DDT: BIOSILICO Vol. 2, No. 6 November 2004

# What the semantic web could do for the life sciences

Eric K. Neumann, Eric Miller and John Wilbanks

**Scientific research is predicated on the effective exchange of knowledge. The effective exchange of data and accompanying interpretation underpin new hypotheses and experimental designs, typically followed by a community-based process of debate and rebuttal. This community-driven process clarifies and strengthens the elements of facts and hypothesis. Within the life sciences, the result of this process is a collective understanding of emerging biological viewpoints. The methodologies for community debate and knowledge transfer have changed little over the past twenty years, although both scientific instrumentation and publishing technologies have undergone revolutionary change. It is proposed that newly published recommendations from the World Wide Web Consortium (W3C), which handle the domain and process-specific semantics of life sciences, would better support the application of peer-reviewed knowledge in discovery research. W3C semantic web technologies support flexible, extensible and evolvable knowledge transfer and reuse, enabling scientists and their organizations to increase efficiency across the scientific process.**

**Eric K. Neumann**
Aventis Pharmaceutics
Bridgewater
NJ 08807, USA
e-mail:
eric.neumann@aventis.com
**Eric Miller**
**John Wilbanks**
W3C, Room 32-G515
32 Vassar St
CSAIL-MIT
Cambridge
MA 02139, USA

▼ Informatics has played a pivotal role in the life sciences over the past decade, and will continue to do so increasingly as a result of the reliance on new emerging technologies that enhance scientific progress. Much has recently been achieved through the computational analysis and assembly of genomic sequences [1], as well as quantitative gene expression data [2]. However, scientists still cannot use this information to record their interpretations of the data, and distribute it among their colleagues. We are witnessing a major transition whereby informatics will not merely store and analyze data, but will also represent and capture the interpretation in a concise and expressive way based on common knowledge and hypothesis formation. This could have immediate consequences on several specific aspects of biological data management: (i) integrating heterogeneous data through common explicit semantics; (ii) applying logic to infer new insights and to propose and/or capture new hypotheses; (iii) expressing rich and well-defined models of biological systems; (iv) annotating findings and interpretations formally (semantically), and sharing with other researchers and their informatics groups; and (v) embedding models and semantics directly within online publications.

Current research has demonstrated that knowledge can be assembled from structured and unstructured sources, and represented in a machine-readable format that is web-compatible using resource description framework (RDF) [3], (http://www.w3.org/RDF/). RDF is a World Wide Web Consortium (W3C) [4] standard that is used to define sets of relations between data and concepts. It is the cornerstone for the semantic web (http://www.w3.org/2001/sw/) [4] that will enable anybody to clearly and commonly define the concepts and logic within any document. Any information expressed in RDF can be connected to any other information expressed in RDF, in much the same manner that any document expressed in HTML can link to any other document expressed in HTML. In this way, the expression of discrete facts in RDF makes them available for sharing and analysis by the scientific community. RDF has the potential to let scientists apply all available knowledge to decision-making, including target prioritization, assessment of compound liabilities and clinical trial design.

Despite recent advancements in technologies that generate and analyze data, methods for interpreting data and sharing its derived knowledge are very much human-based (e.g. meetings, email, report writing and thematic reviews) and have fundamentally remained unchanged. Although information technologies have brought tremendous efficiencies in


1741-8364/04/$ – see front matter  PII: S1741-8364(03)02420-5

the sharing of life sciences information (e.g. world wide web), these for the most part document accessibility efficiencies, and synthesis, interpretation and application of knowledge still occur exclusively in the minds of the scientist and are disseminated in unstructured forms.

Owing to the exponential growth and complexity of life sciences knowledge, in addition to the use of cross-disciplinary information, there is now an unqualified need for computer-based systems to support the logical interpretation and association of life science knowledge in a more manageable form. No place is this more evident than within the pharmaceutics industry, which has some of the most sophisticated data-generating and data-analyses technologies of any group. Drug discovery has been hampered by the fact that, although the list of new potential drug targets has been growing thanks to genomics, the list of well-characterized targets has been decreasing [**5**]. Consequently, the number of new drug applications per year has not increased, but has remained flat. This is having dire consequences for healthcare, and the future roles of informatics and knowledge discovery are further described in this context.

## Data integration (over) emphasis

Over the years, there have been numerous discussions on challenges in life science informatics [6,7]. The two principal issues most cited are: (i) an increasing volume of data generated by new technologies at an unprecedented rate, which require systems with massive storage and throughput capacities; and (ii) a heterogeneous and complex assortment of data, which warrant better means for handling multiple data-types and is supported by interoperable software. Many vendors have emphasized these points repeatedly and have offered major solutions to address these problems. Although these are important issues, this viewpoint on data format ignores the real issues of data utility for most biological research and biotechnology – the consolidation and utilization of diverse yet relevant knowledge for scientific insight and maximum value realization. It is not merely a problem of storing and accessing data, but how scientists perceive meaning around data and how they can intelligently and clearly relate this to other scientists. In the end, the goal is more about knowledge aggregation than data integration.

