The Life Sciences Semantic Web is Full of Creeps!

Benjamin M. Good and Mark D. Wilkinson

Abstract

The Semantic Web for the Life Sciences (SWLS), when realized, will dramatically improve our ability to conduct bioinformatics analyses using the vast and growing stores of web-accessible resources. This ability will be achieved through the widespread acceptance and application of standards for naming, representing, describing and accessing biological information. The W3C-led Semantic Web initiative has established most, if not all, of the standards and technologies needed to achieve a unified, global SWLS. Unfortunately, the bioinformatics community has, thus far, appeared reluctant to fully adopt them. Rather, we are seeing what could be described as 'semantic creep'—timid, piecemeal and ad hoc adoption of parts of standards by groups that should be stridently taking a leadership role for the community. We suggest that, at this point, the primary hindrances to the creation of the SWLS may be social rather than technological in nature, and that, like the original Web, the establishment of the SWLS will depend primarily on the will and participation of its consumers.

Keywords: Semantic Web; Ontologies; RDF; LSID; OWL; Bioinformatics; Creeps

INTRODUCTION

The World Wide Web (WWW) can be envisaged as a graph in which the nodes are documents and the unlabeled edges are the hyperlinks. The Semantic Web (SW) can be thought of as a graph in which the nodes are anything that can be named (a concept, a document, a person) and the labeled edges are meaningful properties that describe the relationships between the nodes. While the WWW is primarily a web of documents, the SW is meant to be a web of data and knowledge.

By providing improved mechanisms for organizing information on a global scale, the SW should benefit all producers and consumers of information. Based on the rapidly expanding role of web-based resources within bioinformatics, the SW should offer particularly substantial benefits to Life Science researchers. Life scientists can already utilize web-accessible databases and analytical tools to conduct hypothesis-driven experiments purely in silico [1] and such resources also provide mechanisms for improving the analysis and interpretation of data produced by inhouse experiments. However, despite their demonstrable utility, the nature of WWW resources is not ideal for conducting integrative biological analyses. The practice of in silico, web-based biological research is hindered by the lack of:

- Globally unique and resolvable names for biological entities
- Consistent standards for data representation
- Consistent standards for knowledge representation
- Standard interface definitions for data retrieval and processing

The SW offers solutions that address each of these issues:

- The Life Science Identifier system (LSID)
- The Resource Description Framework (RDF)
- The Web Ontology Language (OWL)
- Semantic Web Services
While, given their direct applicability to the most critical problems, we might wish for these solutions to be rapidly adopted by the Life Sciences community; the technologies are, instead, slowly creeping into use by a minority of the key players (‘Creeps’) in the field.

Here, we describe the motivation for and the application of each of these technologies, describe the Semantic Creep exhibited by projects within the life sciences community, and discuss the technical and social barriers that are impeding their full penetration.

IDENTIFYING BIOLOGICAL ENTITIES ON THE SEMANTIC WEB—THE LIFE SCIENCE IDENTIFIER SYSTEM

Identifying and naming biological entities is a source of significant consternation in the research community. Disparity of gene names, for example, has resulted in heated public debates [2–8], led to the creation of permanent gene naming organizations [9], and instigated a cottage industry focused on automatic gene name recognition in text, complete with its own annual contest [10]. While biological entities such as gene names have contextual, historical, functional and sometimes even emotional meaning within the biological community, Web identifiers are ‘opaque’, and do not carry the same vested interests precisely because they are not intended to be used as names in human discourse. In the context of the Semantic Web for the Life Sciences SWLS, an identifier is simply a ‘pointer’, and does not carry any innate meaning. As such, SWLS researchers can and should avoid these nomenclature arguments and pursue workable solutions to more critical Life Science research problems.

The first requirement for the formation of the web was the mechanism for global document identification and retrieval—the Uniform Resource Locator (URL). Once this standard was widely accepted and applied, the rest of the web linked rapidly into place. In exactly the same way, the SW is, first and foremost, a web-based technology; as such, the first requirement for the SWLS is to define a mechanism for stable and predictable SW entity identification and retrieval that spans both biological entities (e.g. proteins) as well as conceptual entities (e.g. ontology terms).

URLs offer one very significant advantage over other naming strategies that might be employed. They are guaranteed to be unique by the infrastructure of the WWW. Thus the development of integrative technologies can be undertaken that dependably rely on their uniqueness. It is critical that the SWLS community focus on this most fundamental step—utilization of web-enabled identifiers for all entities and concepts—as its first priority. Though they are being utilized by a small number of Creeps, as illustrated by the National Center for Biotechnology Information’s (NCBI) current use of stable URLs for identifying much of its data [11], the need for ubiquitous application of such identifiers should not be underestimated by the wider community, as it is the fundamental requirement for the creation of the SWLS.

