Biological knowledge management: the emerging role of the Semantic Web technologies

Erick Antezana, Martin Kuiper and Vladimir Mironov

Submitted: 5th January 2009; Received (in revised form): 17th April 2009

Abstract

New knowledge is produced at a continuously increasing speed, and the list of papers, databases and other knowledge sources that a researcher in the life sciences needs to cope with is actually turning into a problem rather than an asset. The adequate management of knowledge is therefore becoming fundamentally important for life scientists, especially if they work with approaches that thoroughly depend on knowledge integration, such as systems biology. Several initiatives to organize biological knowledge sources into a readily exploitable resourceome are presently being carried out. Ontologies and Semantic Web technologies revolutionize these efforts. Here, we review the benefits, trends, current possibilities, and the potential this holds for the biosciences.

Keywords: knowledge management; system biology; Semantic Web; ontology; data integration; knowledge representation

INTRODUCTION

The aim of this section is to introduce some of the most general concepts pertaining to the field of knowledge management (KM).

Data vs. knowledge

Generations of philosophers, beginning with Plato [1], spared no efforts trying to define the concept of knowledge. This definition, however, proved to be elusive and controversial [2, 3].

For the purpose of this review, it is important to make a clear distinction between data and knowledge. The concept of data came into prominence relatively recently, mainly due to the widespread use of the information and communication technologies (ICT) and the advent of modern empirical technologies that outpour huge amounts of data. Data should not be confused with knowledge—the former is just a collection of facts that require interpretation in order to be converted into knowledge. Thus, knowledge is data plus an interpretation of its meaning. Traditionally, the interpretation was carried out by a human being; however, today the interpretation of large-scale data sets is typically only possible with the help of computers because of the sheer volume of data. To give an example, consider the output of a microarray experiment. This is pure data, a matrix of labels and numbers that conveys no meaning to the human mind. A subsequent analysis of the data may reveal that a certain group of genes is over-expressed under certain conditions; if this finding would be based on experimental evidence, obtained through accepted analysis approaches and have statistical significance, this would comply with...
the conditions above and constitute a piece of knowledge. Obviously, the same set of data may afford many alternative interpretations. Therefore, the concept of ‘provenance’, keeping track of how pieces of knowledge came to be, is crucial for KM.

Another concept often used in the context of KM is information. This concept is somewhat loosely defined (in this particular context) and it finds its place between the concepts of data and knowledge, sometimes overlapping with both. Information refers to useful data—that has a potential to become knowledge within a defined context where the relations among the data elements are understood (see [4] for a recent treatment of these three concepts). The three concepts together form a paradigm of gradual understanding—data—information—knowledge [5].

What is KM?

KM is the process of systematically capturing, structuring, retaining and reusing information to develop an understanding of how a particular system (e.g. an organelle or a pathway) works, and subsequently to convey this information meaningfully to other information systems (knowledge distribution). KM has been a key component for governing the productivity and growth in the so-called knowledge-based economies. KM is now also becoming a familiar and important process for the life sciences, as exemplified by its ample coverage in scientific publications. This reflects the considerable investment of technological and human resources, which, in turn, will determine the innovation and success of KM both in academia and industry.

KM fundamentally relies on technological aspects of the so-called information society. This implies the use of resources offered by ICT, such as networks and databases, to support the knowledge handling more effectively.

One concept often used in connection with KM (and occasionally treated as synonymous to KM) is data integration. This is arguably the most important aspect of KM. It is obvious that only by combining data from various sources (typically databases) we will be able to extract maximum knowledge from the data. However, integrating data from different sources is not straightforward. The major impediment towards data integration is the so-called schema mismatch. The way that data is stored in a database (meta-data) is defined by a database schema. The schemas of different databases can be highly idiosyncratic, which makes it very difficult to interrogate multiple databases simultaneously. Another typical problem is the ‘identity crisis’—distributed resources tend to use different sets of identifiers and names for the same real-world entities.

There are two typical approaches to data integration—centralized and distributed. In the former, the schemas of the individual databases are translated into a single unifying schema, and the data are deposited into a single database (warehouse). The second approach (database federation) leaves all the data in the original sources and relies on an agreed protocol to query the data. In both cases, the user is presented with a uniform querying interface (typically SQL-based) in order to access the data from different sources. Neither of the two approaches is perfect and each has specific limitations. For example, warehouses are difficult to keep up-to-date and may be a suboptimal solution, as cautioned by the Towards 2020 science report [6], whereas querying federated databases can be rather inefficient in terms of performance (some queries may take days). In either approach an integrated solution is labour intensive, the developed solution is project-specific and cannot be reused. In any case, the presentational layer should include a powerful and user-friendly interface to hide the implementation details from the user, if it is to become widely accepted (see e.g. Louie et al. [7] for a more detailed discussion of data integration).

Knowledge representation

An adequate Knowledge Representation (KR) has been recognized as an essential prerequisite for effective KM [8–14].

