

Data integration in life sciences: Cell Cycle Ontology

Vladimir Mironov

Dept. of Biology, NTNU

vladimir.mironov@bio.ntnu.no

VILVITE Bergen Science Centre
10 October 2008, Bergen, Norway

Outline

- Background
- CCO implementation
- CCO exploitation
- Conclusions

Data management in life sciences

- Information explosion
- High degree of fragmentation
- Large number of heterogeneous data formats and schemas
- Identity crisis
- Lack of shared semantics
- Complexity of biological data

Semantic Web

- An extension of the current Web
- Enables navigation and meaningful use of digital resources by automatic processes
- Based on common formats that support aggregation and integration of data from diverse sources
- Tools, programming environments, specialized databases available

<http://esw.w3.org/topic/SemanticWebTools>

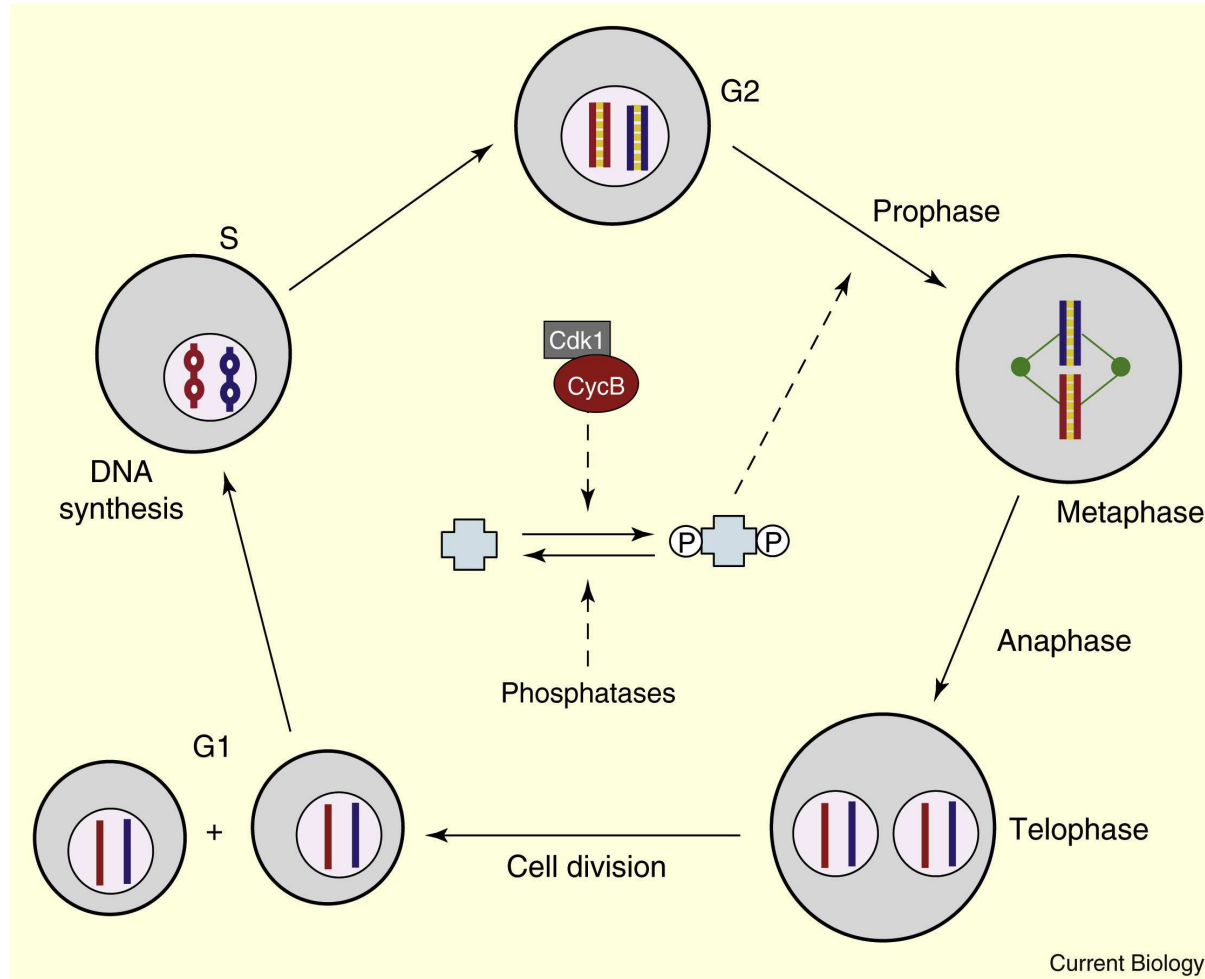
RDF

- Designed to describe Web resources
- Uses Unique Resource Identifiers (URI)
- Statements take the form of triples:
Subject - **Predicate** - **Object**
- Triples form a graph
- Graphs could be queried with SPARQL

RDFS and OWL

- Designed to store, use and exchange knowledge
- Improve interoperability
 - Flexible, extensible and decentralized
 - Self descriptive
 - Use global identifiers
- Support automatic reasoning (OWL):
 - Inference
 - Classification
 - Consistency checking