Unfortunately, URLs offer only a limited solution to the naming problem in that they are intended to be used to identify documents rather than discrete conceptual units, and offer no presumptions of stability over time [12]. As Stein [13] points out, this instability causes problems even on the existing Web, and is equally or more devastating in the context of the SW. As a result, alternative identification systems have been developed in recent years that provide significantly more power and stability than URLs alone, and are more suitable for use on the SW.

Rooted in the Interoperable Informatics Infrastructure Consortium (I3C) activities between 2002 and 2003, the LSID became an OMG [14] standard in 2004. It was designed to be a location-independent, stable, and resolvable identifier for entities within the SWLS [15]. The LSID specification includes three aspects:

(i) Standardized identifier format, including versioning;
(ii) Defined protocols for retrieval of identified data and/or metadata;
(iii) Customary metadata syntax.

LSIDs take the form of Uniform Resource Names (URNs) [12] which are guaranteed to be globally distinct for any named entity; i.e. two different entities are guaranteed not to have the same LSID. Importantly, unlike URLs where a change in the location of data often changes its identifier, the LSID of an entity is independent of its location. As such, LSIDs are extremely stable, giving third-parties the ability to reliably generate annotations for any entity...
named by an LSID, even though they cannot control the ultimate location of the identified data.

A second, and in the context of the SW, more important advance of LSIDs over URLs is that they can be used to identify entities other than documents. Conceptual entities can be identified by LSIDs because there is no requirement for an LSID to resolve to data. An LSID might only identify metadata, and this makes it plausible to use LSIDs to identify ontology nodes, in place of the URL-style ‘document#fragment’ identifiers that are more commonly used. In fact, at least one key developing SWLS resource—UniProt-RDF—already does this [16]. This latter behavior—‘creep by proxy’—is a by-product of SWLS advocates imposing SW technologies onto the resources that are most critical to their users in the absence of uptake of these technologies by the resource curators themselves. Though a workable temporary solution, it is clearly fragile and unsustainable.

There are a variety of technical reasons that might account for the slow and/or incomplete adoption of LSIDs:

(i) None of the most common web clients can natively resolve an LSID to data or metadata.
(ii) LSIDs are perceived to be high-maintenance due to the specification’s requirement to update version numbers for even minor changes in the underlying data, potentially resulting in a high curatorial burden.
(iii) Providers of LSIDs are expected to implement LSID resolver services to take full advantage of the resolution system; this requires dedicated coding efforts.
(iv) The code providing the LSID resolution stack, and its code dependencies, have been somewhat unstable to date, making it unreliable for production environments.

<table>
<thead>
<tr>
<th>Project</th>
<th>Domain</th>
<th>LSID Visibility</th>
<th>Resolver?</th>
<th>Authority</th>
</tr>
</thead>
<tbody>
<tr>
<td>FuGE [74]</td>
<td>Data standards for functional genomics</td>
<td>Internal</td>
<td>Internal</td>
<td>Self</td>
</tr>
<tr>
<td>uBio [75]</td>
<td>Universal organism naming service/authority</td>
<td>Public</td>
<td>Yes</td>
<td>Self</td>
</tr>
<tr>
<td>BioMoby [56]</td>
<td>Interoperability via web services</td>
<td>Public</td>
<td>Yes</td>
<td>Self</td>
</tr>
<tr>
<td>myGrid/Taverna [76]</td>
<td>Interoperability via web services</td>
<td>Internal</td>
<td>Internal</td>
<td>Self</td>
</tr>
<tr>
<td>uniProt-RDF [16]</td>
<td>Protein data repository</td>
<td>Public</td>
<td>No</td>
<td>Self</td>
</tr>
<tr>
<td>Pseudogene [77]</td>
<td>Pseudogene knowledgebase</td>
<td>Public</td>
<td>Unknown</td>
<td>(no example found)</td>
</tr>
<tr>
<td>GBIF [78]</td>
<td>Biodiversity</td>
<td>Unknown</td>
<td>Unknown</td>
<td>(non-responsive)</td>
</tr>
<tr>
<td>HapMap [79]</td>
<td>Haplotype Maps</td>
<td>Public</td>
<td>Unknown</td>
<td>Self</td>
</tr>
<tr>
<td>SLRP [80]</td>
<td>Entities retrieved during in silico experiments</td>
<td>Unknown</td>
<td>Yes</td>
<td>Self</td>
</tr>
</tbody>
</table>

Examples of projects that employ LSIDs either within their internal architectures or as public naming authorities. Noted are the domain of entities that are named by the LSID; the visibility of the LSID—that is, whether or not the LSIDs are meaningfully shared publicly; the existence of an LSID resolver for those LSIDs; and whether or not the authority for the data/metadata being identified by the LSID is owned by the project itself, or if the LSID resolver is acting as a proxy wrapper around third party resources.
The metadata format used in LSID metadata resolution (discussed below) is well-supported in Java, but much less so in Perl or many other languages commonly used by the bioinformatics community, thus alienating many resource providers.