A properly conceived KR formalism is expected to:

- Represent real-world entities, both tangible (e.g. cell) and intangible (e.g. mitosis), in the form of models through the process of abstraction (called surrogation in KR jargon);
- Provide a set of decisions (commitments) about how and what to observe in a particular domain of discourse (e.g. cell cycle). These commitments define a particular view on the domain in question and the way to understand it;
- Provide support for intelligent reasoning over the domain of discourse according to the needs of a particular community;
- Enable efficient organization and processing of information in a computational environment;
KR is enabled mainly by two essential elements:

- A language that can describe ideally all the facts to be modelled;
- Some modelling principles that provide a standard or widely accepted formalizing rationale;

Formal KR languages are means to ensure shared understanding and an unambiguous exchange between systems (interoperability). In general, two conditions are necessary for interoperability:

- Adoption of a common syntax: this enables applications to parse the data.
- Adoption of a means for understanding the semantics: this enables applications to use the data.

Syntax refers to a well-defined set of symbols plus a set of rules for their valid composition into aggregate structures, whereas semantics deals with the relationship between syntactic structures of a symbol system and what they are intended to denote—their intended meaning. A set of syntax rules might help us to represent information; however, without a semantic component, the meaning cannot be communicated or understood by the receptor (human or computer).

In this context, it is important to note that the two languages most often used now for information exchange, HTML and XML, are both devoid of semantics and thus do not support interpretation of the intended meaning. Therefore, a remarkable number of efforts have been undertaken to extend XML with semantics required to support particular research areas in the form of dedicated mark-up languages (e.g. CellML [16], SBML [17], VisML [18], insilicoML [19]). These languages define the semantics of the XML elements used to mark up the model structure and provide format interchange solutions [20]. A more radical solution to the problem of semantics was proposed by the W3C in the form of semantic languages such as RDF/RDFS and OWL (see subsection Semantic Web technologies for KM).

**Ontologies and KR**

Ontologies constitute the very core of the computational KR. The term ontology was originally used in a branch of philosophy (metaphysics) in which the being or existence, and its basic categories and relationships were studied. In this discipline, entities are ordered according to their similarities and differences in categories (also known as concepts) with the resulting conceptual arrangements named ontologies. This philosophical notion differs in some ways from the one that has been used by computer scientists working on KM. The difference mainly follows from their distinctive goals: in philosophy, the ordering or categorization is valued the highest whereas computer scientists attach high importance to the exploitation of such a modelling resource in diverse applications (e.g. data integration, decision support).

There are several modern definitions of the term ontology. Gruber [21] proposed a definition most often used: ‘an ontology is an explicit specification of a conceptualization’. This was refined by Guarino in 1998: ‘an ontology is a shared vocabulary plus a specification of its intended meaning’ [22]. According to this last definition, the term vocabulary corresponds to a collection of symbols (e.g. p53); and the term’s intended meaning corresponds to the entity (a.k.a. referent) the symbol refers to (e.g. the symbol p53 refers to the tumor suppressor protein 53). In real life, the relation between the two is not always one-to-one.

Formal ontologies are built using a logical framework (e.g. description logics, a branch of mathematics). In contrast, in non-formal ontologies the intended meaning is described non-rigorously (e.g. natural language). A higher degree of formality renders ontologies more machine-readable, which in turn enables a better exploitation of knowledge. Apart from this type of ontology classification, other different types of classifications [23] consider domain ontologies that model specific domains (e.g. the Gene Ontology (GO) [24] models the molecular functions, biological processes and cellular component relevant in the domain of molecular biology); application ontologies, which define relevant concepts for a particular application or use (e.g. the Cell Cycle Ontology [25] captures the concepts related to cell cycle control so that they could support a systems biology approach applied to cell cycle research); task ontologies, which model specific task-related concepts (e.g. the Multiple Alignment Ontology [26] provides a task-oriented ontology for data retrieval and exchange in the fields of sequence and structure alignment); and top-level
ontologies, which model common elements that define a generic, integrative framework for domain ontologies (e.g. the Basic Formal Ontology [27] models common objects that provide a scaffold for domain ontologies such as GO).

Ontologies are developed with the use of specialized formal KR languages (ontology languages). These languages typically formalize knowledge as networks composed of nodes linked by relations (semantic nets) [21, 28–31]. Some other types of formal languages, such as the conceptual graphs have also been used for specifying ontologies [32]. With the advent of the Semantic Web technologies, OWL (Web Ontology Language, see Subsection 3.1) has been gaining popularity as a means for developing ontologies (e.g. the Ontology for Biomedical Investigations (OBI) [33] was developed in OWL).

Ontologies serve to provide a shared conceptualization in the domain of discourse and thus the use of ontologies guarantees agreement on the semantics of the integrated information [34–36]. Therefore, ontologies are necessary (if not sufficient) for resolving naming confusions, as they are designed to unambiguously represent things that are known. Consequently, ontologies are becoming an integral part in every general KM initiative, including in the life sciences. Public and private sectors are increasingly investing in this field or at least becoming aware of the prospective advantages of using ontologies.