Representations for handling complex biological data have been discussed by many bio-standards groups, such as Object Management Group (OMG-LSR; http://www.omg.org/lsr), Microarray Gene Expression Database (MGED; http://www.mged.org/), Information Infrastructure Interoperability Consortium (I3C; http://www.i3c.org), Bio-Ontologies (http://www.cs.man.ac.uk/~stevensr/meeting03/) [8], Open-Bio (http://www.open-bio.org) [9], CDISC (http://www.cdisc.org/), and BioPAX (http://www.biopax.org) All these groups aim to provide life science and clinical informaticists with data standards that improve the development of interoperable technologies (e.g. XML and web services). To date, full implementations of such standards efforts have been only partially successful, even when specifications have been produced. In almost all cases, these efforts focus on helping the informaticists directly rather than scientists, yet it is the scientists who in the end would review the informatics analysis. More so, within drug discovery, the current efforts are directed primarily at the early discovery phase, which comprises <10% of the overall drug discovery process.

## Ontologies uniting across domains

Heterogeneous data often need to be integrated together through a common framework to bring potentially related data into proximity that can be mined for deeper insights. Several approaches exist including schema merging (data warehouse), federating databases (multiple databases that are linked via an interconnected query mediator) and common-model indexing (a relational graph database with unique data identifiers for multiple databases) [10]. However, the meaning, or semantics, of why certain data elements are linked to each other through some form of data-definition language (DDL) is not always made explicit (e.g. the use of foreign keys in relational databases, or wrapper multiplexing) (Figure 1). Because it is the meaning around data that most scientists need to understand when interpreting experimental data, it is necessary for information systems to include science-specific semantics and to do so in an extensible manner.

This is where ontologies come in because they allow scientists to specify to any degree of resolution, how data, terminology (i.e. controlled vocabularies), concepts and ideas all relate to each other. Ontologies are an organized system of concepts and the relations between them, along with their logical constraints. For the most part, ontological knowledge systems are either frame based, graph based or description logics (DL). Several examples of research ontological systems have been developed and include Ontolingua, Loom and TAMBIS [11–13]. Within these systems, individual ontologies can be created and applied over a broad range of topics. In addition, an open-knowledge application-programming interface (API) {open knowledge base connectivity (OKBC) [14]} and a knowledge interchange format (KIF; http://logic.stanford.edu/kif/) exist, but both are defined specifically with frame-based systems in mind. The recently completed ontology web language (OWL; http://www.w3.org/2001/sw/WebOnt/) specification from the W3C supports the exchange and use of all forms of ontologies

