

# Semantic Systems Biology: enabling integrative biology via Semantic Web technologies

Erick Antezana

Norwegian University of Science and Technology (NTNU),  
Department of Biology,  
Høgskoleringen 5, 7491 Trondheim,  
Norway.

[erick.antezana@bio.ntnu.no](mailto:erick.antezana@bio.ntnu.no)

Ward Blondé

Ghent University,  
Department of Applied Mathematics,  
Biometrics and Process Control,  
Coupure links 653, 9000 Ghent,  
Belgium.

[ward.blonde@ugent.be](mailto:ward.blonde@ugent.be)

Aravind Venkatesan

Norwegian University of Science and Technology (NTNU),  
Department of Biology,  
Høgskoleringen 5, 7491 Trondheim,  
Norway.

[venkates@nt.ntnu.no](mailto:venkates@nt.ntnu.no)

Bernard De Baets

Ghent University,  
Department of Applied Mathematics,  
Biometrics and Process Control,  
Coupure links 653, 9000 Ghent,  
Belgium.

[bernard.debaets@ugent.be](mailto:bernard.debaets@ugent.be)

Vladimir Mironov

Norwegian University of Science and Technology (NTNU),  
Department of Biology,  
Høgskoleringen 5, 7491 Trondheim,  
Norway.

[mironov@nt.ntnu.no](mailto:mironov@nt.ntnu.no)

Martin Kuiper\*

Norwegian University of Science and Technology (NTNU),  
Department of Biology,  
Høgskoleringen 5, 7491 Trondheim,  
Norway.

[kuiper@nt.ntnu.no](mailto:kuiper@nt.ntnu.no)

## ABSTRACT

The vast amounts of knowledge in the biomedical domain have paved the way for a new paradigm in biological research called Systems Biology, essentially an approach that relies on the integration of all available knowledge of a biological system in a single model. This approach promotes a comprehensive understanding of biological systems, driven by data integration and mathematical modelling. However, the sheer volume, variation and complexity of the current biological data pose a number of hurdles in knowledge management that need to be overcome. The Semantic Web offers various solutions to these challenges. With our initiative, named Semantic Systems Biology (SSB), we augment the systems biology approach with semantic web technologies to enable smooth data integration, rigorous knowledge representation, efficient querying, and hypothesis generation. Here we present an overview of the projects

associated with the SSB initiative. Access to our resources developed within the SSB frame is provided on our website: <http://www.semantic-systems-biology.org>.

## General Terms

Performance, Experimentation, Standardization, Languages.

## Keywords

Semantic Web technology, Bio-ontologies, OWL, RDF, SPARQL, Systems Biology, Semantics.

## 1. Rationale

The concept of the World Wide Web has revolutionised the process of publishing information across the globe and making information more accessible. However, the current web is meant for human consumption and is completely unintelligible for computers. The Semantic Web (SW) was proposed as an extension of the web [1] to enable computational processing of the information presented there. The SW is expected to promote rigorous knowledge management that naturally requires proper structures and formats, re-usability of data and machine interoperability. Ontologies lie at the foundation of this technology, which promises to meet the challenges in knowledge extraction and management by providing sophisticated frameworks to model the knowledge of a given domain. The SW uses at its core the Unique Resource Identifier (URI), and standard data representation formats defined by The World Wide Web Consortium (W3C) and the SW Interest Group (SWIG), namely the Resource Description Framework (RDF) [2] and RDF

---

\*Corresponding Author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WTMS'11, May 25–27, 2011, Sogndal, Norway.

Copyright 2011 ACM 978-1-4503-0148-0/11/05... \$10.00.

Schema (RDFS) [3], the Web Ontology Language (OWL) [4] and SPARQL (SPARQL Protocol and RDF Query Language) [5].

This evolution of the Web has been paralleled in the domain of Life Sciences by the successful development of high-throughput technologies that revolutionised and transformed biological research, resulting in a vast output of data and an exponential increase in the amount of biological knowledge. From the application of these technologies emerged a new integrative approach to biological research, named Systems Biology (SB) [6]. This approach aims at a holistic understanding of a biological system, and it hinges on the extensive use of mathematical and computational models. These models serve to integrate relevant biological knowledge about a process or a system, ultimately helping life scientists to describe and computationally simulate the behaviour of the system they study. Computational predictions of system behaviour result in new hypotheses that may be validated experimentally which in turn yields information that enables the correction and refinement of the computational models.

