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.
<table>
<thead>
<tr>
<th>Project</th>
<th>Keywords</th>
<th>Technologies</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>LinkHub</td>
<td>document ranking, text categorisation, query corpus</td>
<td>RDF</td>
<td><a href="http://hub.gersteinlab.org/">http://hub.gersteinlab.org/</a></td>
</tr>
<tr>
<td>Lipid bibliography</td>
<td>lipids, metabolites, reasoning</td>
<td>OWL</td>
<td></td>
</tr>
<tr>
<td>Neurocommons</td>
<td>uniform access, package-based distribution</td>
<td>RDF, SPARL</td>
<td><a href="http://neurocommons.org/">http://neurocommons.org/</a></td>
</tr>
<tr>
<td>RDFScape</td>
<td>systems biology, cytoscape, reasoning</td>
<td>RDF, SPARL</td>
<td><a href="http://www.bioinformatics.org/rdfscape">http://www.bioinformatics.org/rdfscape</a></td>
</tr>
<tr>
<td>S3DB</td>
<td>lung cancer, omics</td>
<td>RDF</td>
<td><a href="http://www.s3db.org/">http://www.s3db.org/</a></td>
</tr>
<tr>
<td>SWAN - AlzPharm</td>
<td>neuromedicine, alzheimer, neurodegenerative disorders</td>
<td>RDF, OWL</td>
<td><a href="http://swan.mindinformatics.org">http://swan.mindinformatics.org</a></td>
</tr>
<tr>
<td>SEMMAS</td>
<td>web services, intelligent agents</td>
<td>OWL</td>
<td><a href="http://semmas.inf.um.es/prototypes/bioinformatic.s.html">http://semmas.inf.um.es/prototypes/bioinformatic.s.html</a></td>
</tr>
<tr>
<td>SOMWeb</td>
<td>distributed medical communities</td>
<td>RDF, OWL</td>
<td><a href="http://www.cs.chalmers.se/proj/medview/somweb">http://www.cs.chalmers.se/proj/medview/somweb</a></td>
</tr>
<tr>
<td>Thea-online</td>
<td>protein interactions, annotations, pathways</td>
<td>RDF, SPARL</td>
<td><a href="http://bioinfo.unice.fr:8080/thea-online/">http://bioinfo.unice.fr:8080/thea-online/</a></td>
</tr>
<tr>
<td>yOWL</td>
<td>yeast, phenotypes, interactions</td>
<td>OWL</td>
<td><a href="http://ontology.dumontierlab.com/yowl-hcls">http://ontology.dumontierlab.com/yowl-hcls</a></td>
</tr>
<tr>
<td>Project</td>
<td>Keywords</td>
<td>Technologies</td>
<td>Website</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-------------------------------------------------------------------------</td>
<td>-------------------------------</td>
<td>-------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Bio2RDF</td>
<td>mashup, linked data, global warehouse, complex queries</td>
<td>RDF, SPARQL</td>
<td><a href="http://bio2rdf.org/">http://bio2rdf.org/</a></td>
</tr>
<tr>
<td>BioDash</td>
<td>disease, compounds, therapeutic model, pathway</td>
<td>RDF, OWL</td>
<td><a href="http://www.w3.org/2005/04/swls/BioDash/Demo/">http://www.w3.org/2005/04/swls/BioDash/Demo/</a></td>
</tr>
<tr>
<td>BioGateway</td>
<td>semantic systems biology, hypothesis generation</td>
<td>RDF, SPARQL</td>
<td><a href="http://www.semantic-systems-biology.org/biogateway/">http://www.semantic-systems-biology.org/biogateway/</a></td>
</tr>
<tr>
<td>CardioSHARE</td>
<td>collaborative, distributed knowledgebase, reasoning, web services</td>
<td>RDF, SPARQL</td>
<td><a href="http://cardioshare.icapture.ubc.ca/">http://cardioshare.icapture.ubc.ca/</a></td>
</tr>
<tr>
<td>Cell Cycle Ontology (CCO)</td>
<td>cell cycle, protein-protein interactions, reasoning, ontology patterns</td>
<td>RDF, OWL, SPARQL</td>
<td><a href="http://www.cellcycleontology.org/">http://www.cellcycleontology.org/</a></td>
</tr>
<tr>
<td>CViT</td>
<td>cancer, tumor, gene-protein interaction networks</td>
<td>RDF</td>
<td><a href="https://www.cvit.org/">https://www.cvit.org/</a></td>
</tr>
<tr>
<td>FungalWeb</td>
<td>fungal species, enzyme substrates, enzyme modifications, enzyme retail</td>
<td>OWL</td>
<td></td>
</tr>
<tr>
<td>GenoQuery</td>
<td>genomic warehouse, mixed query, tuberculosis</td>
<td>RDF, SPARQL</td>
<td><a href="http://www.lri.fr/~lemoine/GenoQuery/">http://www.lri.fr/~lemoine/GenoQuery/</a></td>
</tr>
<tr>
<td>HCLS W3C</td>
<td>knowledge base, life sciences, prototype</td>
<td>RDF, OWL, SPARQL</td>
<td><a href="http://www.w3.org/2005/04/swls/hcls-kb/">http://www.w3.org/2005/04/swls/hcls-kb/</a></td>
</tr>
<tr>
<td>Kno.e.sis</td>
<td>nicotine dependence, biological pathway</td>
<td>RDF, SPARQL, OWL</td>
<td><a href="http://knoesis.wright.edu/research/semsci/application_domain/sem_life_sci/bio/research/">http://knoesis.wright.edu/research/semsci/application_domain/sem_life_sci/bio/research/</a></td>
</tr>
<tr>
<td>Linked Life Data</td>
<td>pathways, interactions</td>
<td>OWL</td>
<td><a href="http://www.linkedlifedata.com">http://www.linkedlifedata.com</a></td>
</tr>
</tbody>
</table>
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