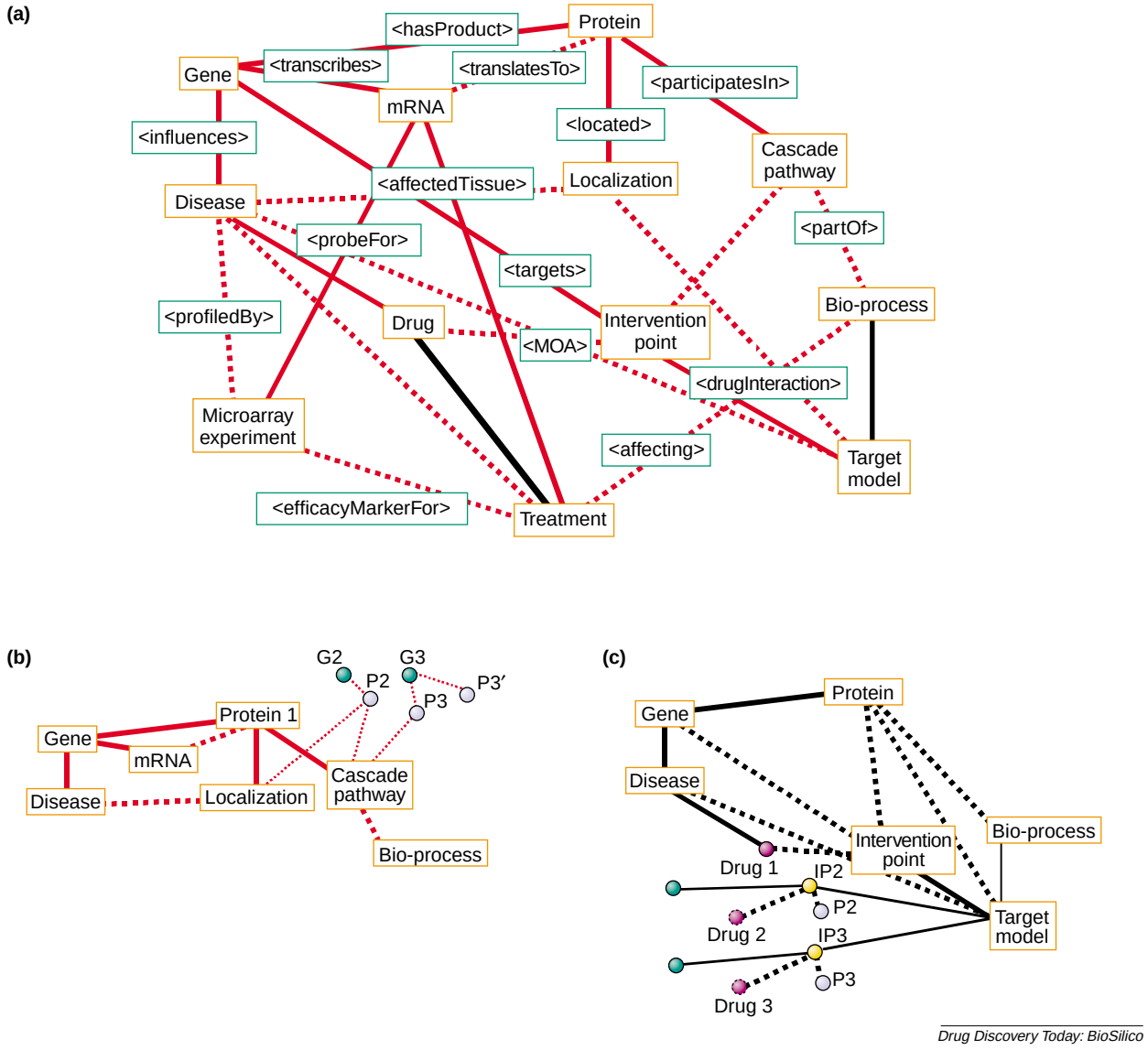


**Figure 1.** Semantic map of molecular data objects. Knowledge assembly of information of multiple but related types can be mapped through the use of an ontology that associates data elements (orange boxes) using semantic relationships (green boxes) into one comprehensive representation via a graph merge **(a)**. Relations can either be direct (solid lines) or aggregated (broken lines). The resulting representations in RDF can focus on different relevance mixtures, such as **(b)** the set of genes involved in a pathway that is associated with a genetic disease, or **(c)** the different druggable targets that can be developed towards a known target model. Note all these instance graphs are subgraphs of the ontology graph (a), which provides a scaffold to organize interpretations and hypotheses, yet from a syntactical point of view (b) and (c) are structurally very different. Abbreviations: G, gene; IP, intervention points; MOA, mechanism of action; P, protein; P3', protein variant; RDF, resource description framework.

(frame based, DL and logic graphs) by web-based technologies, and is an integral part of making the semantic web a reality.

Ontologies not only handle taxonomies (i.e. classification trees), as is evidenced by the Gene Ontology Consortium (http://www.geneontology.org) [15], but can also handle complex sets of relationships (i.e. graph networks) between concepts and facts. Ontologies can be further layered with various concept attributes and restrictions, going well beyond what is possible using a purely object-oriented approach. Indeed, they have already been used to logically organize biochemical pathway information (e.g. BioPAX and BioCYC [16]; http://www.biocyc.org/). As each research community defines their ontologies according to their needs, along with references to already existing ontologies that map onto their domain, they will provide a semantic framework that will allow extensive data and hypothesis exchange between diverse scientific inquiries. Such a community

effort is best implemented through ubiquitous web-based technologies.

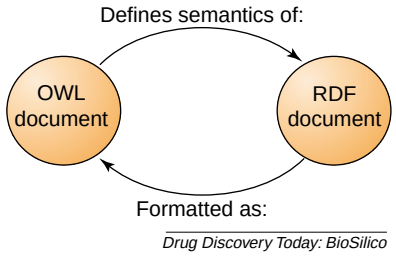


**Figure 2.** The relationship between an OWL document (a special kind of RDF document following the RDF syntax model) and an RDF document, which references one or more ontologies based on OWL. OWL defines the relational logic that RDF documents will refer to and use. Abbreviations: OWL, ontology web language; RDF, resource description framework.

## RDF overview

RDF is an approved W3C specification that relies on semantic relations between elements (meaning) rather than on a syntax grammar (data format). Instead of encoding nested structures that can often be deep or biased (e.g. gene centric versus protein centric), RDF documents are built from subject-verb-object phrases or 'triples'. The verb must be a defined 'property' of a typed object or element, which contains a string, a reference to another element, or a universal resource-identifier (URI). The properties are effectively a semantic mapping of one thing to another (Figure 2). Each fact exists with its own unique meaning, but any part of the fact (subject, verb or object) can be connected to other facts, creating a 'web' of structured, machine-readable facts about a topic domain.