Data on its own remains merely as a collection of facts which need proper interpretation and integration before the knowledge represented by those facts can be elucidated [7]. The SW technologies are viewed as a solution to the knowledge integration problems in the biomedical domain; the biological community is therefore expected to be one of the early adopters of those technologies [8].

We have previously introduced a new paradigm, the Semantic Systems Biology (SSB) approach [9, 10], which aims at putting SW technologies in service of SB. Figure 1 is a schematic representation of the workflow as well as the tool/resource mapping in SSB. Firstly, biological knowledge is extracted from disparate resources and integrated into a Knowledge Base (KB) and the integrated data is checked for consistency. In the next phase the querying of the data leads to new hypotheses. Subsequently, these newly generated hypotheses serve as a basis to design new experiments to confirm or refute them. The experimentally validated hypotheses present new knowledge that can be added to the KB as facts, thereby completing the cycle of SSB. Essentially, the cycle of SSB is similar to the canonical cycle of SB [6] in the sense that in both cases the goal is hypothesis generation, the difference being that it is driven by computational modelling in SB and by computational reasoning and querying in SSB.

Several projects have been initiated under the SSB umbrella, to support the effective and efficient integration of various biological resources. Our efforts have been focused on the various tasks involved in providing the foundation for this paradigm, which essentially involves building specific knowledge repositories that house the integrated information such as BioGateway [10] – a knowledge-based system providing an efficient integration of biological knowledge, the ability to query it via a SPARQL interface and browse it via a simple web interface; and the Cell Cycle Ontology (CCO) [11] – an application ontology built to aid cell cycle researchers to elucidate knowledge about the eukaryotic cell cycle. In addition, to further tune the performance of our KBs and to make implicit knowledge explicit we have developed Metarel [12] – an RDF vocabulary for the formal annotation of ontological relations. The KBs developed under the SSB initiative are essentially RDF stores and require

data transformation tools to convert data into the RDF format. Some triple store systems (such as Open Virtuoso [13]) provide means to skip that step and “connect” those disparate resources to other relevant data without any transformation. However, a definition of mapping of data is still necessary; therefore, we opted for a scenario with a complete data transformation. To facilitate this transformation in an automated manner we have developed ONTO-PERL [14] – a Perl API developed to facilitate the handling of bio-ontologies represented in the Open Biomedical Ontologies format (OBOF) [15]. Furthermore, we have developed the ONTO-Toolkit [16] – a plug-in that enables ONTO-PERL functions from within the popular Galaxy environment [17], thereby providing a user friendly way to manipulate OBOF ontologies. The SSB initiative and associated projects are hosted and developed at the Norwegian University of Science and Technology (NTNU) [18]. In the subsequent sections we provide a brief review of the various project parts and conclude with our perspectives of merging SW and SB.

## 2. The Semantic Systems Biology Platform

We have been among the early adopters of the SW technologies in the biomedical domain. Establishing a workflow for SSB in practice has been challenging and in this section we provide an overview of the challenges we faced during the integration process.

The knowledge integration processes starts with the identification of the domain of discourse and the choice of the knowledge representation formalism. We have chosen to focus on knowledge concerning biochemical processes (regulatory or biosynthetic). Through a trial-and-error process we finally arrived at RDF as the main representation format for reasons explained below. Building a KB requires a tool to process all the required data to be transformed into RDF. Hence, we developed a Perl API suite to facilitate a seamless transformation of data into RDF.

### 2.1 ONTO-PERL and Onto-Toolkit

ONTO-PERL plays a pivotal role in the creation of our KBs. This software suite comprises an extensible set of object-oriented Perl modules to facilitate the manipulation of OBO-formatted ontologies and provides conversion utilities to various SW formats such as RDF and OWL. ONTO-PERL is used to build the core pipeline for all data transformations for our RDF stores. ONTO-PERL remains under active development where we aim for continuously improved performance and versatility. Recently also considerable changes have been made to support better modularization and smoother integration of the process of building a KB. ONTO-PERL is available at [19].