ORGANISMS:

Users:
• Molecular biologist
• Bioinformatician /Computational Systems Biologist
• General audience

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

Knowledge representation in CCO

• Why OBO?
  – “Human readable”
  – Standard
  – Tools (e.g. OBOEdit)

• Why OWL?
  – Web Ontology Language
  – “Computer readable”
  – Reasoning capabilities vs. computational cost ratio
  – Formal foundation (Description Logics)
  – Tools (e.g. Protégé)

• OBO2OWL mapping
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)

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)
Sample knowledge in CCO
Exploring CCO (1/2)

OBO-Edit

Protégé

Cytoscape

visANT
Exploring CCO (2/2)
Advanced Querying

- **RDF = Resource Description Framework**
  - 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
- Example (matching two triples):

  ?protein sp:is_a sp:CCO_B00000000 .

  ?protein rdfs:label ?protein_label

* [http://www.w3.org/TR/rdf-sparql-query/](http://www.w3.org/TR/rdf-sparql-query/)
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 **Manders 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.). 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. (Note: 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 .
}
```

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

**Suggested PREFIXes:**
"all the core cell cycle proteins (S.pombe) participating in a known process"
<table>
<thead>
<tr>
<th>prot_name</th>
<th>biological_process_name</th>
</tr>
</thead>
<tbody>
<tr>
<td>UBC11_SCHPO</td>
<td>G2%2FM transition of mitotic cell cycle</td>
</tr>
<tr>
<td>UBC11_SCHPO</td>
<td>cell cycle</td>
</tr>
<tr>
<td>UBC11_SCHPO</td>
<td>mitosis</td>
</tr>
<tr>
<td>UBC11_SCHPO</td>
<td>mitotic metaphase%2Fanaphase transition</td>
</tr>
<tr>
<td>UBC11_SCHPO</td>
<td>regulation of mitotic cell cycle</td>
</tr>
<tr>
<td>UBC11_SCHPO</td>
<td>cyclin catabolic process</td>
</tr>
<tr>
<td>SRW1_SCHPO</td>
<td>cell cycle</td>
</tr>
<tr>
<td>SRW1_SCHPO</td>
<td>cyclin catabolic process</td>
</tr>
<tr>
<td>SRW1_SCHPO</td>
<td>activation of anaphase-promoting complex during mitotic cell cycle</td>
</tr>
<tr>
<td>SRW1_SCHPO</td>
<td>cell cycle arrest in response to nitrogen starvation</td>
</tr>
<tr>
<td>SRW1_SCHPO</td>
<td>negative regulation of cyclin-dependent protein kinase activity</td>
</tr>
<tr>
<td>DYHC_SCHPO</td>
<td>dhc1-peg1-1 physical interaction</td>
</tr>
<tr>
<td>DYHC_SCHPO</td>
<td>synapsis</td>
</tr>
<tr>
<td>DYHC_SCHPO</td>
<td>meiotic recombination</td>
</tr>
<tr>
<td>DYHC_SCHPO</td>
<td>horsetail nuclear movement</td>
</tr>
<tr>
<td>ORB6_SCHPO</td>
<td>cell morphogenesis checkpoint</td>
</tr>
<tr>
<td>ORB6_SCHPO</td>
<td>regulation of cell cycle</td>
</tr>
<tr>
<td>DED1_SCHPO</td>
<td>G2%2FM transition of mitotic cell cycle</td>
</tr>
</tbody>
</table>
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?”

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
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
  – 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
Semiotic Systems Biology Cycle

Mathematical knowledge

Consistency checking
New information to model
Model Refinement
Automated reasoning

Dynamical simulations and hypothesis formulation

Experimental design

Information extraction, Knowledge formalization

Data analysis

Experimentation, Data generation

Semantic Systems Biology Cycle
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 $A \text{ part_of } B$, and $B \text{ part_of } C$, then $A \text{ part_of } C$ 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

• **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, Inferred from Electronic Annotation.”
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
Select a query in the drop-down box

The query editor
Parameterising the query
The results appear in a separate window.

<table>
<thead>
<tr>
<th>protein_name</th>
<th>disease_description</th>
<th>interacts_with</th>
<th>encoded_by</th>
</tr>
</thead>
<tbody>
<tr>
<td>1C06_HUMAN</td>
<td>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.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NALP1_HUMAN</td>
<td>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.</td>
<td></td>
<td>ASC_HUMAN</td>
</tr>
<tr>
<td></td>
<td>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.</td>
<td></td>
<td>PYCARD</td>
</tr>
</tbody>
</table>
The result: 9 proteins

Labeled arrows to extra information

The URIs in blue.
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
- …

http://www.semantic-systems-biology.org
Extra slides
BioGateway

The homepage of SSB, including BioGateway as a first step towards this approach.
Type a query here.

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
OWL restrictions

Restriction on Nucleus: some part_of Cell

Necessary conditions vs Necessary and sufficient conditions
Sample entry in OBO