RDF is a format for making statements about facts including hypotheses. The RDF model is based on the use of triples, whose types are explicitly defined either within the RDF body or referenced via links to RDF schemas or ontologies. RDF not only allows making associations about multiple molecular and causal influences, but also statements (propositions) about associations, and statements of statements of, for example:

<Abl-Bcr> <is implicated in> <CML>
<CML> <is a type of> <Lymphoma>
<Gleevec> <inhibits> <Abl-Bcr>
<Gleevec> <cures> <CML>
<Sue> <hypothesizes>
'<Gleevec> <inhibits> <c-KIT>'

Such 'inner statements' are known as reifiable statements because they are not taken as fact until some agreed mechanism marks them as valid, and thereby 'reifies' them (i.e. makes them real). Assuming someone's structured hypothesis is valid is one example of reification that RDF easily supports. It is noteworthy that this has been a major impediment for implementation using standard database approaches, yet it is at the heart of scientific activities.

The use of typed objects and their associated properties requires a mechanism for defining and applying them. Such semantic definitions could reside in other (referenced) RDF documents, comprising mainly class and property definitions rather than instance data. The complete set of RDF documents that refer to these definitions form a closed set under logical closure. OWL is a

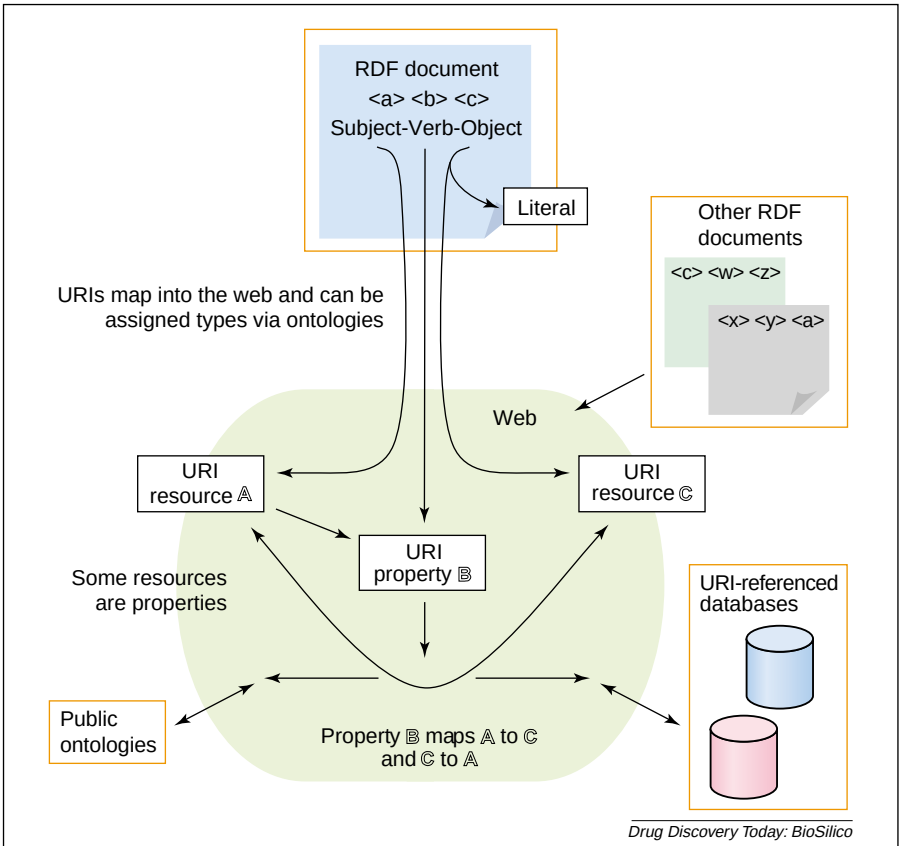


**Figure 3.** A RDF statement can be assigned a graph of A having a property B that maps to C. By using URIs, anything in a document can be semantically mapped (i.e. types with meanings) to anything else on the web, even if it is a specific element within a document. Public ontologies are in OWL format and are web-accessible. Abbreviations: OWL, ontology web language; RDF, resource description framework; URI, universal resource-identifier. Adapted from Ref. [18].

**Box 1. Ontology namespace combination**

Multiple ontologies can be mapped into a document using namespace definitions, as follows:

```
<?xml version='1.0'?>
<rdf:RDF
    xmlns:rdf ='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
    xmlns:rdfs='http://www.w3.org/2000/01/rdf-schema#'
    xmlns:owl ='http://www.w3.org/2002/07/owl#'
    xmlns:umls='http://www.nlm.nih.gov/UMLS#'
    xmlns:bpx='http://www.biopax.org/biopax#'>
```

special case of RDF, and can be referenced by or within RDF documents to form instances of concepts, relations and controlled vocabularies to be used (Figure 3).

