Semantic Web meets Integrative Biology: a survey

Huajun Chen, Tong Yu and Jake Y. Chen

Submitted: 27th December 2011; Received (in revised form): 18th February 2012

Abstract

Integrative Biology (IB) uses experimental or computational quantitative technologies to characterize biological systems at the molecular, cellular, tissue and population levels. IB typically involves the integration of the data, knowledge and capabilities across disciplinary boundaries in order to solve complex problems. We identify a series of bioinformatics problems posed by interdisciplinary integration: (i) data integration that interconnects structured data across related biomedical domains; (ii) ontology integration that brings jargons, terminologies and taxonomies from various disciplines into a unified network of ontologies; (iii) knowledge integration that integrates disparate knowledge elements from multiple sources; (iv) service integration that build applications out of services provided by different vendors. We argue that IB can benefit significantly from the integration solutions enabled by Semantic Web (SW) technologies. The SW enables scientists to share content beyond the boundaries of applications and websites, resulting into a web of data that is meaningful and understandable to any computers. In this review, we provide insight into how SW technologies can be used to build open, standardized and interoperable solutions for interdisciplinary integration on a global basis. We present a rich set of case studies in system biology, integrative neuroscience, bio-pharmaceutics and translational medicine, to highlight the technical features and benefits of SW applications in IB.

Keywords: semantic web; integrative biology; web ontology; web of data

INTRODUCTION

Integrative Biology (IB) lies at the intersection of a multitude of scientific and technological disciplines, and focuses on bridging the gap between different disciplines and the wedding of technological advances to biological insight. In the 1980s it was recognized that biology bounded by traditional disciplines no longer reflected the best way to do science, which created new...
classifications that combined two or more specialties [1]. For example, it is clear that difficult neuroscience problems like mapping gene expression in the whole brain and understanding Parkinson’s disease are too large to be accomplished unless the research of multiple groups working across disciplines can be combined [2]. As Mina Bissell recently commented, ‘Almost three decades later, we are finally ready to integrate, and indeed if the goal is to seek larger advances in biology, then we must “only connect” to other relevant scientific disciplines, especially those that can provide the tools that will give us a much better understanding of biological processes and systems’ [3].

IB is fundamentally integrative science, which adopts an interdisciplinary approach to the study of science [4]. IB typically involves interdisciplinary integration, which is a research paradigm that approaches an issue from a range of disciplinary perspectives, and the contributions of the various disciplines are integrated to provide a holistic or systemic outcome [5]. Therefore, IB needs to bring together researchers of diverse expertise to identify, articulate and structure problems, and involves intense interdisciplinary collaboration and resource sharing [4]. Specifically, IB involves the integration of data, knowledge and capabilities across disciplinary boundaries in order to solve complex problems. It is a long envisioned subject that is far from realized in biology, because of a number of disciplinary gaps such as the language gap, the knowledge gap and the collaboration gap. The language gap refers to the situation that scientists fail to understand each other’s ‘domain languages’, containing jargons, terminologies, etc. The knowledge gap refers to the fragmentation of knowledge and barriers to knowledge sharing. The collaboration gap refers to the cross-disciplinary differences (in interests, objectives and methodologies) which hinders collaboration.

Bioinformatics facilitates interdisciplinary integration with information technologies such as information sharing, knowledge management and workflow tools [6]. It also supports *in silico* experimentation with the ability to digitize biological output, and the computational power to analyze comprehensive and massive data sets [7]. With the advent of Bioinformatics, there has been an explosion of biomedical data, and their integration has proved problematic [8]. Most traditional solutions, e.g. data warehouses, can be characterized as local integration solutions, in that they can enhance the resource sharing and collaboration inside one organization or one discipline, yet fail to interoperate with each other to achieve a global solution, which is crucial to support the interdisciplinary integration.