Beyond KR

Vast amounts of scientific information are present in an unstructured form, typically as free text, and therefore can only be understood by humans. Before this information can be converted to a formal KR format suitable for machine processing, knowledge must be extracted from free text. One way of doing that is manual curation of literature by human experts. There are a number of successful examples of large-scale manual curation in the life sciences, for example, SwissProt [37], Reactome [38] and Gene Ontology Annotation [39]. While the knowledge extracted through curation is typically of high quality, these efforts do not scale up to the current needs of the life sciences [40]. Therefore, they are being complemented by computer-assisted approaches such as information extraction and text mining [41, 42]. In spite of continuous efforts and gradual advancements, text mining remains one of the bottlenecks in scientific KM [43]. Although a significant improvement in terms of algorithms, automatic processing tools, etc. [44–46] has been witnessed, more efficiency (time) and consistency (accuracy) are needed [47]. A comprehensive coverage of this field, however, is outside the scope of this review and the interested reader is referred to recent reviews [48–50].

In the subsequent sections, we give an overview of the current state of KM in the life sciences, we introduce the tool-box offered by the Semantic Web and provide examples of the applications of these tools to the management of biomedical information, and finally outline some of the challenges lying ahead.

INFORMATIONAL LANDSCAPE IN LIFE SCIENCES

From data to knowledge

In the post-genomic era, life sciences (LS) turned into a very data-intensive domain. Therefore, scientists in this domain are facing the same challenges as in many other disciplines dealing with highly distributed, heterogeneous and voluminous data sources. However, these problems are more acute in the LS compared to many other domains due to a number of factors. Those factors could be divided into two groups. One group that could be called natural, reflects the specifics of biological data and knowledge, first of all its complexity [51]. The other group of factors could be categorized as cultural. The avalanche of data outpouring from the high-throughput genome-wide technologies caught life scientists unprepared. Traditionally, LS had no culture of using formal languages and standards and little notion of KM. The result was an abundantly idiosyncratic domain specialization. In particular, the biomedical domain is plagued with extremely high fragmentation. The number of databases in the LS is outrageous compared to many other data-intensive disciplines like particle physics, and these databases are typically autonomous and disconnected [52]. The extreme heterogeneity of data models (often of marginal quality) and formats [10, 53–55] made interoperability and data integration hardly feasible [56]. Finally, the notoriously widespread incidence of synonymy, homonymy and polysemy in the LS [57, 58] negatively affects both the precision and recall of queries over biological data. The difficulties associated with the integration of biological data have been discussed repeatedly [13, 54, 59–62].
As a result, there is a growing disparity between the wealth of data and the modicum of knowledge being extracted from these data [63, 64], even though the volume of biological data is rather moderate by the standards of some other disciplines (e.g. astronomy and particle physics). To narrow this gap, a seamless data integration is required that supports complex, detailed and targeted queries over multiple distributed resources to facilitate data analysis, hypothesis generation and experimental design [65]. Quite a number of biological data-integration efforts were undertaken to this end, and while some of them succeeded in achieving their objectives (e.g. Entrez [66], Ensembl [67], BIOZON [68], SRS [69], Integr8 [70], Galaxy [71], BioMart [72]), many failed due to a vagueness of the objectives, poorly designed user interfaces, or a lack of agreement on the semantics [51, 62, 73].

In spite of the problems with data integration outlined earlier, it is fair to say that the rigour of KM practises in LS has been steadily improving in the course of the post-genomic era. In particular, we have witnessed the emergence of many biological information standardization efforts, mainly driven by synergistic communities ranging from experimentalists to computer scientists that recognized the importance of talking the same language. Some of those efforts define standardized protocols for describing biological experiments [55, 74–81]. Others aim at the development of XML dialects for data exchange in specific sub-domains in LS, e.g. CellML [16], SBML [17], MAGE-ML [82], mzXML [83]. These efforts support sharing, dissemination and re-analysis of datasets, and they provide assistance while establishing and promoting best practices in specific technical areas.

Apart from that, formal graphical representations (e.g. the Systems Biology Graphical Notation (SBGN) [84]) and solutions based on existing standards (such as the Unified Modeling Language (UML) [85]) have also been proposed for representing knowledge formally in specific domains.

Even a more important trend is the increasing adoption of bio-ontologies as a means to provide standardized domain conceptualization in the LS.

Bio-ontologies

Bio-ontologies capture the entities and their inter-relationships within the domain of LS [86]. Currently, most of the bio-ontologies are hosted under the umbrella of the OBO (Open Biological Ontologies) foundry. The OBO foundry [87] is a collaborative effort not only for gathering bio-ontologies but also to provide a set of foundational principles [88] to structure the further co-ordinated development of bio-ontologies (e.g. ontology orthogonality, i.e. no overlap between ontologies in the foundry). The OBO foundry now constitutes a set of 53 domain-specific candidate ontologies, which are becoming widely accepted as a reference by the life science community.