The definitions are defined using RDF schemas and/or OWL ontologies depending on the level of detailed specificity required. OWL is based on DARPA's Agent Markup Language (DAML) and OIL [17], and is a RDF vocabulary that provides a richer constraint language (rules about concepts and objects) that facilitates integration and interoperability of data and concepts among more interactive communities. Web ontologies are referenced within a RDF document using namespace tags, which can easily support multiple ontologies by associating each with its own namespace (Box 1).

This allows the combination of different ontologies from various organizations into one composite semantic view that connects definitions within a set of assembled data. RDF can be used to merge data and facts, and make statements about the assembly, including hypotheses, formulas and proofs. RDF was designed to 'say' things (statements) about any group of typed objects or their properties.

RDF can be used as an interchange form that aggregates data, organizes the pieces into a single semantic document and applies assertions onto the document. However, RDF is a lot more than just swapping and aggregating data; it is a fully defined graph model, which enables contents to be stored, typed and managed within any persistent or inferential system in a generalized manner.

Furthermore, RDF is a logical graph model, providing the basis for a set of logical relations and conditions that could be specified by the structure and components of the graph. RDF data can incorporate OWL constructs and be checked for logical soundness, completeness, and analyzed for properties using query and inference tools. In this way, RDF can serve as a common exchange format among many different kinds of knowledge bases.

The model behind RDF is a logical graph, meaning that a set of logical relations and conditions are specified by the structure and components (nodes and edges) of the graph (http://www.w3.org/TR/rdf-mt/) [18,19]. A few assumptions need to be made to represent data and logic appropriately throughout the web, and guarantee its connection to the real world: (i) a RDF graph entails all its subgraphs – a document graph is true if all the component graphs (embedded and referenced) are true; (ii) a RDF graph is entailed by any of its instances – instance to concept mapping; (iii) a set of RDF graphs is equivalent to the merged RDF graph – internet-distributed RDF documents work together as one; (iv) a RDF graph is entailed by another graph if there is a subgraph that is an instance when all variables (e.g. anonymous nodes) are bound validly; and (v) if a subgraph S' entails E, than its parent graph S entails E as well - rule of monotonicity, because a more complete description (S) can never remove validity.

RDF graphs are statements about things one knows or might want to say about the world. Specifically, it can capture a researcher's interpretation of a set of experimental results in some scientific study. This could be an aggregation of commonly believed facts, newly made associations based on a scientist's beliefs (axiomatic and semantic perspectives), or a set of hypotheses the researcher wishes to make about some data. For example:

We propose that:

- the termination and modulation of:
  - the JAK/STAT signaling pathway,
  - is mediated by
    - tyrosine phosphatases,
    - the SOCS (suppressor of cytokine signaling) feedback inhibitors and
    - PIAS (protein inhibitor of activated STAT) proteins

(Pathway example taken from Ref. [20]).

More concrete examples of potential applications for RDF will be described after a brief review of some of the life science issues surrounding data integration.

## A RDF model for the life sciences

A crucial element for future life science informatics initiatives is the availability of an expressive language that can describe biological phenomena and theories in a rich and concise way. A system of life science ontologies can be used to directly tag elements within a RDF document, so that the triples of typed data and/or object instances and their typed relations are consistent with the ontologies referenced (logically validated). This can be demonstrated in the area of molecular biology and bioinformatics. Genes and their potential relations to proteins can be defined as follows (Box 2).

The above ontology can be used to state (in the same or in another RDF document) that a specific gene has a specific protein product. The equivalent inverse relationship

**Box 2. Simple gene–protein relation in RDF and/or OWL**

The following represents an example of a gene being mapped to a protein and its regulator.

```
<owl:Class rdf:ID='Gene'>
        <rdfs:subClassOf rdf:resource='#BioEntity'/>
</owl:Class>
<owl:ObjectProperty rdf:ID='hasProduct'>
        <rdfs:domain rdf:resource='#Gene'/>
        <rdfs:range rdf:resource='#Protein'/>
        <rdfs:inverse rdf:resource='#encodedBy'/>
</owl:ObjectProperty >
<bpx:Gene rdf:ID = 'G27376'>
        <bpx:hasProduct>
                <bpx:Protein rdf:about = '#P74991'/>
        </bpx:hasProduct>
</bpx:Gene>
<bpx:Protein rdf:ID = 'P74991'>
    <bpx:encodedBy>
        <bpx:Gene rdf:about = '#G27376'/>
    </bpx:encodedBy >
</bpx:Protein >
<bpx:TransFactor rdf:ID= 'TF5352'>
…
</bpx:TransFactor>
<bpx:Gene rdf:ID = 'G27376'>
        <bpx:transRegulatedBy rdf:resource= '#TF5352'/>
</bpx:Gene>
```

('encodedBy') can also be stated. The example continues by showing how RDF can be used to state that a specific transcription factor regulates a particular gene. In addition, context-sensitive information about regulation conditions could be added within a context statement.