While some of these problems, such as code instability, will no doubt diminish over time, others, such as native support for LSIDs in common web tools, will only be resolved when the utility of these identifiers becomes more broadly recognized, and their usage more widespread.

There may also be sociological factors limiting the success of LSIDs. There was certainly some initial mistrust of the I3C by the largely open-source-oriented bioinformatics community and, justified or not, this seems to have affected up-take of its spin-off technologies even now that they are ‘unencumbered’. In addition, many of the major players, for example much of the ontology development community, were (and are) focused on centralization of their resource-building efforts. As such, the benefits offered by a location-independent naming system are not compelling. We believe, however, that such centralized efforts are unsustainable, and that the SWLS will eventually be created in much the same way as the web itself was—by its users. As such, the power of LSIDs (or some other URI-based identifier system) will soon become indispensable. Recent discussions within the newly formed Semantic Web Health Care and Life Sciences Interest Group [18] support this view.

Data representation—RDF

Once data can be uniquely identified on the web, it should be in a form interpretable by the software that retrieves it. On the WWW, that form is predominantly HTML; on the SW, that form is envisaged to be RDF. In February 2004, the WWW Consortium announced their recommendation of RDF as a ‘lightweight ontology system for knowledge exchange on the web’ [19]. RDF represents data as a set of directed graphs; URIs are assembled into triples composed of a subject URI, a predicate URI, and an object URI. The predicates of RDF triples on the SW are similar to the hyperlinks of the web; however the key advantage of RDF triples over HTML hyperlinks is that the links are explicitly labeled—the intent (semantics) of the relationship between the two entities is thus computationally accessible through URI resolution. These Triples of URIs are the most fundamental concept in the SW.

The properties of URIs allow all participants on the SWLS to refer to and annotate any entity in the sphere of public information. Moreover, it is possible to assume named-entity-equivalency based entirely on syntactic matching of the URI strings, thus two groups utilizing RDF containing the same URIs can merge their information safely and automatically. This simple, yet incredibly powerful feature dramatically reduces the effort required to integrate distributed sources of information; however it is critically dependent on the community’s adoption of, and adherence to, consistent standards for naming.

LSID metadata is, by convention, provided in RDF format. In this case, the RDF document explicitly describes the relationship between the entity named by the LSID, and other entities on the SWLS (named by LSIDs or URLs), and this is made available in a standardized way by the LSID metadata resolution protocol. This makes LSIDs distinctly well-suited for use on the SWLS, despite the current lack of native support in most web browsers and agents.

The lack of client-side support for SW technologies is not limited to LSIDs. Just as the potential of the HTML web was not apparent to the world prior to the advent of the web browser, the potential of the SW will not be readily apparent without the development of interfaces that facilitate user interaction with RDF. At present, mechanisms for ‘surfing’ through RDF graphs are just starting to be realized in a variety of emerging Semantic Web Browsers [20–27]. Many of these systems are tested using biological data (notably Haystack’s use within the myGrid project [28]), however only BioDash has specifically targeted the life sciences community [29]. Though the SW’s key enhancement to the current web is its improved computational accessibility, the success of semantically enabled browsers in connecting human users with information integrated through RDF will likely prove pivotal to the success of the SWLS in exactly the same way that early browsers such as Mosaic were vital to the adoption and rapid growth of the original web.

Knowledge representation

Despite the utility of URIs and explicit relationships between them, a URI alone cannot indicate what
kind of entity it identifies. Thus, on top of RDF, it is possible and sometimes critical to impose a layer of computationally tractable meaning, such that the identified entities are explicitly defined, and the relationships that hold between them can be accurately interpreted and utilized in an automated fashion. On the SW, OWL ontologies [30] provide these definitions. For a general review of ontology-based knowledge representation in biology, see [31], and for a recent introductory review of SW ontology languages and some of their applications in biology, see [32].