Initially, the development of ontologies for LS research communities was mainly performed by domain experts possessing limited notions of formal ontology development. However, such developments are now typically supported by multi-disciplinary teams with backgrounds in biology, computer sciences and philosophy (like the Gene Ontology [24] and the Relationship Ontology [89]). Therefore, the ontology-building process takes into account a range of issues such as philosophical aspects (e.g. the partonomy paradoxes—partonomy refers to an ontological model built with ‘part of’ relations. This type of relation may have a variety of semantics so that their indiscriminate use may result in a paradox. For example, ‘Mick Jagger’s thumb is part of Mick Jagger’ and ‘Mick Jagger is part of the Rolling Stones’. Is Mick Jagger’s thumb part of the Rolling Stones?), the machine enabled functions (e.g. automatic consistency checking) and the biological applications (e.g. supporting annotation tasks). One recent trend in the development of bio-ontologies is the application of text-mining techniques for building ontologies from text [90–95].

Applications of bio-ontologies

While the most widely known and successful application of bio-ontologies is for annotating gene products with terms from GO, an ever-growing number of other applications of bio-ontologies are currently under way.

Workflow management systems (like those used by Taverna [96] and BioMOBY [97]) take advantage of ontologies to support tasks such as sequence analysis and genome annotation. BioMOBY, in particular, provides an ontology of bioinformatics data types (such as amino acid sequences, DNA sequences, SNPs) that not only specifies the semantic relationships between those data types, but also the serialization of each data type into XML. This enables a community of services that are capable of working together [98]. Likewise, ONDEX [99]
uses ontologies to link, integrate and visualize data from diverse biological data sets. The ONDEX functionality is exposed as Web services, which can be used in workflow systems like Taverna.

Text mining is another area where bio-ontologies are increasingly often used (see [44, 100] for review).

Thus, ontologies are now pervading and underpinning the field of knowledge management in the life sciences. Ontologies are also deemed indispensable for the implementation of the Semantic Web (see section Semantic Web and Life Sciences).

**SEMANTIC WEB AND LS**

Whereas bio-ontologies are constructed by rigorously formalizing knowledge contained in the scientific literature, vast amounts of omics data residing in various databases also constitute a rich resource to infer biological functions and relations between biological entities. A powerful integration of all available biological data and knowledge depends on an efficient information retrieval and management system to deal with numerous distributed resources [101] and one of the most promising proposed solutions is the use of the Semantic Web technologies [35].

**Semantic Web technologies for KM**

Semantic Web technologies [102] were designed to meet the challenge of KM in a world with highly distributed resources. The Semantic Web promises an infrastructure that comprises machine understandable content and, therefore, a world wide Web made of semantically linked data instead of a mere collection of HTML documents. Indeed, computational systems based on a semantic integration of raw data and ontological relationships will provide a sophisticated framework to interrogate and retrieve pertinent information. The Semantic Web tool-box comprises four essential components: RDF, RDFS, OWL, and SPARQL, which are described below.

The World Wide Web Consortium (W3C) developed a number of standard KR languages to describe explicitly the content underlying the current Web as well as to specify the implicit information from media resources (e.g. images, videos.). A hierarchy of such languages defines the so-called Semantic Web Stack [103], in which each layer exploits and uses capabilities of the layers below.

The simplest language in the stack is the Resource Description Framework (RDF) [104], which can be used to represent information in the form of the so-called triples: subject, predicate, object (see Figure 1). Although RDF is sufficient for many applications, it contains no mechanism for describing predicates, nor does it support the description of relationships between predicates and other resources. For this, the RDF Schema (RDFS [105]) has been developed, which is an RDF vocabulary description language, layered on top of RDF. RDFS allows the hierarchical specification of classes and properties, which, in turn, enable simple inferencing (e.g. class membership). RDFS is actually an ontology language, although its expressive power is quite limited and it is unable to support a number of commonly required features, such as negation or disjunction. The Web Ontology Language (OWL [106]), which was designed as an extension of RDF/RDFS, fulfills such types of demands and provides several constructs to create, for instance, new class descriptions as logical combinations (e.g. unions) of other classes. W3C provided three specifications for OWL (a.k.a. OWL 1): OWL-Lite, OWL-DL and OWL-Full (in the order of increasing expressiveness). A recent document [107] of W3C drafts the new Web Ontology Language (OWL 2) and its differences relative to OWL 1. OWL-DL, in particular, offers an interesting trade-off between expressiveness and computability. OWL-DL is equivalent to a well-understood fragment of Description Logics (DL) [108] (hence its name), which guarantees computability. As a consequence, a number of DL reasoners, developed by the artificial intelligence community, are currently available for deployment in the Semantic Web for: (1) consistency checking, (2) inferencing, (3) classification and (4) querying (FaCT++ [109], Pellet [110] and RACERPro [111] are among the most commonly used DL reasoners).