RDF serves as a well-defined mechanism for using ontologies in conjunction with data and statements about the data, which then can be exchanged in a machine-readable form to other knowledge systems or applications. Using RDF, it is possible to state that 'Eric Neumann does not believe that gene G27376 is regulated by transcription factor TF5352'.

With a web-based ontological and semantic system in place, several key applications for the life sciences are possible. These range from extending the way scientists do research to how all scientific knowledge across multiple domains can be made accessible and reusable throughout intranets and internets.

## Integration of heterogeneous data through common semantics

Gene data from a genomic database and corresponding data from a protein database can be collected and aggregated (merged) within RDF documents. As long as the gene to protein relation has been verified, a semantic relation stating one is the product of the other can be inserted, thereby linking together all gene attributes at most two steps away from the protein attributes. This approach could subsequently take microarray data and associate it with protein information for the same-targeted genes (see Box 2). Functional analysis of genes would therefore be easily associated with localization of the proteins in a context-specific (e.g. tissue) manner.

By using RDF-based services, users and informatics systems would be able to query and request data in a RDF form along with any referenced ontologies. Informaticists could use and store the RDF as it is, or dynamically merge it with additional resources through other RDF web portals. Software tools such as information aggregators (Urchin; http://nurture.nature.com), and web spiders that search the internet and assemble selected data into a single representation, could regularly gather high-quality, relevant knowledge for researchers by using the metadata within RDF and OWL key content sites. Finally, the information gained can then be made available on the web (as RDF) to other scientists who would like to use such captured and interpreted knowledge.

## Applying logic to infer new insights

Another powerful use of RDF (once data are translated into RDF) is the ability to perform inferential logic to discover new insights within the data, in addition to proposing new hypotheses based on prior knowledge and testable models. Once data are represented in RDF, inferences can be made on the databased sets of biologically sound rules (themselves represented in RDF):

'If Gene X is implicated in Disease D, and its Protein Product Y is a functional component of only Pathway P → Then Disease D directly perturbs Pathway P'

which translates into RDF as:

```
<rdf:Description>
<log:is rdf:parseType='Quote'>
      <rdf:Description rdf:about= 'variable#Gene_X'>
        <hasProduct rdf:resource= ' variable#Protein_Y'/>
        <isImplicatedIn rdf:resource= ' variable#Disease_D'/>
</rdf:Description>
      <rdf:Description rdf:about= ' variable#Protein_Y'>
        <inPathway rdf:resource= ' variable#Pathway_P'/>
</rdf:Description>
</log:is>
<log:implies rdf:parseType='Quote'>
      <rdf:Description rdf:about= ' variable#Disease_D'>
        <D_perturbs rdf:resource= ' variable#Pathway_P'/>
</rdf:Description>
</log:implies>
</rdf:Description>
```

Findings can be phrased as hypotheses and used to annotate the original dataset. In medicine, this can be used to infer relations between known disease symptoms and identified molecular targets. Because knowledge can be segmented using RDF (via domains), different belief systems and local models can be concurrently managed to compare and contrast the implications of different hypotheses.

### Creating rich and well-defined models of biological systems

Another RDF application would be the definition and creation of complex biological models using its extensible semantic framework. Model descriptions rely on languages that have two main features: expressiveness and consistency. An expressive language can be used to build complex biological models, both quantitatively and qualitatively.

The definition of models is a key challenge for system biology researchers, and is contingent on having a proper set of expressive biological relations (protein activation, gene expression, partial agonists, genetic predisposition, downstream effects, limited compartments). Organizations such as BioPAX will be able to help guide the inclusion by scientists of all relevant relations as part of an OWL-based ontology. These can then be used to define the molecular and cellular processes, and the causal (transforms-to, activates) and associative (part-of, attached-to) relations within RDF-based documents and knowledge models [16].

### Formal (semantic) annotations that can be shared

Rather than relying on using simply free-text in comment fields, a detailed model and instance definition can be the primary form of an annotation system. This could be a potential way to augment the distributed annotation service (DAS) system developed by genomic researchers [21], thereby making large sets of genomic facts machine-readable. This model can be expanded to more nascent areas of research, such as systems biology, that rely more heavily on complete biological and molecular knowledge, and the use of logical relations [22].

Annotations that include logical relations will be accessible to a new generation of search engines that can locate 'types' of phenomena or mechanisms. Formal annotations will be able to include proposed hypotheses that would be in a form for others to use in inference engines. Indeed, the boundary between classic search engines and inference engines would begin to blur. Annotations could also readily refer to earlier logical annotations, either as an extension to them or as a refutation to an earlier hypothesis.