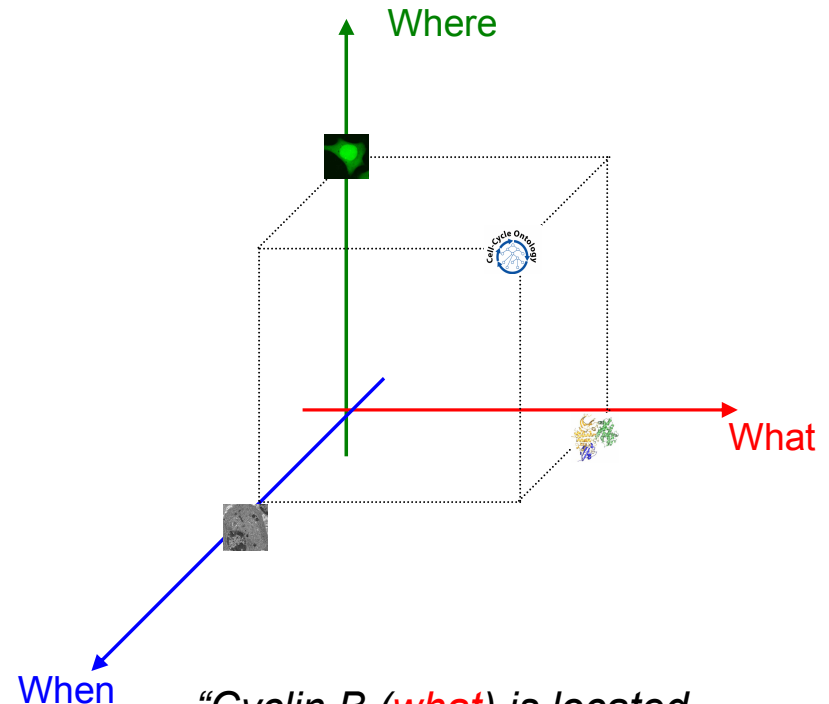
Cell Cycle



- One of the most fundamental processes
- Evolutionary conserved
- Very complex
- Health Care relevant

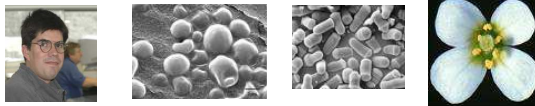
Cell Cycle Ontology

- Capture the knowledge of the CC process
- Enable advanced querying and deployment of automatic reasoners
- Promote sharing, reuse and integration with existing resources



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

ORGANISMS:

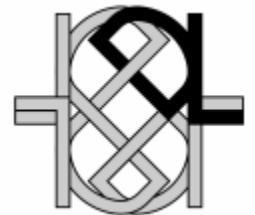
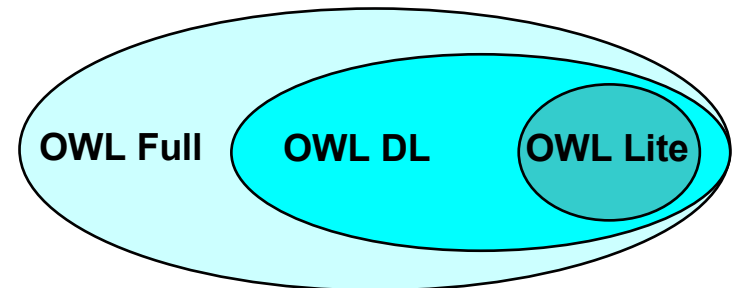