Semantic Web frameworks are typically conceived as being devised on the open world (OW)
assumption (the knowledge is incomplete: no inferences are drawn from the lack of evidence). This is in contrast to the closed world (CW) assumption, which allows the derivation of conclusions from the absence of evidence to the contrary (standard database implementations are the typical examples of such a paradigm) [112]. The OW paradigm seems to be particularly well suited for LS frameworks where the knowledge is typically incomplete and constantly evolving.

Another distinguishing feature of OWL is the so-called non-unique name assumption. For example, in database systems, two entries with different names are normally assumed to be different. However, in a Semantic Web framework this unique name assumption could be too stringent since many entities are known under multiple names (‘Cellular tumor antigen p53’, ‘Tumor suppressor p53’, ‘Phosphoprotein p53’, ‘Antigen NY-CO-13’, etc); therefore, the Semantic Web would assume that these names may refer to the same entity, unless stated otherwise. Instead of relying on unique names for identification the Semantic Web makes use of ontologies. For example, in a universal framework, different semantic interpretations might arise within different communities—a group working with plants may use ‘apex’ to refer to the tip of a plant organ while another group working with invertebrates may say that ‘apex’ is the tip of the spire of the shell of a gastropod. However, a proper use of ontologies will easily resolve such an ambiguity if there is an assertion in the ontology that differentiates the different parts of an organism according to the kingdom they belong to (animal vs. plant).

The fourth essential component of the tool-box of the Semantic Web is the querying language SPARQL (SPARQL Protocol and RDF Query Language) [113]. The use of SPARQL ensures that the information stored in the form of triples can be retrieved from the repositories (triple stores) with high precision and recall, similar to the use of SQL in traditional relational databases.

A number of features implemented by design in the Semantic Web languages are particularly relevant for the subject of this review:

- The Semantic Web languages are based on the mathematical graph model.
- They all rely on global identifiers (URI).
- RDFS and OWL are self-descriptive, in the sense that data and meta-data are represented in the same way and are not separated.
- RDFS and OWL are flexible, extensible and decentralized due to the hierarchical structure overlaid on the graph model.

These properties are expected to overcome some of the old nagging troubles in data integration in general: eliminate the schema mismatch problem, solve naming conflicts and make the efforts of data integration both much less labour intensive and reusable [35].

Semantic Web technologies are steadily gaining acceptance both in the public and private sectors and therefore a wide range of enabling tools is already available (see e.g. [114]). In particular, a number of storage solutions for scalable triple stores with support for SPARQL have been developed and some of them are available as freeware (e.g. Virtuoso [115], Sesame [116]).

Applications of Semantic Web technologies in LS

The following list is a non-comprehensive set of resources and projects that make use of Semantic Web technologies within the life science domain. A brief description summarizing their scope and main features is given (we refer to the respective reference for more information, see Table 1):

- The Cell Cycle Ontology (CCO) project aims at capturing the dynamics of cell cycle control by using Semantic Web technologies. It integrates data from existent ontologies (such as GO) and offers their application ontologies in diverse formats (such as RDF, OWL). A SPARQL endpoint [117] (a service that enables users (human or other) to query an RDF knowledge base (KB) via the SPARQL language) exploits the RDF files enabling the construction of complex queries over the integrated data. CCO represented in OWL enables reasoning.
- BioGateway provides a centralized KB holding data from the diverse public sources such as the GOA project and the candidate OBO foundry ontologies. The ultimate aim of this initiative is to support Semantic Systems Biology (see section Semantic systems biology). A SPARQL endpoint and a graphical interface provide access to such data.
### Table 1: Projects using Semantic Web technologies within the life science domain