The formal representation does not make the interpretation of such annotations any more difficult; in fact, it will be easier to understand because the actual semantics are already in place, thereby reducing misinterpretation. Browser-based tools could easily render these annotations in clear textual and graphic representations, with pathways as graphs and genomes as linear constructs, by using current web-based scalable vector graphics (SVG). The same display would be used by scientists at 'annotate an annotation' and publish that composite view through a DAS-like system.

### Embedding models and semantics within online publications

As a logical next step for HTML publications, RDF can be embedded in scientific literature and used to integrate biological models within a publication, link them as an extension to prior versions of the models (active references), and capture the primary conclusions of a paper in a form that can be queried by a machine, which also relates the biological entities to typed data web-objects (e.g. public genomic databases). Similar to the goal of the semantic web, scientific and medical articles will be intelligently and semantically linked to each other, forming a web of scientific research [23], whose accessibility will revolutionize how science is carried out, debated and rationalized.

Currently, text mining has become a very active area for scientists to find and aggregate the knowledge contained within scientific articles. It will soon be possible for semantic embedding to be part of the publishing process through the creation of new, user-friendly authoring tools. Text mining will only be necessary for extracting from the finite set of legacy literature, and then it will become a thing of the past. Such an undertaking will require sets of ontologies that will cover the different scientific areas.

The National Institutes of Health (NIH; http://www.nih.gov) has hundreds of datasets as part of their sophisticated public health knowledge network, including links to published scientific papers through PubMed. Yet with all this open and multifarious connectivity, it is often difficult to find the right pieces of information for a given topic. Returning the right item is about having enough knowledge of a topic to understand what is really requested and how it should be retrieved. Short of just having humans in the loop, what is needed now is a system that understands and links meaning (i.e. for researchers), not just matching strings of letters during a query. The inclusion of semantics with these databases would greatly enhance the search capability of these resources, and allow them to be tied to the other databases with more meaning than just a URL link. It would also enable agent-based harvesting to be more productive and comprehensive.

Several online journals already are including resources to deeper representations of pathways and mechanisms (http://www.signaling-gateway.org/, http://stke.sciencemag.

org/). A RDF-based approach could be embedded directly into the publishing mechanisms, and the online papers or summary articles viewed through the web-based tools described above. The documents would include formal models within them, in addition to links to preceding models (active references) that could be activated by any modelling or simulation tool capable of parsing RDF statements. Not only would researchers view biological pathways from within scientific papers, but they could also query, analyze and evaluate them directly through the web-based tools and interfaces.

## Key steps that need to happen

The applications outlined above are – at a technical level – achievable today, but they require an ontological and web-centric perspective to be in place for the life sciences. Forethought and consistency in ontological definitions are necessary in the construction of useful ontologies. However, this process must be driven by the way scientists view their scientific world and, because this is certainly an evolutionary process (or possibly even revolutionary), the ontologies do not need to be perfect from a conceptual point of view; they should first and foremost be practical and be able to address contemporary use such as cases around databases and communication. An open forum for discussing practical approaches is therefore necessary, so that a genuine community of practice can emerge similar to what is already applied in organizational development [24]. To begin addressing this need, a Semantic Web for Life Sciences discussion group has been created at W3C (http://www.public-semweb-lifesci@w3.org), which will help coordinate activities between the W3C's Semantic Web Initiative and life science interests. The next step is to formalize such a group to actively develop RDF use-cases and basic implementations to some key life science informatics issues that will be placed in the public domain. Activities in this effort are already underway at the public-semweb-lifesci mailing list.

Several groups have already formed around developing ontologies using RDF and OWL, including GeneOntology [17], BioPAX and GenomeKnowledge (http://www.genomeknowledge.org/) [25]. Most have defined their ontologies using OWL, yet have not explored the uses of RDF to realize their goals. However, once rendered in OWL, it is easy to implement a RDF model to exchange and merge data that is covered by the ontology's semantics. Finally, there are many RDF-related resources freely available to the public, including information [3] (http://www.w3.org/TR/rdf-concepts/) and tools (e.g. JENA; http://www.hpl.hp.com/semweb/), which can be used for fast implementation.

## Conclusion

The web has had an important effect on how science is now practised: (i) research documents are rapidly distributed throughout a community for review and comment; (ii) experimental data are easily shared with others, thus accelerating its analysis and interpretation; (iii) internal databases containing the distillations of scientific research are publicly accessible and easily queried through user-friendly web pages; (iv) scientific groups can share computational resources with each other through grid computing practices [26]; and (v) peer-reviewed journals offer online access allowing scientists to harvest large sets of relevant articles.

Scientific publications and curated databases together hold a vast amount of actionable knowledge. However, their full value is realized only in the context of such resources being connected together by meaning, such that machine processes can traverse and identify these links intelligently. As has been outlined here, RDF is sufficiently expressive and can be linked to other web resources in order to be the foundation for knowledge sharing and complex scientific transactions. Its applications will be driven by how well the life sciences community can realize its benefits and potential in light of its current challenges.