The prototypical example illustrating the need for the shared semantics provided by ontologies is the task of issuing and processing a single query that spans multiple databases [33]. Somehow, the relationships between entities in different databases (e.g. equivalency) must be identified automatically in order for the distributed queries to succeed. The paradigmatic solution to this problem, implemented many times in the bioinformatics domain, is the creation of an ontology into which entities from each participating database can be mapped [34–38]. Such ontologies embody the knowledge needed to perform the necessary alignment of entities in different databases to allow for successful distributed query and/or data integration. Table 2 enumerates a list of some of the key ontologies in bioinformatics, indicating their domain, the type of representation they employ, and whether or not they implement resolvable URIs.

### Table 2: Key biomedical ontologies.

<table>
<thead>
<tr>
<th>Name</th>
<th>Domain</th>
<th>Representation</th>
<th>Resolvable URIs</th>
<th>Accessibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foundational Model of Anatomy [82]</td>
<td>Human Anatomy</td>
<td>Protégé Frames</td>
<td>No</td>
<td>Academic License</td>
</tr>
<tr>
<td>OpenGALEN [83]</td>
<td>Medicine</td>
<td>GRAIL [84]</td>
<td>No</td>
<td>Open</td>
</tr>
<tr>
<td>National Cancer Institute Thesaurus [85]</td>
<td>Cancer biology</td>
<td>OWL</td>
<td>Yes</td>
<td>Open</td>
</tr>
<tr>
<td>Gene Ontology [86]</td>
<td>Gene Products</td>
<td>OBO, OWL, RDF, MySQL</td>
<td>No</td>
<td>Open</td>
</tr>
<tr>
<td>BioPax [87, 88]</td>
<td>Pathways, interactions,</td>
<td>OWL</td>
<td>Yes</td>
<td>Open</td>
</tr>
<tr>
<td></td>
<td>protein modifications</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MGED [89]</td>
<td>Microarray experiments</td>
<td>OWL</td>
<td>Yes</td>
<td>Open</td>
</tr>
<tr>
<td>UMLS Semantic Network [90]</td>
<td>Medicine</td>
<td>Internal</td>
<td>No</td>
<td>Open</td>
</tr>
<tr>
<td>biOzen OWL ontology framework [91]</td>
<td>Bioinformatics</td>
<td>OWL</td>
<td>Yes</td>
<td>Open</td>
</tr>
<tr>
<td>myGrid [57]</td>
<td>Bioinformatics Web Services</td>
<td>OWL, RDF</td>
<td>Yes</td>
<td>Open</td>
</tr>
<tr>
<td>BioMoby Object Ontology [92]</td>
<td>Bioinformatics Data Types</td>
<td>RDF</td>
<td>Yes</td>
<td>Open</td>
</tr>
<tr>
<td>BioMoby Namespace Ontology [93]</td>
<td>Bioinformatics Data Resources</td>
<td>OWL</td>
<td>Yes</td>
<td>Open</td>
</tr>
</tbody>
</table>

Listed are a selection of ontologies for bioinformatics and medicine, their domain-scope, their representation, and whether or not their components are identified by resolvable URIs. The citation for each provides a link to either the project website or directly to the ontology. Numerous additional key ontologies not listed can be found on the OBO website [81].

As Table 2 demonstrates, many expertly crafted and professionally curated biomedical ontologies exist; however the utilization of these ontologies within the SWLS depends on independent parties having the ability to link their local data to ontologically defined concepts through RDF Triples. Unfortunately, many key biomedical ontologies are hidden from the SW because they do not utilize resolvable URIs to name their components. Like un-hosted html pages, they are invisible and unusable without context-specific software because the ontology developers have focused on ‘Semantic’ rather than on ‘web’.

Efforts have been made to migrate some of the largest and most important extant, non-SW ontologies, such as the Foundational Model of Anatomy (FMA) and the Unified Medical Language System (UMLS), to the SW paradigm; however, this is a slow and technically challenging process [39, 40]. In addition to the difficulties of converting between knowledge representation standards (e.g. from Protégé Frames to OWL Description Logic), the often tremendous size of these ontologies makes them unwieldy to process with existing SW reasoning systems. The NCI Thesaurus, for example, is composed of over 500 000 triples describing more than 27 000 classes [41], and the OWL representation of the Gene Ontology is over 10 Megabytes. Thus, despite the presence of the ontology on the web and its representation in a SW language, the size...
of these resources often makes them unfeasible to retrieve and process at run-time as is envisaged for their utilization on the SWLS. To address this problem, a number of research groups are investigating mechanisms for decomposing such large ontologies into smaller, more portable pieces that still retain the semantics of some of the complete versions [42–45]. As indicated above, we believe that LSID metadata resolution may hold the key to utilizing these ontology fragments in an open and scalable manner.