The functionality of ONTO-PERL is now also available as a plug-in for Galaxy [16, 17].

### 2.2 Metarel

Ontologies provide the necessary scaffold for a KB on which complex queries may be executed. In the biomedical domain the building of ontologies is mainly performed under the umbrella of the OBO Foundry, an initiative that has resulted in a set of standards and guidelines for bio-ontologies [20]. Hence, most ontologies are expressed in the OBO format (OBOF) which is primarily meant to be human-readable, even though there have been some attempts to perform simple automated reasoning over OBOF ontologies [21]. On the other hand, OWL has been

designed from its inception to enable computer-assisted processing. Sadly enough, we and others observed that reasoning over large OWL ontologies suffers from severe computational tractability problems [11, 22].

As an alternative, we considered RDF. RDF makes it relatively easy for domain experts to represent and integrate large amounts of knowledge but it was not originally designed to support reasoning tasks. We have developed a method to make semi-automated reasoning in RDF possible with the development of the Metarel vocabulary [12]. Metarel is an RDF vocabulary that utilizes some of the language constructs offered by OWL Full, and it provides logical semantics to relations between classes. Deployment of Metarel in combination with the SPARQL/Update language [22] on an RDF store implements a query system that makes the implicit knowledge explicit and augments the store with inferred knowledge. This is achieved through the use of five rules (closures): *Reflexivity*, *Transitivity*, *Priority over Subsumption*, *Super-relations* and *Chains*. Together, these rules allow the inference of all knowledge present implicitly in the OBO ontologies and any RDF store using those ontologies. The application of these rules boils down to reasoning in a tiny logic-based RDF language that avoids the usage of logically defined classes and reasons only through direct relations between classes with an *all-some* semantics. This effort essentially brings RDF closer to the expressiveness of OWL and sets a new paradigm: all the inferences are pre-computed before querying the KB instead of making inferences on the fly as Description Logics (DL) reasoners attempt to do.

### 2.3 BioGateway

The BioGateway KB serves as a window to distributed, generic resources. Currently, BioGateway contains about 1.8 billion RDF statements and integrates the entire set of OBO Foundry ontologies; the annotations from the complete set of Gene Ontology Annotation (GOA) [24] files; fragments of the NCBI taxonomy [25]; and SWISS-PROT [26]. The RDF store is built from scratch at regular intervals, in a fully automated way using the functionality offered by ONTO-PERL, to accommodate the latest data from the distributed resources. In this process, also the rule-based closures are computed with the use of Metarel and added to the store. The use of Metarel allows the execution of queries like “*Give me all the mammalian proteins located in the nucleus or any sub-part thereof*”, which are not possible (with a single query) in any other KB.

The ONTO-PERL translations to RDF were tested and optimised by querying and exploring BioGateway with SPARQL. This has made the store much more accessible and well integrated compared to an arbitrary upload of different RDF resource in the same RDF store. However, even with an optimised integration of RDF, productive query efforts require both an in-depth knowledge of the ontologies in BioGateway and SPARQL querying skills. For this reason the store was explored beforehand and made accessible through a library of example queries in SPARQL that are easy to parameterize.

### 2.4 Cell Cycle Ontology

Application ontologies define relevant concepts for a particular application. They are built by combining parts of domain ontologies which can be further extended according to the needs of the application in question and are embedded in KBs to facilitate data mining and hypothesis generation. Application

ontologies make use of the formalisation of domain knowledge, thereby facilitating the integration of different types of information. The utility of application ontologies has been convincingly illustrated with the development of our Cell Cycle Ontology (CCO) [27]. CCO is a knowledge management system that facilitates the analysis of the cell cycle process and was developed to serve the needs of anybody interested in this field of research, be it a molecular biologist, a cancer investigator or a student. CCO is a protein-centric ontology providing information pertaining to a given cell cycle protein such as its cellular location, molecular function, biological process it is involved in, its protein-protein interactions and orthology relations with other proteins in other species. CCO is available in various formats (including OWL-DL) to support diverse ways of exploration. In particular, the RDF version of CCO enables versatile queries about the cell cycle via a SPARQL endpoint [28].