The first truly global integration solution is the World Wide Web. In 1990, Tim Berners-Lee invented the Web, in support of the cross-boundary information sharing and collaborative research in CERN [9]. Since its inception, the World Wide Web has changed the ways scientists communicate, collaborate and educate [10]. The Web enables the development and maintenance of cyber infrastructure for e-Science, which facilitates data sharing and interdisciplinary collaborations on a global basis [11]. However, the current Web still lacks a widely-accepted and standard way to publish and share structured data, leading to the difficulty of achieving global data integration [12].

In order to fill the data gap on the Web, Tim Berners-Lee *et al.* envisioned the Semantic Web (SW) as a web of data that is meaningful and understandable to any computers [13, 14]. As they have predicted, the Web of data will enable Web users to share structured data as easy as they share documents, photos and videos today. As shown in Figure 1, the Web of data can be conceptualized as a global graph of things, or the graph layer on top of the Web [15]. *Intelligent agents* can operate directly on the Web of data in order to solve complex problems and accomplish intelligent tasks. This new layer leads to the emergence of Web 3.0 applications, which use the Web of data to augment the underlying Web system’s functionalities such as information retrieval and knowledge sharing [16].

Technically speaking, the SW is closely associated with the notion of ‘ontology’, which refers a computational model that can be used to explicitly represent the meaning of terms and the relationships between those terms [17–19]. The SW can support the collaborative engineering of domain ontologies that are shared by a community, and the use of ontologies to describe Web resources including knowledge, data and services. This approach not only enables digital resources to be shared and interconnected beyond the boundaries of applications and websites, but also supports the implementation of various machine learning and automatic reasoning methods.

Whereas SW technologies were originally designed to work globally, they were originally adopted by organizations to resolve the problems of internal integration. For example, SW technologies can be used to build a ‘semantic data warehouse’, which integrates the legacy and heterogeneous data sets internally, and
supports advanced functions such as semantic search \cite{20, 21}. SW technologies really took off with the emergence of the Linked Open Data (LOD) project, which has unleashed a revolution of data publication and interconnection in a plurality of domains such as e-education, e-health and e-science \cite{14}.

In particular, the SW has gained significant uptake in the Life Sciences to connect the various data sets in this field (e.g. Bio2RDF \cite{22}, Linking Open Drug Data (LODD) \cite{23}). A series of projects has adopted the SW technologies for the ontology integration \cite{24}, knowledge management and sharing \cite{25} and collaboration \cite{26}. Combined with semantics-driven data analysis workflow orchestration and distributed execution (e.g. Taverna \cite{27}), the new framework for in silico biomedical experimentation has the potential to add a new dimension to the way biomedical research is conducted \cite{28}. Indeed, the SW is increasingly becoming the hub of biological research, and is regarded as the promising platform for interdisciplinary integration by the bioinformatics community \cite{29}.

In this review, we provide insight into how SW technologies can be used to build open, standardized and interoperable solutions for interdisciplinary integration on a global basis. We first present a brief overview of the SW technologies and the LOD project. We then identify the major disciplinary gaps that are hindering interdisciplinary integration, and review the SW technologies that can bridge these gaps. Next, we present a rich set of case studies in systems biology, neuroscience, drug discovery and translational medicine, to highlight the technical features and benefits of SW applications in IB. Finally, we discuss the perspectives, challenges and visions of SW technologies and their applications in IB.

**Figure 1:** The architecture of the Internet contains three major levels of abstraction: Net, Web, Graph (Inspired by Tim Berners-Lee). The Internet is built on the infrastructure of telecommunications, consisting of various types of networks (Ethernet, WiFi, 3G, etc.), routers and satellites. The Net Layer was thought of as the ‘Internet Cloud’, which encapsulates the underlying communication power between computers, and allows a program on one computer to talk to a program on another computer. The Web Layer allows the exchange and sharing of Web resources while hiding the details of IP addressing and message transportation. The Graph Layer was thought of as the SW, it allows Web users to explore the connections between the things without the awareness of the Web documents. The value of this architecture is that each layer leverages the social components of the lower layer’s architecture.
SW IN A NUTSHELL
In a nutshell, the SW facilitates the integration of heterogeneous data on the World Wide Web by making the semantics of data explicit through formal ontologies [30]. The SW community has proposed core languages such as Resource Description Framework (RDF) [31], RDF vocabulary description language (RDF schema) [32], Web Ontology Language (OWL) [33] and SPARQL (a recursive acronym for SPARQL Protocol and RDF Query Language) [34]. Since 2007, the SW community has launched the LOD project, aiming to convert open data into RDF and OWL format, and publish them on the Web [14].