Knowledge representation

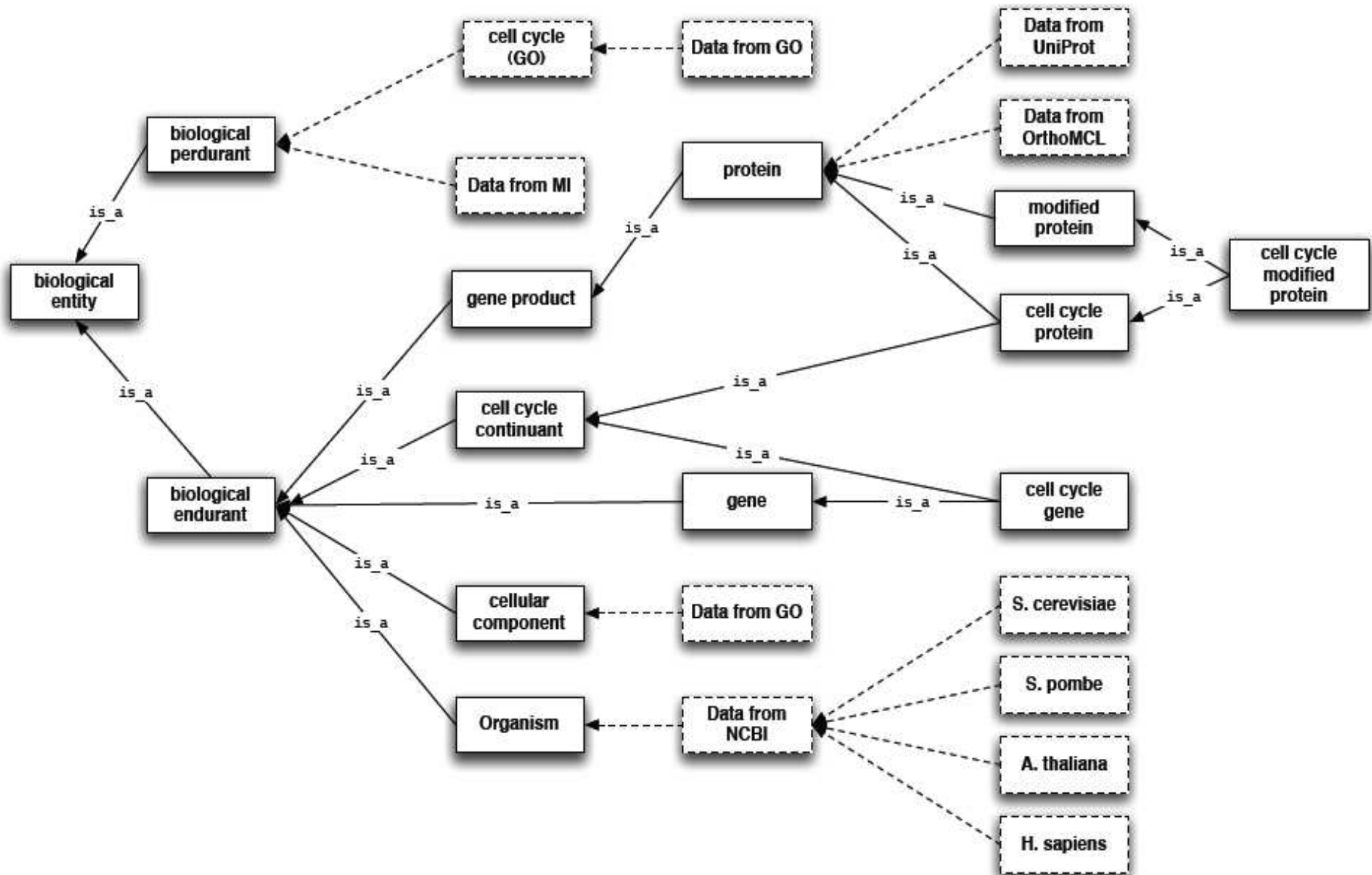
- Why OBO?
 - “Human readable”
 - Standard
 - Tools (e.g. OBOEdit)
 - <http://obo.sourceforge.net>



- Why OWL?
 - “Computer readable”
 - Reasoning capabilities vs. computational cost ratio
 - Formal foundation (Description Logics: <http://dl.kr.org/>)
 - <http://www.w3c.org/TR/2004/REC-owl-features-20040210>
 - **Reasoning:** RACER, Pellet, FaCT++