## 3. Discussion

Despite the fact that currently many tools and resources supporting these new technologies are available [6, 7], a full implementation of the Semantic Web is still out of reach. Many factors confirm this undesired situation:

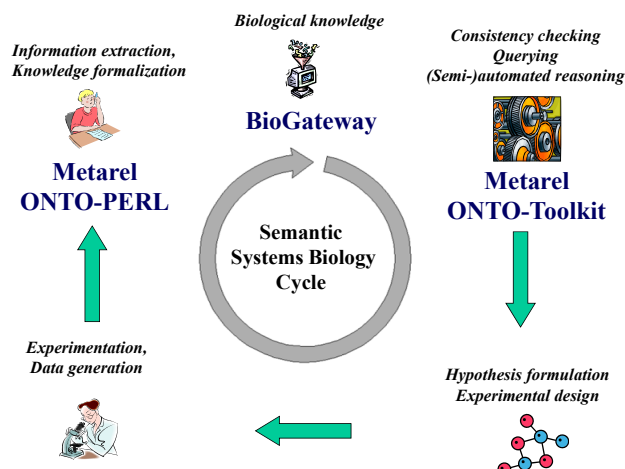
- the considerable upfront time investment,
- the limitations of triple stores,
- poor or simply non-existing visual user interfaces.

The Semantic Web bears great promise for the life science community. Computer scientists have been introducing with some success the semantic web technologies to their life-scientist colleagues, but in order to create a truly intuitive semantic web platform for the biomedical domain computer scientists need to take into account the particular needs and computational skill levels of the average life scientist. Platforms that are developed through a combination of providers ‘push’ and users ‘pull’ are bound to excel in adequate problem solving for the biomedical domain.

We have reviewed our efforts to develop a platform for semantic systems biology. As in Kitano’s original description of Systems Biology [6] two phases exist in an iterative process, corresponding to a computational, data processing phase (Figure 1) and an experimental, wet-lab phase (Figure 1, bottom). Applications and systems are evolving based on the requirements and feedback provided by the bio-community. Biological data encoding efforts are further pursued to accommodate existing data in any of the semantic web languages. Also, unification processes (such as gene name and IDs mappings) are taking place as well as ontology merging. All this has been calling for different and new ways to the biological knowledge management by providing data in formats that comply with semantic web standards. Once the technical challenges of uploading knowledge have been solved, users can validate the system. Unfortunately, they will still need the assistance of a Semantic Web specialist to be able to first translate their questions into a machine readable format (using SPARQL [5] for instance) which in turn will be able to extract relevant information from the system.

Several bio-ontologies are already incorporated in knowledge systems opening new research opportunities in SB. In particular, the EU and national funding agencies have been investing in such multidisciplinary projects (e.g. CCO). Moreover, the industry is

also exploiting such systems (e.g. Sentient from IO-informatics [29], Anatomy Lens from IBM [30]). Recent initiatives are now calling for world-wide community contributions (e.g. [31], [32]), whereas some other current initiatives are focusing on a particular domain (e.g. Neurocommons [33]). All these initiatives recognize the added value of the SW technologies for diverse domains of discourse. Successful use cases will drive further technology development to overcome conceptual, design and hardware limitations that currently limit the effectiveness and breadth of the approach.



**Figure 1: The Semantic Systems Biology cycle.** The approach depends on the biological knowledge which is gathered and integrated into a knowledge base: a) Data is checked for consistency, and querying and automated reasoning can be applied; b) hypothesis are formulated that can be used for the design of new experiments; c) the experimentation generates new data which needs to be interpreted to validate or negate the experimental hypothesis; d) interpreted data is integrated into the knowledge base, further enhancing the knowledge base for further querying and reasoning and establishing the knowledge base for further querying and reasoning and establishing a cyclical process.