Besides the technical challenges of translating, transporting and reasoning over the enormous ontologies characteristic of biomedical informatics, licensing issues (for example with many of the source vocabularies of the UMLS), may also hinder the ability of semantic content providers to release their ontologies according to the open-access paradigm of the SW. Though the technical challenges seem surmountable, this social obstacle, characteristic of a centralized view of knowledge management that is in opposition to the nature of the web, may prove the most difficult to overcome. Nevertheless, the rapid adoption and clear benefits provided by the open source software movements and open access scientific journals provide hope that the providers of some of the most computationally useful knowledge in biology will eventually follow suit. The open access criteria for inclusion under the umbrella of the Open Biomedical Ontologies (OBO) project of the National Center for Biomedical Ontology (cBIO) and the same initiative’s creep towards providing OWL versions of their ontologies indicates a clear push by these community leaders towards the realization of the SWLS.

**Interface definitions—accessing the Hidden Semantic Web**

Using ontologically described RDF and a consistent global naming strategy such as URLs or LSIDs, it is theoretically possible to automate web-based data integration to a level never before realized. However, much of the richest content on the web remains ‘hidden’ in databases and analytical methods that are only accessible via the activation of a server-side program. Thus, to fully realize the SWLS, consistent mechanisms for accessing this hidden web must be implemented.

With respect to automation, the limitations of standards designed to facilitate human interaction with dynamic Web content such as the Common Gateway Interface (CGI) have been well-described [13]. The key limitation is that the nature of the data to be entered into any given Form field is indicated by visual cues on a web page, and as such the primary target for both the input and the output of these services is intended to be a human operator, rather than a user-agent. Importantly in the context of the SW, the computational opacity of web forms creates a barrier to the indexing of that data by search engines. Though the inability of search engines to index data resources is likely considered a ‘feature’ rather than a ‘bug’ by many data providers, it defies the SW’s ability to automate discovery and retrieval.

**The rise of Web Services and Semantic Web Services**

Web Services are a formal specification that proposes to more tightly define the input and output of web interfaces through a machine-readable specification called Web Services Description Language (WSDL), and moreover, defines a registry-based discovery system—Universal Description, Discovery and Integration (UDDI)—for brokering interactions between service consumers and appropriate service providers. Through UDDI/WSDL, a client program can look-up services that execute a particular type of operation, and precisely determine the input and output fields for that service. Moreover, the communication between client and service utilizes a novel XML-based standard, the Simple Object Access Protocol (SOAP).

There are several clear advantages to the Web Service approach. First, the use of XML-based messaging makes the Web Services infrastructure platform- and language-independent. Unlike direct-access systems, it enables the provider to extensively change their architecture, data model, or data location without affecting their public interface. Consumers do not need to know anything about the implementation of the service or any details of its operations in order to invoke the resource. Further, changes to the interface can immediately be detected by client software.

Unfortunately, Web Services do not go far enough to solve the most critical interoperability problems in the Life Sciences domain. The promised interoperability between platforms and languages is certainly achieved, but other promises of Web Services are only successful within well-defined, closed-world situations. There are two primary problems with archetypal Web Services, both of
which share a common root cause in the lack of semantics:

(i) The discovery of appropriate services cannot be done entirely by machine, because machines cannot determine the ‘intent’ of a service or its interface components.

(ii) Though a client can detect when an interface has changed, it cannot (except in some specific pre-anticipated circumstances) react to this [46]; thus the invocation of that service is blocked exactly as it would be with a CGI form.

In recent years, Semantic Creep has invaded these interface definitions, in the form of Semantic Web Services\(^1\). A number of high-profile Life Science Semantic Web Service initiatives have emerged from these activities: TAMBIS [49], BioMoby [50], caCORE [51], PathPort [52], myGrid [53] and DiscoveryNet [54]. Each of these projects has enjoyed some degree of success, and interestingly, the location of the Semantic Creep has been different in each case. We will focus on the three most broadly-scoped of these projects—BioMoby, myGrid, and caBIO—and examine the semantics that allow them to function more effectively than Web Services alone.