<table>
<thead>
<tr>
<th>Project</th>
<th>Keywords</th>
<th>Technologies</th>
<th>Web site</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<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>
<td>[165]</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>
<td>[166]</td>
</tr>
<tr>
<td>CardioSHARE</td>
<td>Collaborative, distributed knowledgebase, reasoning, web services</td>
<td>RDF, SPARQL</td>
<td><a href="http://cardiohare.icapture.ubc.ca/">http://cardiohare.icapture.ubc.ca/</a></td>
<td>[160]</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>
<td>[168]</td>
</tr>
<tr>
<td>FungalWeb</td>
<td>Fungal species, enzyme substrates, enzyme modifications, enzyme retail</td>
<td>OWL</td>
<td><a href="http://www.lri.fr/~lemoine/GenoQuery/">http://www.lri.fr/~lemoine/GenoQuery/</a></td>
<td>[169]</td>
</tr>
<tr>
<td>GenoQuery</td>
<td>Genomic warehouse, mixed query, tuberculosis</td>
<td>RDF, SPARQL</td>
<td><a href="http://www.w3.org/TR/hsds-kb/">http://www.w3.org/TR/hsds-kb/</a></td>
<td>[171]</td>
</tr>
<tr>
<td>HCLS W3C</td>
<td>Knowledge base, life sciences, prototype</td>
<td>RDF, SPARQL, OWL</td>
<td><a href="http://knoesis.wright.edu/research/semsci/application">http://knoesis.wright.edu/research/semsci/application</a>.</td>
<td>[172]</td>
</tr>
<tr>
<td>Kno.e.sis</td>
<td>Nicotine dependence, biological pathway</td>
<td>OWL</td>
<td><a href="http://www.semantic-sci/bio/research/">http://www.semantic-sci/bio/research/</a></td>
<td>[173]</td>
</tr>
<tr>
<td>Linked Life Data</td>
<td>Pathways, interactions</td>
<td>RDF</td>
<td><a href="http://hub.gersteinlab.org/">http://hub.gersteinlab.org/</a></td>
<td>[174]</td>
</tr>
<tr>
<td>LinkHub</td>
<td>Document ranking, text categorization, query corpus</td>
<td>OWL</td>
<td>[128]</td>
<td></td>
</tr>
<tr>
<td>Lipid bibliosphere</td>
<td>Lipids, metabolites, reasoning</td>
<td>RDF, SPARQL</td>
<td><a href="http://neurocommons.org/">http://neurocommons.org/</a></td>
<td>[175]</td>
</tr>
<tr>
<td>Neurocommons</td>
<td>Uniform access, package-based distribution</td>
<td>RDF, SPARQL</td>
<td><a href="http://www.bioinformatics.org/rdfscape">http://www.bioinformatics.org/rdfscape</a></td>
<td>[132]</td>
</tr>
<tr>
<td>RDFScape</td>
<td>Systems biology, cytoscape, reasoning</td>
<td>RDF</td>
<td><a href="http://www3db.org/">http://www3db.org/</a></td>
<td>[176]</td>
</tr>
<tr>
<td>S3DB</td>
<td>Lung cancer, omics</td>
<td>RDF</td>
<td><a href="http://www.mindinformatics.org">http://www.mindinformatics.org</a></td>
<td>[177, 178]</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>
<td>[177, 178]</td>
</tr>
<tr>
<td>SEMMAS</td>
<td>Web services, intelligent agents</td>
<td>RDF</td>
<td><a href="http://semmas.inf.um.es/prototypes/bioinformatics.html">http://semmas.inf.um.es/prototypes/bioinformatics.html</a></td>
<td>[179]</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>
<td>[129]</td>
</tr>
<tr>
<td>Thea-online</td>
<td>Protein interactions, annotations, pathways</td>
<td>RDF, SPARQL</td>
<td><a href="http://bioinfo.unice.fr:8080/thea-online">http://bioinfo.unice.fr:8080/thea-online</a></td>
<td>[180]</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>
<td>[181]</td>
</tr>
</tbody>
</table>
The Bio2RDF project aims at providing a system that would enable SPARQL queries over the Semantic Web to answer questions that are currently not possible or practical. This web of linked data will mimic a global data warehouse even if it is distributed over the Internet. Bio2RDF gathers documents from public bioinformatics databases (such as KEGG, PDB, MGI, HGNC and several of NCBI’s databases).

NeuroCommons is an ambitious project that aims at making all scientific research material (such as articles) available and usable. Currently, it provides some resources in RDF (such as GO), which can, in turn, be loaded in triple-based stores enabling querying. This project also fosters the standardization of biological resources.

The Health Care and Life Sciences Interest Group (HCLS W3C) has built a prototype KB for the LS. This prototype is seen as the original major effort that used Semantic Web technologies for the benefit of the LS and inspired many other projects (e.g. BioGateway).

The Linked Life Data project bears on the achievements of the LarKC [118] project and uses SW technologies to integrate multiple biomedical resources (UniProt, Entrez Gene, GO, NCBI taxonomy, BioGRID, Reactome, National Cancer Institute - Pathway Interaction Database, The Cancer Cell Map, BioCyc Database collection, KEGG).

CardioSHARE presents a prototypical framework for querying distributed data taking advantage of Semantic Web technologies. Their SPARQL query engine enables the required data to be retrieved dynamically from Web services. This project was originally focused on the analysis of clinical data on heart disease; however, its design leaves room for integrating any type of biological data. CardioSHARE [119] is based on the BioMOBY experience and success. It provides a prototype that shows how a distributed life-science Semantic Web could be achieved by bringing machine-readable semantics to Web services [120].

RDFScape aims at bridging Semantic Web technologies and Systems Biology. This Cytoscape plug-in allows querying, visualizing and reasoning over ontologies represented in RDF or OWL.

GenoQuery presents a relational genomic warehouse. It provides a graphical interface as well as a SPARQL-based one to query the data. GenoQuery collects mainly data from UniProt, KEGG and prokaryote genome-related data.

SEMMAS is another initiative that combines Semantic Web technologies with multi-agent systems for integrated access to biological resources.