SW languages: RDF, OWL and SPARQL
The RDF is a language for representing information about resources in the World Wide Web [31]. RDF is based on the idea of identifying things using Web identifiers (called Uniform Resource Identifiers, or URIs) [35], and describing resources in terms of simple properties and property values. In this framework, a knowledge base (KB) contains a set of statements in the form of Subject–Property–Object triple. Subjects are in practice (though not restricted to) resources, Objects can be resources or literals and Properties define binary relations between two resources or between a resource and a literal. The intuitive meaning of a statement <S, P, O> is that the S has a property of the type P, and the property value is the O. A set of RDF triples, also called a 'RDF graph', can be encoded in RDF/XML and exchanged via the Web, enabling the sharing, integration and reuse of data on a global basis.

The SW community provides standard languages and practical tools for working ontologists [19]. The RDF vocabulary description language (RDF schema) [32] extends RDF to a resource typing system, which can be used to specify domain ontologies and complex biomedical taxonomies (such as an ‘is-a’ hierarchy). RDF schema allows classes, properties and types of resources to be explicitly declared. Generalization between classes/properties, and domain and range of properties can also be defined. In addition, the OWL adds more vocabulary for describing properties and classes: among others, relations between classes (e.g. disjointness), cardinality (e.g. ‘exactly one’), equality, richer typing of properties, characteristics of properties (e.g. symmetry) and enumerated classes [33]. There is also a rich set of practical tools that support the engineering of Web ontologies. For example, Protégé is a free, open source ontology editor and KB framework that supports a variety of formats including RDF(S) and OWL, and is widely adopted by life scientists [36]. In summary, the SW community has established a coherent ontology infrastructure for the representation, publishing and merging of shared ontologies in a decentralized manner.

A RDF Triple Store is a database that is specialized in the storage and retrieval of RDF graphs. Triple Stores that are widely used include Jena TDB [37] and Sesame [38]. An application developer can store RDF data in a Triple Store and retrieves it via SPARQL queries. SPARQL is the query language for the SW, providing the ideal and standard way to query large amount of machine-readable data between heterogeneous systems over the Internet. A SPARQL query essentially specifies a graph-matching pattern against RDF graphs. Besides querying single RDF graphs, SPARQL also provides for querying sets of Named Graphs. The SPARQL languages are explained in detail in the SPARQL Recommendation [34].

The Linked Data
The basic idea of Linked Data is to apply the general architecture of the World Wide Web [39] to the task of sharing structured data on global scale [14]. Tim Berners-Lee introduced the term Linked Data in 2006, and proposed the following Linked Data principles [40]:

(i) Use URIs as names for things.
(ii) Use HTTP URIs, so that people can look up those names.
(iii) When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
(iv) Include links to other URIs, so that they can discover more things.

In 2007, the W3C initiated the LOD community project to realize the SW vision by publishing various open data sets according to the linked data principles. The existing data sets may be in different formats, such as XML files, spread sheets, micro-formats and relational databases. In order to serve them as Linked Data on the Web, they must undergo an ‘RDFizing’ process that converts heterogeneous data into RDF. In the cases where data is stored in a relational database, it is desirable to retain the existing data management infrastructure and software, so as not to
disrupt legacy applications, and instead simply publish a Linked Data view of the relational database [14]. One widely used tool designed for this purpose is D2R Server [41]. As of September 2011, the resulting Web of data, also known as the LOD cloud, contains 295 data sets, 31,634,213,770 RDF triples and 503,998,829 RDF links [42].