BioMoby: Semantic Creep appears in the data-type definitions and the service annotations of BioMoby Web Services [50, 55]. BioMoby uses an ontology to define data types that can be ‘understood’ and passed between the Web Services of data providers and data consumers. In addition to data typing, a simple hierarchical vocabulary of bioinformatics service types is available such that the nature of the Web Service operation can be described. The BioMoby Web Service registry utilizes these ontologies to enrich its ability to match-make between clients requesting data services, and service providers that are capable of responding. Currently there are more than 300 data types and 630 data retrieval and analysis tools available via the BioMoby Web Service standard [56].

myGrid: The Semantic Creep of myGrid is distinct in that it involves ontology-based annotation of interface descriptions [53], rather than defining data structures, themselves, as BioMoby does. myGrid services utilize archetypal WSDL and W3C XML Schema to describe service interfaces and data structures, respectively, and extends these standards by employing custom designed ontologies [57] to annotate the ‘intent’ of each field in the interface, the nature of the underlying resource, and/or the algorithm being used to execute the service. This allows these paradigmatic interfaces to be discovered in semantically rich ways by their registry query system compared with traditional UDDI queries.

caBIO: Semantic Creep appears within Cancer Bioinformatics Infrastructure Objects (caBIO) as a thesaurus used to annotate the data accessible through the caCORE system [51]. caBIO is an object-oriented model for representing and computing over objects in this biomedical domain. Accessible via Java, SOAP, and HTTP–XML web interfaces, it forms the core computational infrastructure of caCORE, the foundation of the National Cancer Institute’s (NCI) bioinformatics platform.

Comparing these approaches, we see that BioMoby is distinct in that the creep of semantics into the data-type definitions allows the process of service pipelining to be entirely automated; knowledge exploration tools such as the Ahab BioMoby client are capable of discovering and executing services in an unattended manner [58], and the manual creation of workflows can be guided by the data semantics, as seen in the BioMoby plug-ins to myGrid’s Taverna [59] workflow management tool, and in other client applications such as Remora [60] and MOWserv [61]. The cost of this high degree of interoperability is that every participating service must be specifically coded to utilize BioMoby data types and messaging formats. Thus the key advantage of the myGrid approach is that legacy Web Services can be annotated and entered into the myGrid registry without modification of the service code. Moreover, myGrid Web Service annotation can be done by third parties with no participation by the service provider themselves. The limitation of the current myGrid implementation is that the myGrid ontology is ‘closed’—service providers cannot extend it as they do in the BioMoby system. Thus, if a required ontological term is missing, the service cannot be fully annotated.

In sharp contrast to the BioMoby and myGrid projects, which focus on the development of standards and registries designed to enable interoperability of publicly available web services, caCORE is the primary provider of the services that are accessed via caBIO interfaces. As such, the caBIO
service ‘registry’ operates within a closed world. The principal benefits of this are that the pre-existing knowledge of the system components allows for the precise definition of an extremely rich Application Programming Interface (API) the complete set of valid Web Service interfaces, and what can be done with any given data type available within caBIO, is detailed in the project documentation. This enables the construction of extremely powerful query interfaces. For example, it is possible for the system to answer questions like

‘In brain tissue from patients diagnosed with glioblastoma multiforme subtype of astrocytoma, which genes in the p53 signaling pathway are over or under expressed in cancerous versus normal tissue?’ [51]

The ability to natively answer such powerful queries distinguishes the caBIO approach from any of the other bioinformatics Semantic Web Service projects, but it comes at the expense of requiring additional curation to utilize external resources, and rigidly defining what operations are possible on any given piece of data. It is possible that, when Semantic Creep has sufficiently penetrated the data description layer of open Web Service solutions, as it has with the NCI data resources, it will be possible to develop queries over open systems that produce similar results. As more data is published as RDF and annotated with the growing number of publicly available OWL ontologies, we believe it likely that the power of queries over open Semantic Web Service frameworks will improve to a level comparable to that seen in caBIO.

Unfortunately, even when Semantic Creep has fully penetrated every aspect of Semantic Web Services, a technical gap between the Semantic Web of documents and the Semantic Web of services will remain. Many semantic relationships—the relationship between a sequence and its homologues, for example—are not usefully represented on the SW as static documents. It is sometimes better for these relationships to be calculated dynamically to ensure that the most up-to-date information is available. Such relationships can be instantiated by Web Services, where the Web Service provides the semantic link between a subject (the input) and an object (the output) at run-time. This web of potential links is the ‘Hidden Semantic Web’. If these hidden links could somehow be traversed by SW queries or crawlers in a manner analogous to static RDF relationships, the Hidden Semantic Web of data would be available for exploration using extensions of existing mechanisms. At present no such standard, convention or protocol exists and thus the worlds of Semantic Web Services and the Semantic Web remain largely distinct.

DISCUSSION

In a recent article about the SWLS, the question was posed ‘Are we there yet?’ [62]. We would answer, ‘We have barely started!’