Other projects also demonstrated the power of using Semantic Web technologies such as the KNO.E.SIS [121], which integrate data from Entrez Gene and HomoloGene and three pathway resources: KEGG, Reactome and BioCyc.

As it can be noted, most of the projects/efforts just described share a common, challenging objective: answer complex questions over diverse sources of biological data. To that end, many of them opt to first collect disparate data and transform it into RDF (or OWL in some cases, e.g. in CCO), then this data is stored in appropriate repositories (such as Virtuoso), which, in turn, enables the formulation of queries that could combine several sources. The construction of those systems not only shows that the usage of Semantic Web technologies in biology can be a significant advancement but also demonstrates that they can be easily understood and adopted in spite of some limitations (e.g. performance).

While the projects described earlier clearly attest the high potential of Semantic Web technologies for solving data integration problems in LS, most applications so far use a centralized approach mainly because of the performance limitation. However, the available solutions, such as Virtuoso, do support querying distributed resources and, in the future, this will become a preferred way of data integration as the performance improves.

Semantic Systems Biology
As initially argued by Ludwig von Bertalanffy, biological systems properties cannot be described accurately by means of context-free elements [122]. This notion is now pervading the LS and has led to the widely embraced integrative discipline of systems biology [123–127]. Essentially, systems biology hinges on a mathematical or computational model that allows the simulation of the behaviour of complex biological systems. Simulations will help to validate the model, but also to predict the behaviour of the system under new conditions and to come up with new hypotheses than can be validated experimentally.
Ruttenberg et al. [35] predicted that systems biology would be among the earliest adopters of the Semantic Web technologies due to its highly integrative interdisciplinary nature. Indeed, we can see already how the methodology of systems biology is now becoming supplemented with the Semantic Web technologies [128-132] to help build high-quality models of biological systems. We have named this fusion of systems biology and Semantic Web technologies: Semantic Systems Biology [133].

Building a biological knowledge system is a crucial first step to enable Semantics Systems Biology. As we explained, ontologies delineate the structure of such knowledge systems. Despite the current limitations posed by limited expressive depth, reasoning paradigms, and database performance, the powers of semantic technologies have already been illustrated in a number of interesting use cases [25, 130, 134, 135]. In these cases, the cyclic workflow of a Semantic Systems Biology approach is followed as depicted in Figure 2. As in Kitano’s original Yin and Yang of Systems Biology [123], two phases exist in an iterative process, highlighted within the cycle. These phases correspond to a computational, data processing phase (top) and an experimental, wet-lab phase (bottom). Applications and systems are evolving based on the requirements and feedback provided by the bio-community. Biological data and knowledge are encoded in any of the semantic web languages. This enables unification processes (such as gene name and IDs mappings) as well as ontology merging. In this way, efficient data and knowledge exploitation is assured. All this opens new ways to biological knowledge exploitation through compliance with Semantic Web standards.

**FUTURE CHALLENGES**

**Preamble**

We have introduced the Semantic Web technologies as the most promising approach to solving the current challenges in the management of life science knowledge. Yet, it is important to keep in mind that despite the advances the Semantic Web has made, the original vision [136] from Tim Berners-Lee still remains to a large extent unrealized: infrastructure scalability (e.g. underlying consistency checking over massive and evolving resources), community commitment (mostly groups working with Web technologies dedicate efforts to develop semantically-enriched systems), semantic enrichment of content (few databases having data annotated semantically), and a global unique identification of resources are still very much in the discussion phase. These challenges are currently being tackled by multi-disciplinary research teams (e.g. the Semantic Web Health Care and Life Sciences Interest Group (HCLSIG) [137]) striving to solve the problems on the way towards the realization of the Semantic Web vision.

If the Semantic Web is to become a critical technology pillar in the LS, two elements are essential: (1) a technological infrastructure for data representation, exchanging and querying, and (2) a standard convention for the identification of biological entities and relationships, for instance, based on uniform resource identifiers (URI), underpinning the above-mentioned infrastructure. RDF and OWL provide the key components for the infrastructure, but the identification of biological entities and relationships has not yet been consolidated despite some proposals (such as URIs-based [138], LSIDs [139], OKKAM IDs [140] and MIRIAM URIs [141]). Only when a consensus has been reached on a commonly accepted identification scheme will it be possible to refer unambiguously to specific biological resources.
Towards HCLS 3.0

Alongside the steady progress towards the Semantic Web, the WWW has also been evolving in the direction of the so-called Web 2.0. The Web 2.0 refers to the use of the Internet as a collaborative environment, thus transcending the traditional Web (Web 1.0), which is a static content web. The concept of Web 2.0 has been rather satisfactorily adopted by biologists and bioinformaticians [142]. It enables a better scientific sharing and communication in comparison with the Web 1.0. Within this arena, one of the most widely used tools is the so-called wiki, with many examples to demonstrate its power (e.g. Proteinpedia [143], WikiProteins [144], OpenWetWare [145], YeastPheromone Model [146]). Recent initiatives are now calling for worldwide community contributions (e.g. Mons et al. [147]). Another typical Web 2.0 tool, blogs, has been widely adopted by the scientific community for the purpose of scientific communication, e.g. nature.com blogs [148] or The Seven Stones—The Molecular Systems Biology Blog on Systems & Synthetic Biology [149]. Yet another approach being adopted within the Web 2.0 community concerns mashups, web applications for flexible and light-weight integration of distributed Web resources on demand (see [150] for LS examples).