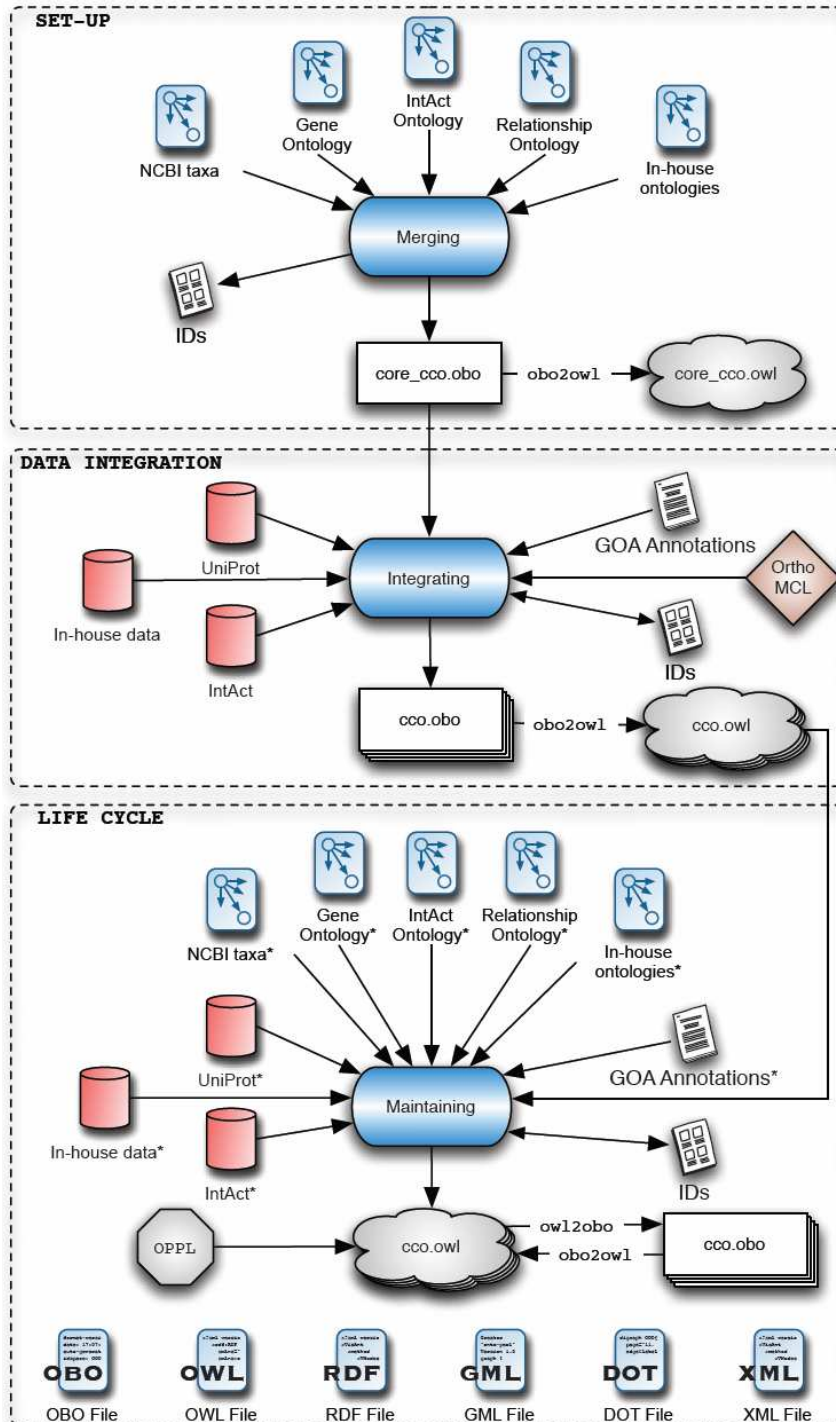


Upper Level Ontology

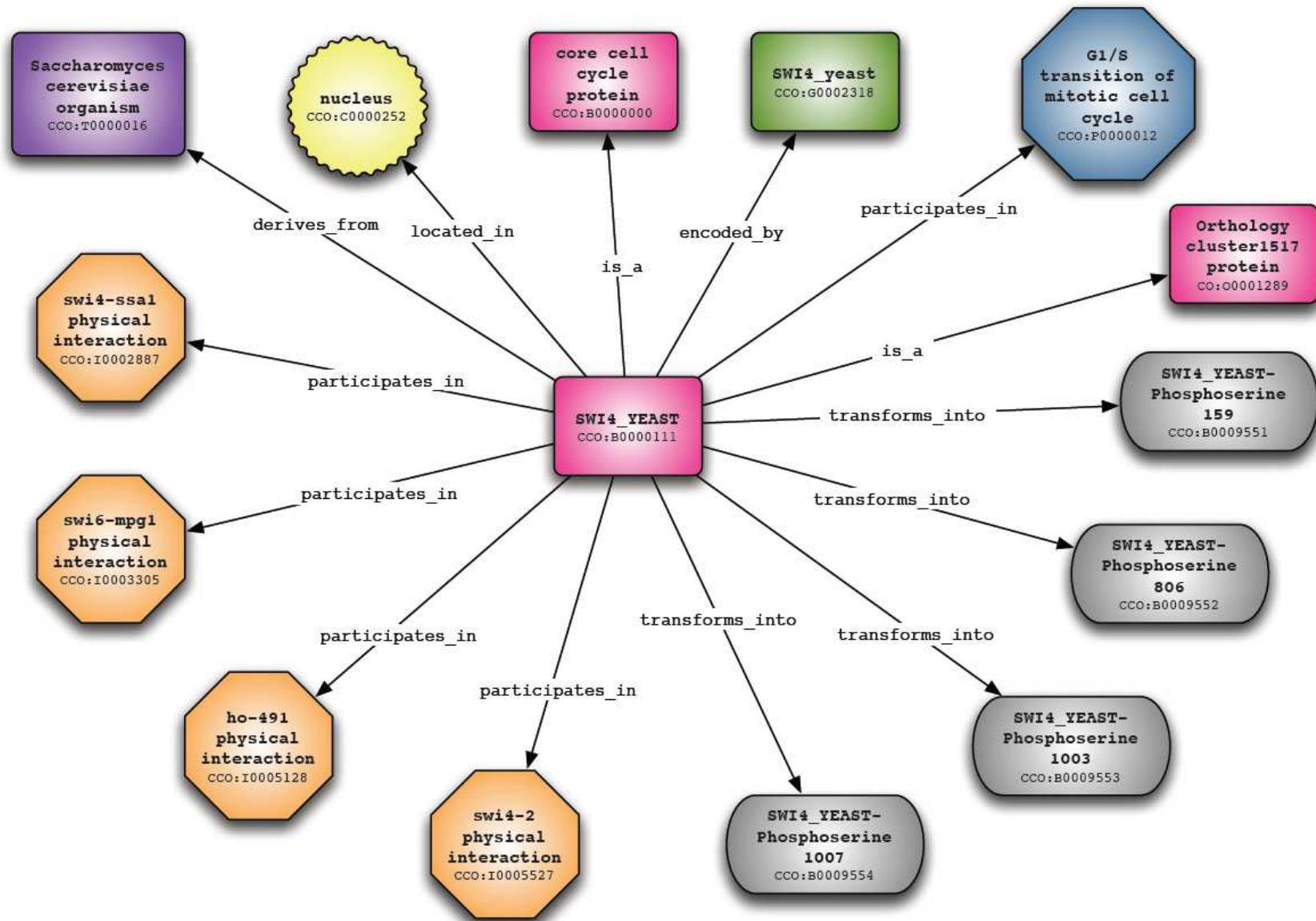


CCO Pipeline

- ontology integration
 - format mapping
-
- data integration
 - data annotation
 - consistency checking
-
- maintenance
 - data annotation
 - semantic improvement: OPPL
 - ODP (BMC BioInf – *in press*)



Current knowledge



Some figures

Entity	Ontology				
	At	Hs	Sc	Sp	CCO
Proteins	252	5829	7069	930	24541
Genes	222	1806	3148	852	6028
Interactions	76	2394	5162	399	8031
