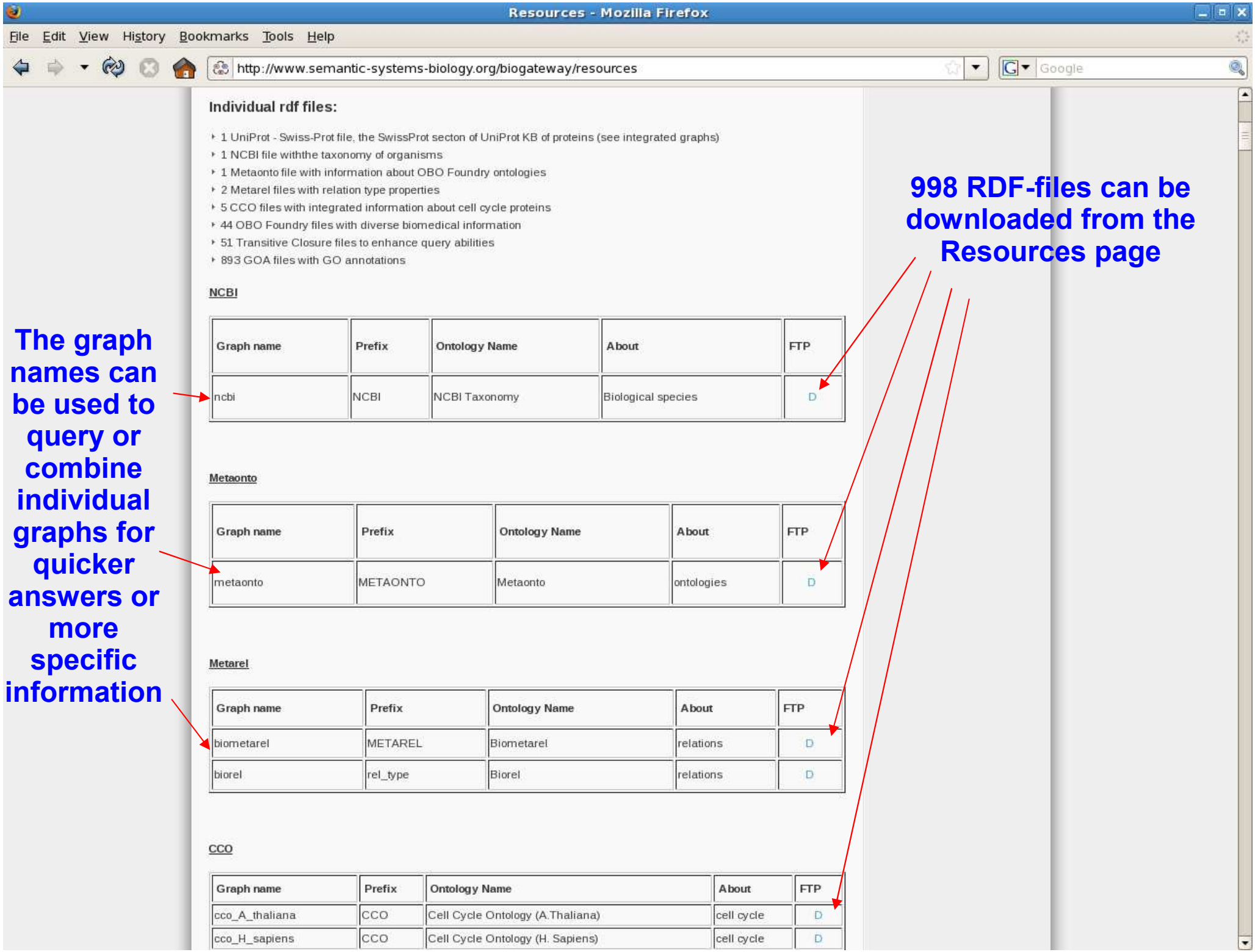
The resulting triples (arrows) are represented as a small grammatical sentence: subject, predicate, object.