The vision of the future Web 3.0 combines both the advances of the Web 2.0 and the ideas of the Semantic Web [51, 150, 151] and will lead eventually to HCLS 3.0 as an implementation thereof for the LS.

Reasoning over complex biological information

Even though some successful cases of the application of automated reasoning to biological data were reported [129, 152–154], limitations still exist at several levels:

- **Popularity**: the technology and available tools are still esoteric for many prospective users and only a few research groups have applied it on real data.
- **Performance and scalability**: current infrastructures only allow for handling relatively small KBs, whereas the handling of complex, massive and distributed resources, a situation more typical for the LS, still presents problems. Even minor changes in the problem size (e.g. new data integrated, more assertions included) can result in a serious deterioration of performance [155].

- **Data complexity**: the inherent intricacy of biological data presents serious challenges in representing [51] or modelling [156] such data (e.g. temporal aspects of biological processes).
- **Data fuzziness**: Biological data are often imprecise and incomplete, typical examples being transcriptomics and proteomics data. Although some studies [157–159] have shown how fuzzy set theory [160] may enhance capturing and modelling imprecise biological data, large parts of the available raw data remain under-exploited.

Biological knowledge evolution

Biological ‘facts’ are constantly changing. Biological experiments yield new insights, which will eventually be included in KBs. However, a new piece of information may destabilize the existing set of accepted facts by creating an inconsistent KB. The new knowledge could refute previously ‘accepted’ information. Thus, it is important to keep in mind that the accepted ‘facts’ held by a KB may be incomplete, incorrect or simply false. Therefore, an adequate support for keeping track of the knowledge evolution is needed to ensure an acceptable degree of knowledge consistency and provenance. A number of theoretical frameworks [161–164] have been proposed for dealing with this issue and should be considered in KB implementations.

CONCLUSIONS AND RECOMMENDATIONS

The LS need a versatile and comprehensive knowledge integration framework [101]. Although the concept of a web portal is useful, and several portals with significant amounts of data and information are currently in use, it is rather difficult to ask simple questions like, for example, which human diabetes-related proteins are located in the nucleus (or a part of it) and are interacting with proteins related to pancreatic cancer. Such questions require the integration of data from multiple sources. The Semantic Web provides some of the necessary technologies for integrating distributed resources. The knowledge resources integrated in this way support advanced querying and may even allow the deployment of automated computational reasoners to
generate new hypotheses about the functionality of biological systems, thus enabling a new concept that we have named Semantic Systems Biology. Some current initiatives have already gone quite a way forward in this direction by providing the first concrete steps towards interoperable and universal knowledge architecture.

The development of standards for representation and exchange of biological data has been largely driven by the evolving needs of the LS. Research is still needed to harmonize existing efforts so that a unique, interoperable, universal framework be exhibited to the end users. To achieve that, it is important that researchers who generate data and report new findings work closely together with knowledge management experts to build scalable data repositories that allow handling the contents of those repositories in a natural and efficient manner. Finally, although the Semantic Web infrastructure has not yet been implemented in full, a number of significant advances towards it have been achieved, mainly in the form of middleware (e.g. reasoners, triple storage repositories). These achievements pave the way for the development of successful applications that allow the life-science community to take full advantage of the integrative and analytical potential of knowledge management.

**Key Points**

- Data vs. knowledge: there is a growing disparity between the massive production of biological data and the modicum of knowledge being extracted from these data.
- An adequate knowledge representation has been recognized as the cornerstone for effective knowledge management.
- Semantic Web technologies standardize both data identification and knowledge representation.
- Research is still needed to harmonize existing efforts so that a unique, interoperable, universal framework arises.
- All this leads to future uses of computers for heterogeneous data integration, querying, reasoning and inference, which in turn will support knowledge discovery.

**Acknowledgements**

We thank the editors and three anonymous referees for providing constructive suggestions. We also thank Cecilia Rodriguez for preparing the figures.

**References**


105. RDF Vocabulary Description Language 1.0: RDF Schema 2009. [http://www.w3.org/TR/rdf-schema/ (10 April 2009, date last accessed).](http://www.w3.org/TR/rdf-schema/ (10 April 2009, date last accessed).)

106. OWL Web Ontology Language Reference 2009. [http://www.w3.org/TR/owl-ref/ (10 April 2009, date last accessed).](http://www.w3.org/TR/owl-ref/ (10 April 2009, date last accessed).)


The emerging role of the Semantic Web technologies