A rich set of tools, such as semantic browsers and semantic search engines, are created to consume linked data. The following Linked Data browsers can serve as the entry points for the Linked Data: Tabulator [43], Marbles [44] and Disco [45]. A number of search engines aggregate the Linked Data from the Web by following RDF links and provide query capabilities for Web clients, such as Sig.ma [46], Falcons [47] and SWSE [48].

The benefits of SW technologies

As we have mentioned, the major benefit of SW technologies is to achieve data integration. Traditional solutions to data integration include data warehouses, data marts and data federations. Most of these technologies are centralized in nature, and not scalable for the global data integration. By contrast, the SW relies on a distributed, use-as-you-go approach to data integration, which enables the integration of data between different parties worldwide. Whereas the SW is often seen as a global database, it is not going to replace the traditional relational databases, but to provide a platform for the publishing and interlinking of relational databases. SW technologies and standards achieve an interoperable representation of data and the seamless integration of data from different sources. They also provide the languages for expressing the meaning of resources (data, information, documents, links, etc.) in a machine-processable way. Together, these two aspects facilitate the sharing of data and allow their accurate interpretation [8] when they are passed between different communities of different background or levels of expertise.

In addition to data integration, the SW also facilitates the integration of ontologies, experimental results, knowledge and service descriptions, etc. All these digital resources are expressed as data, and therefore data integration lays at the foundation of all integration solutions. Indeed, computer scientists are exploring the possibilities to combine SW technologies with other Web-based technologies (e.g. service-oriented architecture [49], grid computing [50] and cloud computing [51]), to create more powerful integration solutions.

Finally, as Berners-Lee et al. predicted in 2001, the machine-understandable content on the SW will unleash a revolution of intelligent agents [13]. In Artificial Intelligence (AI), an intelligent agent is an autonomous entity which observes and acts upon an environment and directs its activity towards achieving goals [52]. The SW community, which has a close tie with the AI community, has been actively explored the possibility of implementing intelligent agents on the Web (referred to as SW agents) [53]. For example, project Halo are developing SW technologies, e.g. Semantic MediaWiki (SMW+) [54] and Semantic Inferencing on Large Knowledge (SILK) [55], towards the ultimate goal of creating a ‘Digital Aristotle’ (a reasoning system capable of answering novel questions and solving advanced problems) that can serve as a research assistant with broad, interdisciplinary skills to help scientists and others in their work [56, 57]. As shown in Figure 2, a typical SW agent contains the following major components: a KB, a reasoner and a SW connector. SW agents can access to data from a wide range of data sources, and communicate with each other via the SW, which could make them much smarter than agents with closed KBs. A key feature of an SW agent is that it would not simply exploit a predetermined set of information sources, but would search the LOD cloud for relevant information in much the same way that a human user might do when planning a vacation [58]. Another key feature of SW agents is automatic reasoning, which means the generation of new triples from existing ones based on several rules. The SW supports OWL reasoners such as Racer [59] and Pellet [60], and rule-based reasoners such as Prova [61] and Jena [37]. These reasoners are already being successfully used in many applications.

As far as IB is concerned, SW agents can be a powerful personal assistant for biologists, and facilitate integrative studies. The adoption of agent technologies

![Figure 2: The basic structure of SW agents (Inspired by Danny Ayers). An agent maintains an RDF model as its KB, use a reasoner to accomplish intelligent tasks such as problem-solving and knowledge inference and communicate with the SW via HTTP protocol. The SW serves as a platform for inter-agent communication and collaboration.](image-url)
and multi-agent systems constitutes an emerging area in systems and computational biology [62]. SW agents can participate in this process, by performing a variety of functions. They can facilitate knowledge discovery and management; support systems modeling and simulation, look up trusted healthcare services, retrieve medical records, check medical insurance, schedule therapy appointments and so forth. A concrete use case is GoPubMed, an ontology-based literature search engine empowered by the agent technology [63]. In this system, agents automatically generate semantic annotations for PubMed literature abstracts in terms of Gene Ontology (GO), and effectively transform textual contents into SW contents. GoPubMed also allows task automation by providing agents with machine-understandable knowledge.