## 4. ACKNOWLEDGMENTS

This work was originally funded by the EU FP6 (LSHG-CT-2004-512143) and the European Science Foundation (ESF) for the activity entitled Frontiers of Functional Genomics. WB was funded by Blondé Engineering. VM was funded by FUGE Mid-Norway. We wish to thank the HPC team at NTNU for their help in setting up the BioGateway server, the ONTO-PERL users for their feedback, and the OBO, Life Science and Semantic Web communities for interesting and motivating discussions.

## 5. REFERENCES

- [1] Berners-Lee T and Hendler J. 2001. Publishing on the Semantic Web. *Nature*. 410:1023-1024.
- [2] Resource Description Framework (RDF): <http://www.w3.org/RDF/>

- [3] RDF Vocabulary Description Language : RDF Scheme: <http://www.w3.org/TR/rdf-schema/>
- [4] OWL Web Ontology Language Overview: <http://www.w3.org/TR/owl-semantics/>
- [5] SPARQL Query Language for RDF, Oct 2010: <http://www.w3.org/TR/rdf-sparql-query>.
- [6] Kitano H. 2002 Systems biology: a brief overview. *Science*. 295:1662-1664.
- [7] Antezana *et al.* 2009. Biological knowledge management: the emerging role of SW technologies. *Brief Bioinform.* 10:392-407.
- [8] Ruttenberg *et al.* 2009 Life sciences on the Semantic Web: the Neurocommons and beyond. *Brief Bioinform.* 10: 193-204.
- [9] <http://www.semantic-systems-biology.org>
- [10] Antezana E. *et al.* 2009 BioGateway: a Semantic Systems Biology tool for the life sciences. *BMC Bioinformatics*. 10(Suppl 10):S11.
- [11] Antezana, E. *et al.* 2009. The Cell Cycle Ontology: an application ontology for the representation and integrated analysis of the cell cycle process. *Genome Biology*. 10: R58.
- [12] Blondé, W. *et al.* 2009. Metarel: an Ontology to support the inferencing of SW relations within Biomedical Ontologies. *Proceedings of the International Conference on Biomedical Ontologies (ICBO)*. 79-82.
- [13] OpenLink Virtuoso: <http://virtuoso.openlinksw.com/>
- [14] Antezana, E. *et al.* 2008. ONTO-PERL: An API for supporting the development and analysis of bio-ontologies, *Bioinformatics*. 24:885-887.
- [15] The OBO Flat File Format Guide: [http://www.geneontology.org/GO.format.obo-1\\_4.shtml](http://www.geneontology.org/GO.format.obo-1_4.shtml)
- [16] Antezana, E. *et al.* 2010. Onto-ToolKit: enabling bio-ontology engineering via Galaxy. *BMC Bioinformatics*. 11(Suppl 12):S8.
- [17] Goecks *et al.* 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 11(8):R86.
- [18] <http://www.ntnu.edu/biology/semantic-systems-biology>
- [19] <http://search.cpan.org/dist/ONTO-PERL/>
- [20] Smith *et al.* 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol.* 25:1251-1255.
- [21] Mungall C. J. 2004. Obol: Integrating language and meaning in bio-ontologies. *Comp Funct Genomics*. 5:509-520.
- [22] Holford *et al.* 2010. Using semantic web rules to reason on an ontology of pseudogenes. *Bioinformatics*. 26:i71-78.
- [23] <http://jena.hpl.hp.com/afs/SPARQL-Update.html>
- [24] Camon *et al.* 2004. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.* 32: D262 - D266.
- [25] Wheeler *et al.* 2005. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 33:D39-45.
- [26] UniProt Consortium: The universal protein resource (UniProt). *Nucleic Acids Res* 2008, 36:D190-D195.
- [27] CCO: <http://www.cellcycleontology.org>.

- [28] <http://www.semantic-systems-biology.org/cc0/queryingcc0/sparql>.
- [29] IO Informatics: <http://www.io-informatics.com/products/index.html> .

- [30] Anatomy Lens:  
<http://services.alphaworks.ibm.com/anatomylens/> .
- [31] Freebase: <http://www.freebase.com> .
- [32] OmegaWiki: <http://www.omegawiki.org> .
- [33] Neurocommons: <http://neurocommons.org> .