The enormous efforts that the life science community has to-date undertaken regarding data management and standards could be efficiently re-focused to rapidly take advantage of the semantic web. Many of the elusive informatics challenges facing scientists and informaticists in life sciences might well be solved by this unique paradigm. While RDF will not directly improve success in high-throughput chemical screening, animal studies or clinical trials, it could sufficiently consolidate important knowledge across multiple disciplines employed by the pharmaceutical industry to better enable decision-making at all levels of the drug discovery pipeline [27].

Modern science is built on top of the shoulders of earlier published research. Specifically, the validity and importance of scientific work depends on the clear connection to related past and future research (either to corroborate or refute) because this is central to the scientific method itself. The linking of systematic meaning with interpretation between related research is precisely what the semantic web can uniquely offer on a global- and domain-focussed scale. If scientists can more effectively discover research findings that offering new insights into genes or diseases, or identify collections of researchers investigating common topics, then science will be able to take advantage of the swelling tide of research-generated data, rather then succumbing to it.

## References

1 Waterston, R.H. *et al.* (2002) On the sequencing of the human genome. *Proc. Natl. Acad. Sci. U. S. A.* 99, 3712–3716
2 Alter, O. *et al.* (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U. S. A.* 97, 10101–10106
3 Powers, S. (2002) *Practical RDF: Solving Programs with the Resource Description Framework*, O'Reilly Press
4 Berners-Lee, T. *et al.* (2001) The Semantic Web. *Sci. Am.* 284, 29–37
5 Cambridge Healthtech Institute (2003) *Post-Genomic Target Validation,* Cambridge Healthtech Institute
6 Moore, S.K. (2001) *Harmonizing Data, Setting Standards*, Institute of Electrical and Electronic Engineers Spectrum, Jan 2001
7 Neumann, E. and Thomas, J. (2002) Knowledge assembly for the life sciences. *Drug Discov. Today* 7, S160–S163
8 McEntire, R. *et al.* (2000) An evaluation of ontology exchange languages for bioinformatics. *Proc. Intl. Conf. Intell. Syst. Mol. Biol.* 8, 239–250
9 Open Bioinformatics Foundation, 17 Clinical Data Interchange Standards Consortium
10 Common Warehouse Metamodel Specification v1.1 OMG TC 2003-03-02.
11 Farquhar, A. *et al* (1997) The Ontolingua Server: A tool for collaborative ontology construction. *Intl. J. Human-Computer Studies*, 46, 707–727
12 MacGregor, R. (1991) Inside the LOOM description classifer. *SIGART Bull.* 2, 88–92
13 Goble, C.A. *et al.* (2001) Transparent access to multiple bioinformatics information sources. *IBM Systems J.* 40, 532–552
14 Chaudhri, V.K. *et al.* (1998) *Open Knowledge Base Connectivity 2.0*, Knowledge Systems Laboratory
15 Harris, M.A. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32, D258–D261
16 Ideker, T. *et al.* (2004) *Introduction to Systems Biology*, Proceeding of the Pacific Symposium on Biocomputing, 9, 471–473
17 Fensel, D. (2000) OIL in a nutshell. In *Knowledge Acquisition, Modeling, and Management, Proceedings of the European Knowledge Acquisition Conference (EKAW-2000) (Lecture Notes in Artificial Intelligence)* (Dieng, R. *et al.*, eds), Springer-Verlag
18 Hayes, P., ed. (2004) *RDF Semantics*, W3C TR Doc /2004/ REC-rdf-mt-200402102004.
19 Bollobas, B. (1998). *Modern Graph Theory*, Springer-Verlag
20 Heinrich, P.C. *et al.* (2003) Principles of interleukin (IL)-6-type cytokine signalling and its regulation. *Biochem. J.* 374, 1–20
21 Dowell, R.D. *et al.* (2001) The distributed annotation system. *BMC Bioinformatics* 2, 7
22 Karp, P.D. (2001) Pathway databases: a case study in computational symbolic theories. *Science* 293, 2040–2044
23 Hendler, J. (2003) Science and the Semantic Web. *Science* 299, 24
24 Wenger, E. (1998), *Communities of Practice*, Cambridge University Press
25 Joshi-Tope, G. *et al.* (2003) *The Genome Knowledgebase: A Resource for Biologists and Bioinformaticists (Cold Spring Harbor Symposia on Quantitative Biology)* (Vol. 68), Cold Spring Harbor Laboratory Press
26 Stevens, R.D. (2003) myGrid: personalised bioinformatics on the information grid. *Bioinformatics* 19 (Suppl. 1), 302–304
27 FDA (2004) *Innovation or Stagnation*, FDA Report