The current SW agents have shown various characters of ‘intelligence’ such as reasoning, learning and question-answering, but they are mostly prototypes. As Hendler J. has commented, there has been much research and talk about intelligent agents, but few real-world implementations [64]. A series of research issues need to be addressed, in order to build practical SW agents for integrative studies.

SW MEETS IB
The central theme of IB is to remove the barriers of interdisciplinary integration, such as the language gap, the knowledge gap and the methodological gap. The root of these gaps is the data gap, and the fundamental solution lies in global data integration. Just as the Web has brought about a revolution in the publication and consumption of documents, the SW has the potential to enable a revolution in how data is accessed and utilized, and help to bridge the disciplinary gap more effectively. Since the inception of the SW in 2001, there has been a growth in the applications of SW technologies in life sciences, a majority of which are related to IB. In 2005, the World Wide Web Consortium (W3C) launched an interest group named Health Care and Life Sciences Interest Group (HCLS IG) to coordinate these activities [65]. In this section, we first outline the activities of the HCLS IG, and then discuss the role of SW applications in supporting IB.

The HCLS IG community
The HCLS IG community aims to develop, advocate and support the use of SW technologies for translational medicine and its three enabling domains: life sciences, clinical research and health care. The HCLS IG has developed a set of SW applications (most of which are prototypes and demos) that demonstrate the value of formalizing and sharing knowledge using SW technologies. As a major task force of the HCLS IG, Linked Life Data (LLD) aims to use SW technologies (e.g. RDF and OWL) to represent, publish, query and integrate the data and knowledge in life sciences [66, 67]. The widely used data resources, including UniProt (the Universal Protein Resource) [68], KEGG (the Kyoto Encyclopedia of Genes and Genomes) [69] and CAS (the Chemical Abstracts Service) [70], are available in different formats including relational databases, structured flat files, HTML and XML. LLD focused on converting these data sets into RDF and OWL, and integrating them into KBs that support intelligent query and search. The LLD group has constructed a prototypical KB named HCLS KB, which demonstrates the effectiveness of SW technologies through use cases such as linking the data sets between Traditional Chinese Medicine and biomedicine [71, 72]. This group also launched a prototype service named LLD, which enables Web users to perform complex SPARQL queries and explore over RDF statements from various sources [73]. The HCLS IG community is also engaging in the development of Web ontologies and applications in various domains such as systems biology, translational medicine and drug discovery.

Bridge the data gap
As we have mentioned, the biological community attempted to use SW technologies to address the problem of data integration [74]. This process can be roughly divided into the pre-LOD period (2000–06) and the LOD period (2007–today). The pre-LOD period was characterized by the building of semantic data warehouses, which represent, store and query both metadata and data across life sciences data sets using SW technologies. YeastHub is a data warehouse allows integration of different types of yeast genome data provided by different resources in different formats including the tabular and RDF formats [20]. Once the data are loaded into the data warehouse, RDF-based queries can be formulated to retrieve and query the data in an integrated fashion. Other data integration efforts with similar approach include the FungalWeb [75], the BioLit [76], etc.

The LOD period was characterized by the publishing of open biological data sets on the Web.
according to the Linked Data principles [14]. As of
September 2011, the LOD cloud contains 41 data
sets in Life sciences, including 3 036 336 004 triples
(9.6% of total LOD triples) and 191 844 090 RDF
links (38.06% of total LOD links) [42]. In particular,
Belleau et al. built the SW repository named
Bio2RDF, which published a multitude of open
data resources according to the linked data rules
[22]. The Bio2RDF repository has ‘rdfized’ more
than 30 widely used data sets, including:

- Human Genome databases, e.g. NCBI Entrez Gene
- Protein databases, e.g. KEGG (the Kyoto Ency-
clopedia of Genes and Genomes) and PDB
(Protein Data Bank) [77]
- Pharmacogenomics databases, e.g. pharmGKB [78]
- Chemical informatics database, e.g. CAS (the
Chemical Abstracts Service) [70], PubChem [79].