Orthology groups	—	—	—	—	1649

CCO is the composite ontology = At + Hs + Sc + Sp + orthology

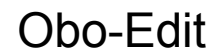
Today: over 50,000 terms in CCO

CCO main features

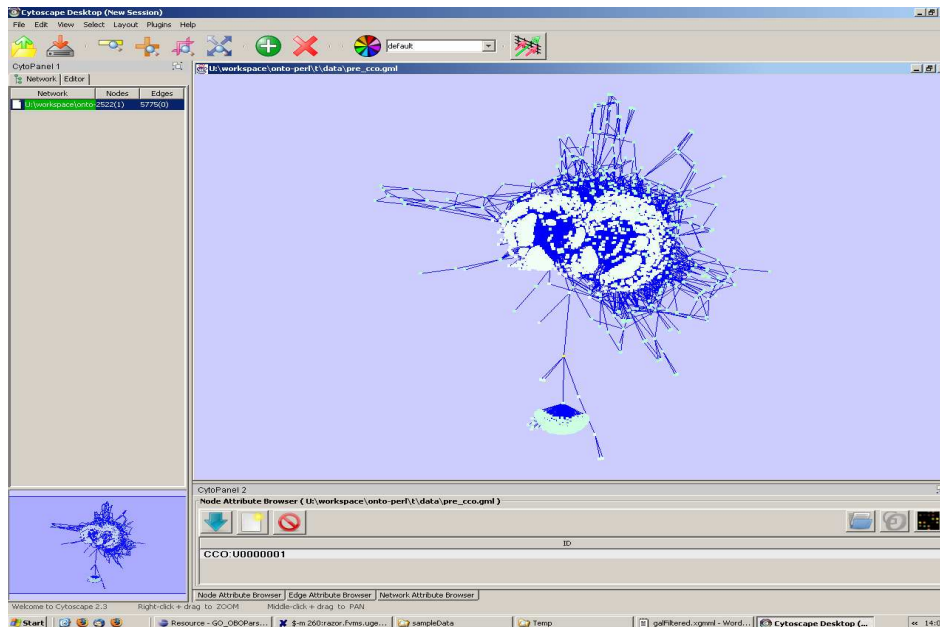
- Ontology driven
- Protein centric
- Semantic Web compatible
- Range of data formats

Tools

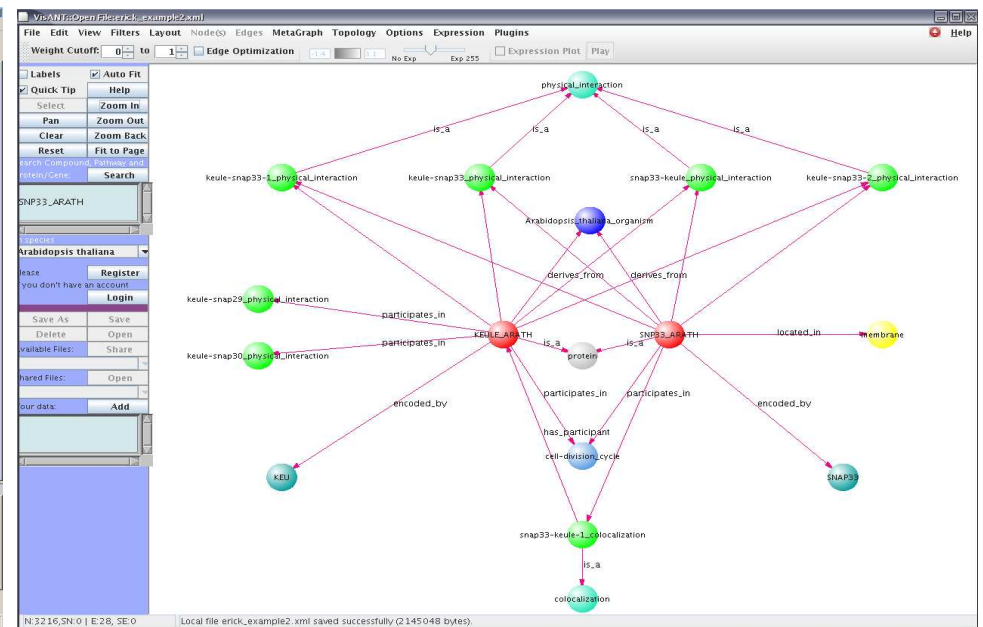
- [OBO-Edit](#) (OBOF)
- [Protege](#) (OWL, RDF)
- [Cytoscape](#) (GML)
- [Graphviz](#) (DOT)
- [VisANT](#) (XML)