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


Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://ccgi.arutherford.plus.com/website/flex/dbPedia/fb3/sparqlQueryViewer/#
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.
}
Limit
Execute
The prefixes
The SPARQL-endpoint
The query without the prefixes
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 URI's in blue.
The results:
9 proteins
Results
Degrees of Separation
Scaling
Link Length
AutoFit
Labeled arrows to extra information


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 RDF export specifications

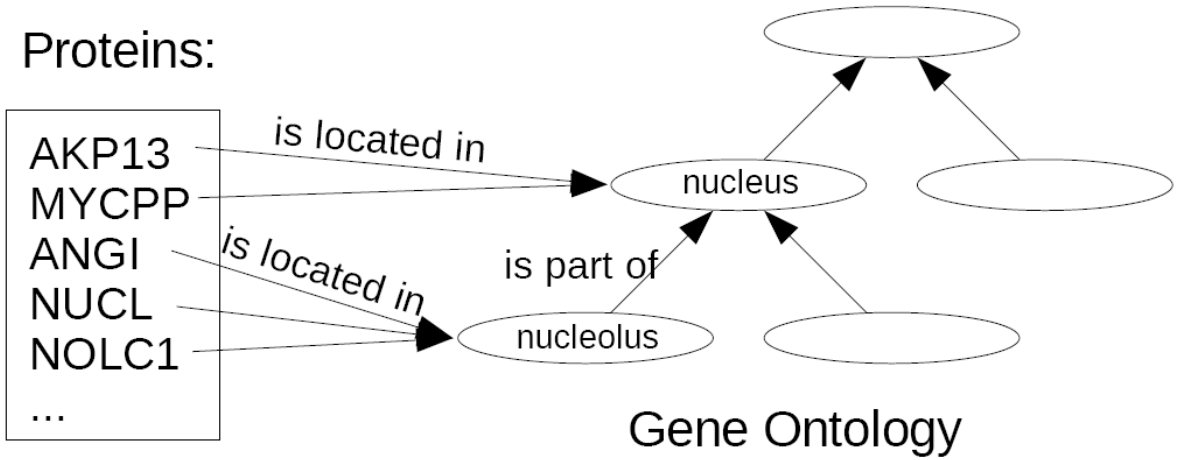
- The RDF is automatically generated with onto-perl, our own ontology API.
- Many choices for the RDF specifications were made during the testing of the queries.
- The resources are available either as part of an integrated graph or as individual graphs.
- BioMetarel, a relation ontology, provides labels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

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***.

Transitive closure graphs

- A transitive closure was constructed for the subsumption relation (is a) and the partonomy relation (part of)
- If ***A is a B***, and ***B is a C***, then ***A is a 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.



Conclusions / Results

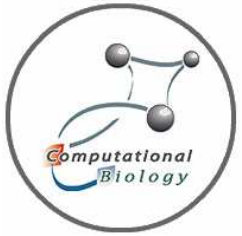
- BioGateway: RDF store for Biosciences
- Data integration pipeline: BioGateway
- Queries and knowledge sources and system design go hand-in-hand (user interaction)
- Existing integration obstacles due to:
  - diversity of data formats
  - lack of formalization approaches
- Calls for '**foundry**' type initiative for RDF

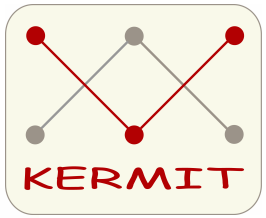
Next steps

- More data sources (e.g. Nutrigenomics, pathways etc.)
- RDF rules
- User interface development
- Reasoning…
- …

Acknowledgements

- Martin Kuiper (NTNU, NO)
- Vladimir Mironov (NTNU, NO)
- Mikel Egaña (U Manchester, UK)
- Robert Stevens (U Manchester, UK)
- Ward Blonde (U Ghent, BE)
- Bernard De Baets (U Ghent, BE)
- Alan Ruttenberg (Science Commons, US)
- Alistair Rutherford (www.netthreads.co.uk)
- Users





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