We are not ‘there’ because the large majority of biomedical data and knowledge is not yet on the SW. Most data sources do not utilize stable URIs or RDF for entity identification, description, and exchange; many key ontologies are not web-accessible, and much of the content remains hidden behind computationally opaque interfaces.

Though we have emphasized the social aspect of the problems facing the development of the SWLS, it is important to note that SW technologies have a much higher complexity than the technologies that they are seeking to replace (e.g. HTML, CGI and relational databases). Even the leading groups in the field have struggled to implement SW solutions to what would be relatively straightforward problems for traditional web technologies. For example, the BioMoby and myGrid groups have spent more than 6 months, and two face-to-face meetings, defining an acceptable and extensible RDF/OWL representation for Web Service metadata (approximately 20 Triples); the BioPAX ontology has been under development since 2002 and, though early versions have been published and are already being utilized, it remains a work in progress due to both the high barrier to entry for construction of valid Description Logic ontologies (i.e. OWL), and constraints on the conceptual relationships that can be represented in the OWL language [63]. With respect to representation of, and querying over, instance data in RDF, the experiences of several groups such as the Pathway Knowledge Base (PKB [64, 65]) and CombeChem [66], indicate that there is significant room for improvement in the query algorithms over RDF graphs, where it has required extensive optimization to achieve workable response times against queries that would be trivial using analogous relational databases. Thus both sociological and technological barriers are acting to inhibit widespread adoption of SW technologies.
In the face of these high technological and social barriers, perhaps the best example of widespread institutional adoption of SW standards within the life sciences community is the development and expanding use of the BioPAX [67, 68] standard for the OWL/RDF-based distribution of biological pathway data [68]. Key data providers in this domain, BioCyc [69], Reactome [70] and KEGG [71] now provide access to their data according to this standard thus facilitating the formation of integrated pathway resources such as PKB [64].

Other pioneering projects, including UniProt-RDF [16], YeastHub [72], and CombeChem, also demonstrate the utility of the RDF paradigm for the manual and/or semi-automated aggregation of data from distributed information sources. Nevertheless, even these leading-edge projects do not fully exploit the Web aspect of Semantic Web technologies, in that their aggregated data is stored in a local warehouse, there are relatively few links to entities outside of these central repositories, and in some cases only locally resolvable URIs are utilized. As such, these projects provide interesting examples of the power of localized semantic webs, but do not yet fulfill the requirements of, nor demonstrate the power of, a singular Semantic Web for the Life Sciences.

This distinction between semantic webs in plural and the Semantic Web in singular is key to understanding why the universal application of standards for identifying, representing, and describing data is so important. The ultimate goal of the SWLS is not to create many separate, non-interacting data warehouses (as we already can), but rather to create a single, ‘crawlable’ and ‘queriable’ web of biological data and knowledge, similar to the existing WWW. To achieve this vision, we must go beyond building pre-compiled RDF warehouses, or ‘wrapping’ major content providers in SW-friendly disguises. Rather, the key players in the life sciences community need to take a leadership role and participate in the provision of their own biological data and knowledge according to SW standards.

CONCLUSION

The emergence of the SWLS is dependent on the ability of independent data providers and consumers to identify and retrieve entities on the Semantic Web, and to utilize public, web-accessible, shared and open ontologies to describe and interpret them. This vision is currently being delayed by the timid and partial creep of semantic technologies and standards into the resources provided by the life sciences community. Through adoption of proven standards for data and knowledge interchange on the Internet—URIs, RDF, OWL—these problems can largely be addressed, but this will represent a fundamental change in the way biomedical information is currently managed.

This change involves a migration away from centralized paradigms of control, and toward open and inter-linked resources, where third-parties are invited and encouraged to add-to, and thus enrich, the data space. The Semantic Web is, first and foremost, a Web-based technology; thus, as we seek to create it, all Creeps should pay heed to the experiences of those who created the original WWW. According to Berners-Lee et al. [73],

‘The essential property of the World Wide Web is its universality. The power of a hypertext link is that “anything can link to anything.” Web technology [including the SWLS], therefore, must not discriminate between the scribbled draft and the polished performance, between commercial and academic information, or among cultures, languages, media and so on.’

The Web came into being because it was universally accessible, both readable and writable. To achieve the SWLS vision, the same principles must be applied.