Analyse graphs



Cytoscape



VisANT

Explore on the Web

THE NATIONAL CENTER FOR BIOMEDICAL ONTOLOGY
BioPortal

Version 1.0

[Browse](#) [Search](#) [Home](#) [Sign In](#) [Register](#) [Help](#)

Cell Cycle Ontology (A. thaliana)

Tree View

Tree view constructed based on is_a hierarchy

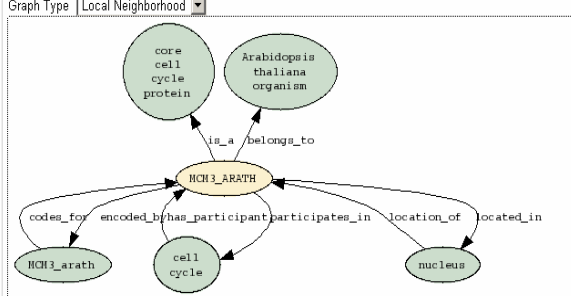
- entity
 - continuant
 - cell cycle continuant
 - cellular_component
 - gene
 - gene product
 - protein
 - cell cycle modified protein
 - cell cycle protein
 - core cell cycle protein
 - APC10_ARATH
 - ARTE_ARATH
 - ATK1_ARATH
 - ATK3_ARATH
 - ATM_ARATH
 - ATR_ARATH
 - BRE1A_ARATH
 - BSH_ARATH
 - CCA11_ARATH
 - CCA12_ARATH
 - CCA21_ARATH

Class/Type Details

| General | Attributes |
|------------------------------------|--|
| Class/Type Name: MCM3_ARATH | Definition: DNA replication licensing factor MCM3 homolog |
| Id: CCO:B0002385 | EXACT SYNONYM: "Minichromosome maintenance protein 3 homolog" |
| | Database_References: UniProt:Q9FL33 |

Graph View

Graph Type: Local Neighborhood



```

graph TD
    MCM3_ARATH -- is_a --> CoreCellCycleProtein
    MCM3_ARATH -- belongs_to --> ArabidopsisThalianaOrganism
    MCM3_ARATH -- codes_for --> MCM3_arath
    MCM3_ARATH -- encoded_by --> cell_cycle
    MCM3_ARATH -- participates_in --> nucleus
    MCM3_ARATH -- located_in --> nucleus
  
```

BioPortal

http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=CCO

EMBL-EBI

OLS - Ontology Lookup Service

CCO Ontology Browser

- entity
 - continuant
 - cellular_component
 - reference
 - organism
 - gene product
 - protein
 - core cell cycle protein
 - SEM1_YEAST
 - CDC28_YEAST
 - TBB_YEAST
 - TOP1_YEAST
 - DNLL1_YEAST
 - CDC25_YEAST
 - NOT2_YEAST
 - CDC37_YEAST
 - SC160_YEAST
 - PIS_YEAST
 - KIN28_YEAST
 - CDC7_YEAST
 - KAPA_YEAST
 - CDC31_YEAST
 - RAD1_YEAST
 - RAD52_YEAST
 - TOP2_YEAST
 - CALM_YEAST
 - RAD10_YEAST
 - IIF4E_YEAST

Help (hide)

Double-click a term to see its children. The ontology browser is populated dynamically. If there are many children for a given term, there may be a small delay while the browser fetches. Click to highlight a term to see any information associated with it. Hover over a term to see its relation with its immediate parent. Root terms will not display any relational information.

Relations

SEM1_YEAST is_a core cell cycle protein

Term Information

ID:

Name:

Zoom

Associated information

Highlight a term to view its associated information.