```
[Term]
id: CCO:B0002153
name: ZW10_ARATH
def: "Centromere/kinetochore protein zw10 homolog" [UniProt:O48626]
xref: UniProt:O48775
is_a: CCO:U0000005 : protein
relationship: derives_from CCO:T0000033 : Arabidopsis thaliana organism
relationship: located_in CCO:C0000251 : nucleus
relationship: participates_in CCO:P000064 : cell cycle
relationship: encoded_by CCO:B0004438 : ZW10
```
Sample entry in OWL

```xml
<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: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>
  </oboInOwl:hasDefinition>
  <rdfs:subClassOf rdf:resource="http://www.cellcycleontology.org/ontology/owl/CCO#CCO_B0000000"/>
  <rdfs:subClassOf rdf:resource="http://www.cellcycleontology.org/ontology/owl/CCO#CCO_B0000000"/>
  <owl:Restriction>
    <owl:onProperty>
      <owl:ObjectProperty rdf:about="http://www.cellcycleontology.org/ontology/owl/CCO#belongs_to"/>
    </owl:onProperty>
  </owl:Restriction>
</owl:Class>
```
OBO2OWL mapping sample
CCO checked with…

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


A reasoner (RACER) was used to identify those inconsistencies
and with…

OPPL in CCO

# Add a class called "interaction".
# Add the following necessary 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
I₁, I₂: 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?”
Systems Biology Cycle

Data analysis
Information extraction

Mathematical model

New information to model
Model Refinement

Dynamical simulations and hypothesis formulation
Experimental design

Experimentation, Data generation
Semantic Systems Biology Cycle

Biological knowledge

Information extraction, Knowledge formalization

Consistency checking
Querying
Automated reasoning

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.
    }
    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

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

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

The parameters are indicated in red.
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

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

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

**Outgoing arrows**

**Incoming arrows**
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*  
data driven

*bottom-up*  
hypothesis driven

**Biological Process**

Genome-scale functional genomics data  
Predictive mathematical model

**Statistics**  
**Mining**

**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

**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 relations $s_4$ (is_a) which has been changed into part_of (right side).

<table>
<thead>
<tr>
<th>CCO ID</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCO:P0007049</td>
<td>cell cycle</td>
</tr>
<tr>
<td>CCO:P0000096</td>
<td>centrosome cycle</td>
</tr>
<tr>
<td>CCO:P0000227</td>
<td>regulation of centrosome cycle</td>
</tr>
<tr>
<td>CCO:P0000221</td>
<td>regulation of centriole replication</td>
</tr>
<tr>
<td>CCO:P0000228</td>
<td>negative regulation of centrosome cycle</td>
</tr>
<tr>
<td>CCO:P0000222</td>
<td>negative regulation of centriole replication</td>
</tr>
</tbody>
</table>