Acknowledgements

This work was supported by Genome Alberta and Genome British Columbia, in part through Genome Canada, a not-for-profit corporation which is leading a national strategy on genomics with $375 million in funding from the Government of Canada; and the Michael Smith Foundation/Canadian Institutes for Health Research Bioinformatics Training Program. We would like to thank the manuscript reviewers for their time and advice, Dr. Raymond Ng for his insight and comments on the manuscript, and Dr. Sean Martin for his comments on the manuscript, his enthusiasm and support for the SWLS vision, and his leadership in generating the LSID specification.

Note

last standards. To date, no significant adoption of these standards has been observed in the Life Sciences community, and thus the utility of these standards cannot be properly judged.

References

*indicates a key recent article of particular interest to the SWLS
**indicates a web resource of high value to the SWLS
***indicates a direct link to one or more SWLS ontologies

16. **Jain, E. Uniprot as RDF. http://expasy3.isb-sib.ch/~ejain/rdf/ (14 June 2006 last accessed). This website provides access to the UniProt knowledgebase in RDF/OWL format. This representation of one of the most important databases in bioinformatics on the Semantic Web is an invaluable resource for researchers in the SWLS.
18. **W3C. Semantic Web Health Care and Life Sciences Interest Group website. http://www.w3.org/2001/sw/hch/ (14 June 2006 last accessed). This website hosts discussion regarding the SWLS and provides links to many ‘early adopter’ projects. It is a key starting point for learning about the current status of the SWLS.
21. **SIMILE. Semantic Interoperability of Metadata and Information in unLike Environments. http://simile.mit.edu/ (14 June 2006 last accessed). The SIMILE group is actively developing applications such as Haystack and Piggy Bank that provide users access to SW resources. To our knowledge, this is the most active group in the world in the development of SW applications for the end user.
28. *Stevens R, Tipney H, Wroe C, et al. Exploring Williams-Beuren syndrome using myGrid. Bioinformatics 2004;20:303–10. This is one of the most complete examples of web-based in silico biology ever demonstrated. It illustrates how a biologist can, aided by the myGrid Semantic Web service platform, find answers to real, biologically relevant questions using the SWLS.
32. *Mukherjea S. Information retrieval and knowledge discovery utilising a biomedical Semantic Web. Briefings in bioinformatics 2005;6:252–62. This article provides a good introduction to the technical details of the RDF and OWL languages. It also provides some motivational examples of the application of the SWLS in the context of information retrieval.
39. *Golbreich C, Zhang S, Bodenreider O (2005). ‘Migrating the FMA from Protege to OWL’, 8th International Protege Conference, Madrid, Spain. This article describes a protocol used to translate from the Protege Frame representation to the OWL description logic. Rules for performing the translations and results of their application to the very large Foundational Model of Anatomy should prove useful to others attempting similar migrations.
63. *Ruttenberg A, Rees JA, Luciano JS (2005), ‘Experience Using OWL DL for the Exchange of Biological Pathway Information’, OWL: Experiences and Directions, Galway, Ireland. This article provides a detailed description of the problems faced during the past several years of work on creating the BioPAX ontology using OWL DL. It gives an excellent explanation of the technical difficulties of creating useful OWL ontologies for use in bioinformatics.
64. **PKB, Pathway Knowledge Base. http://pkb.stanford.edu/(3 July, 2006 last accessed). To our knowledge, PKB is the first public, integrated pathway repository to emerge out of the efforts of the BioPAX group.


81. *OBO. Open Biomedical Ontologies. http://obo.sourceforge.net (14 June 2006 last accessed). The OBO website is currently the primary source for obtaining curated bio-ontologies. Most ontologies are now provided in OBO format, but many are currently being translated into OWL.


The FMA, though not SW compatible as yet, clearly represents a profoundly important semantic resource in the biomedical domain.


89. ***MGED. MGED OWL Ontology. http://mged.sourceforge.net/ontologies/MGEDOntology.owl (14 June 2006 last accessed). An OWL ontology that can be used to represent information about microarray experiments.

90. ***UMLS. Unified Medical Language System. http://umlsks.nlm.nih.gov/(14 June 2006 last accessed). The UMLS is the largest biomedical terminologysystem in the world, subsuming most of the controlled vocabularies in the domain. Not SW compliant at the moment, it remains one of the most powerful and useful semantic resources in existence.

91. ***biozen OWL ontology framework. http://neuroscientific.net/ont/bio-zen.owl (14 June 2006 last accessed). The biozen ontology is used to integrate several other ontologies organized by the Neuronscientific initiative, including OWL versions of MESH, GO, the OBO CellType ontology, NCBITaxonomy, the Sequence ontology, and the INOH Molecule role/type ontology. This is one of the most ambitious and exciting developing projects in the SWLS.