Term Hierarchy

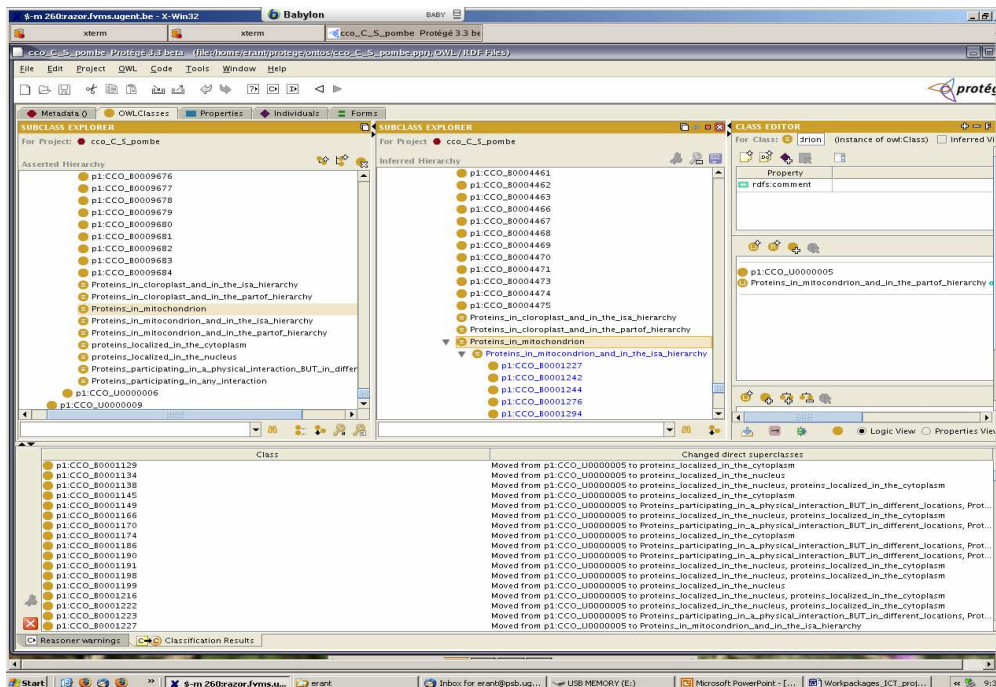
Paths to Root: ☒ Child relationships: ☐

Legend:

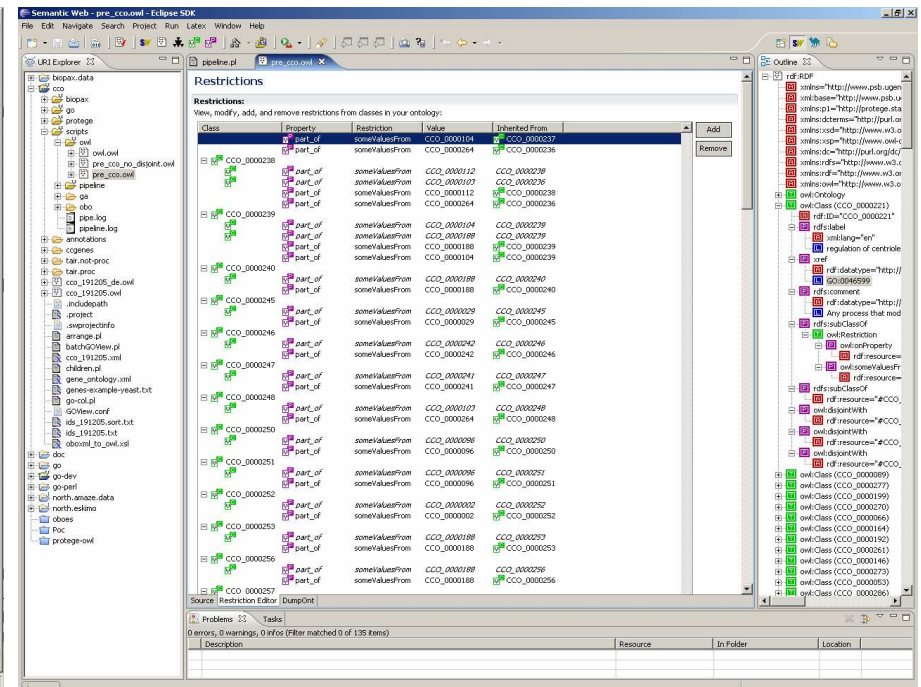
is a

Ontology Lookup Service

Check consistency

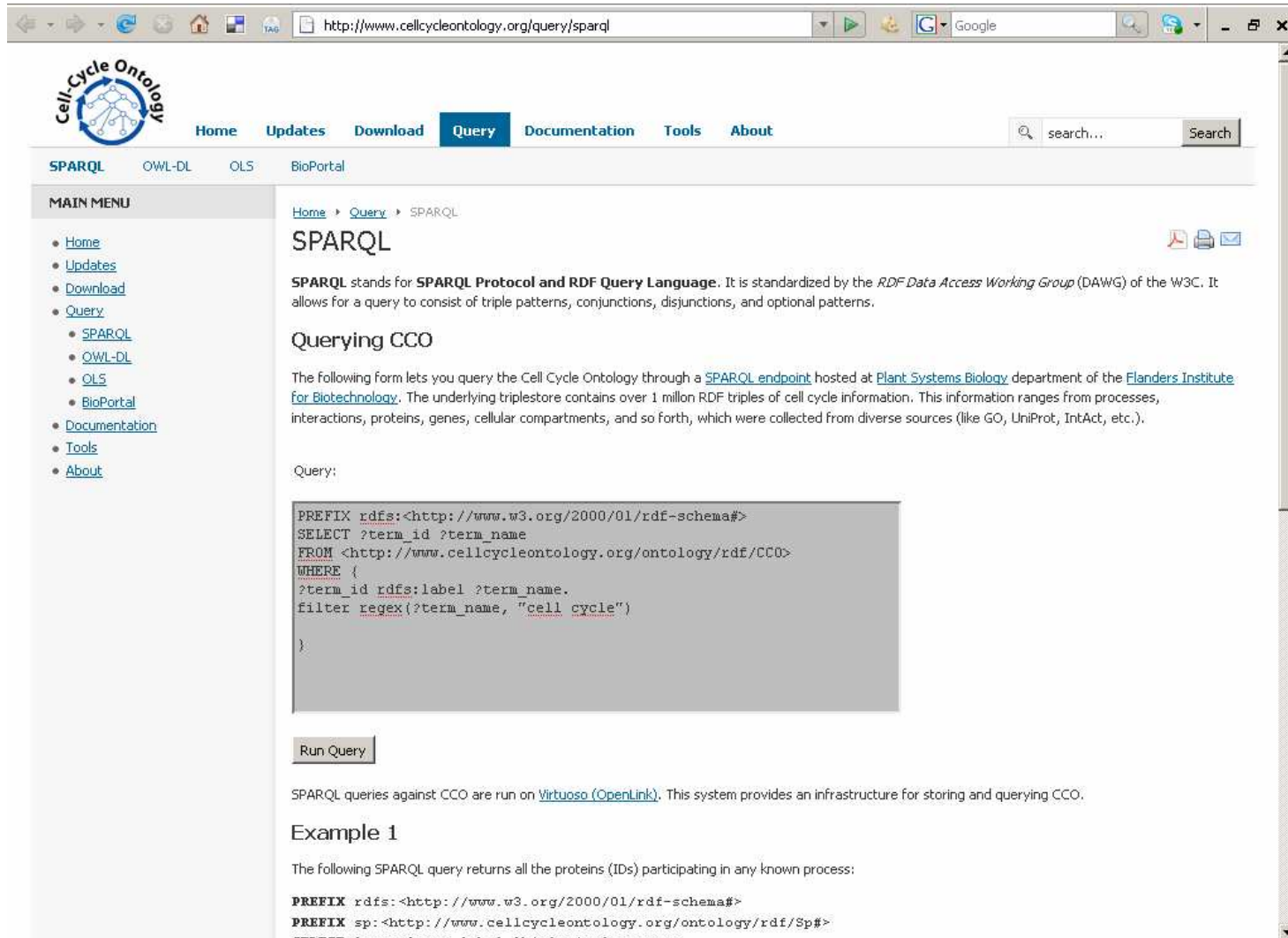


RaceroPro plug-in for Protege



SWeDE plug-in for Eclipse

Query with SPARQL



The screenshot shows a web browser window at the URL <http://www.cellcycleontology.org/query/sparql>. The page features the Cell Cycle Ontology logo and a navigation menu with links to Home, Updates, Download, Query, Documentation, Tools, and About. A search bar is located in the top right corner.

On the left side, there is a "MAIN MENU" with a list of links: Home, Updates, Download, Query, SPARQL, OWL-DL, OLS, BioPortal, Documentation, Tools, and About.

The main content area is titled "SPARQL" and includes a sub-header "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."

Below this, there is a section titled "Querying CCO" which states: "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 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.)."

A "Query:" label is followed by a text area containing a SPARQL query:

```
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
SELECT ?term_id ?term_name
FROM <http://www.cellcycleontology.org/ontology/rdf/CCO>
WHERE {
  ?term_id rdfs:label ?term_name.
  filter regex(?term_name, "cell cycle")
}
```

Below the query area is a "Run Query" button.

Underneath the button, it says: "SPARQL queries against CCO are run on [Virtuoso \(OpenLink\)](#). This system provides an infrastructure for storing and querying CCO."

The next section is titled "Example 1" and states: "The following SPARQL query returns all the proteins (IDs) participating in any known process:"

Below this, there is a SPARQL query:

```
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX sp:<http://www.cellcycleontology.org/ontology/rdf/Sp#>
SELECT ?term_id ?term_name ?term_label ?term_description
```

Conclusions

- integrated turnkey system: CCO evolving toward a one-stop shop for cell cycle researchers;
- exploratory analysis: browsing, visualizing and searching;
- querying facilities: advanced ways for retrieving data;
- reasoning exploitation: classification, consistency checking, and inferencing.

Major conclusion

CCO provides proof of concept:

the same approach could be applied to
other biological domains to develop
interoperable Knowledge Bases

Availability

- CCO web site

<http://www.cellcycleontology.org/download>

- Sourceforge.net CVS

<http://cellcycleonto.cvs.sourceforge.net/viewvc/cellcycleonto/ONTOLOGIES>

- Sourceforge.net SVN

<http://cellcycleonto.svn.sourceforge.net/viewvc/cellcycleonto/ONTOLOGIES>

- BioPortal

<http://bioportal.bioontology.org/ontologies>

Acknowledgements

- Erick Antezana (U Ghent/VIB)
- Mikel Egaña (U Manchester)
- Robert Stevens (U Manchester)
- Ward Blonde (U Ghent)
- Bernard De Baets (U Ghent)
- Martin Kuiper (NTNU)

www.cellcycleontology.org