This integrated repository is openly available as a
part of the LOD cloud, and has been used in use
cases such as exploring the implication of four tran-
scription factor genes in Parkinson’s disease.

Bridge the language gap
A domain-specific language is a language system
dedicated to a particular problem domain, consisting
of jargons, idioms, terminologies, etc. Experts from
different disciplines fail to understand each other’s
language or concepts, and even experts from the
same discipline can develop different ‘dialects’. The
language gap becomes a serious problem when
scientists want to share scientific data with their
descriptions. One approach to bridge this gap is
through the collaborative engineering of shared
donain ontologies, which have moved from a
niche activity to one that is, in all respects, a main-
dstream activity within bioinformatics [18, 80]. A
successful example is the use of the GO [81] to an-
notate the data being generated by high-throughput
technologies. BioPortal [82] is a central repository for
accessing a large collection of biomedical ontologies,
such as the GO, the Medical Subject Headings
(MeSH) [83], the NCBI Taxonomy [84], the cell-
type ontology [85] and the sequence ontology [86].
Biological ontologies are used in the search and
query of heterogeneous biomedical data, the represen-
tation of encyclopedic knowledge and computer
reasoning with data [80].

The community of bioinformatics aims to integrate
these ontologies, and provide an expanding family of
ontologies that are interoperable and logically
well-formed and incorporate accurate representations
of biological reality [24]. Traditional approaches failed
to meet this goal. For example, the Unified Medical
Language System (UMLS) is a compendium of some
100 source vocabularies, for applications such as
indexing and retrieval of clinical documents [87].
However, the vocabularies in UMLS were not
refactored into a common structure. Therefore,
UMLS is not a coherent language system, and remains
a federation of heterogeneous components.

The major benefit of SW technologies is to en-
force semantic integration of biological ontologies.
They have been adopted in a few large-scale ontol-
yogy platforms such as the NCBO (National Center
for Biomedical Ontology) [88]. NCBO Resource
Index provides a ‘semantic mashup’ of more than
200 publicly available ontologies in order to support
integrated exploration of biomedical the knowledge
resources. In addition, the OBO foundry (Open
Biomedical Ontologies consortium), a large-scale
collaborative ontology engineering project, adopts
SW technologies to achieve the interoperability of
biological ontologies.

There are two basic approaches of ontology integra-
tion: retrospective mapping and prospective standard-
ization. Retrospective mapping is the approach of
mapping existing ontologies into SW ontologies, and
integrating existing biomedical ontologies based on
foundational ontologies such as the Descriptive
Ontology for Linguistic and Cognitive Engineering
(DOLCE) [89] and Basic Formal Ontology (BFO)
[90]. There were efforts to convert GALEN [91],
OBO [92] and the UMLS Semantic Network [93] to
OWL. Hoehndorf et al. implemented a free software
that converts ontologies in the OBO Flatfile Format to
OWL, and also provide a prototype to extract rela-
tional patterns from OWL ontologies using automated
reasoning [94]. Notably, Samwald et al. describe an
ontology framework called bio-zen, which provides a
sound ontological basis for the life sciences through the
tailoring and integration of several existing ontologies
in the Open Biomedical Ontologies repository [95].
The ontology framework adheres to the OWL
format and reuses existing foundational ontologies
like DOLCE. As shown in Figure 3, Bio-zen adopts
the design pattern of separating ‘realist’ ontological
descriptions and ‘conceptual’ taxonomies and concept
hierarchies, which provides guidelines for other ontol-
ogy engineering projects aiming to merge realistic
ontologies with taxonomies. In addition, Smith et al.
promoted interoperability of ontologies by engineering the Relation Ontology (RO), which provides consistent and unambiguous formal definitions of the relational expressions [96].

Prospective standardization is the approach of setting up principles, guidelines and systems to engineer new ontologies that are complaint to the SW standards. For example, the OBO foundry attempted to achieve ontology interoperability based on the voluntary acceptance by its participants of an evolving set of principles [97]. An ontology that abides by these principles is the Ontology for Biomedical Investigations (OBI), which is a cross-disciplinary, integrated ontology for the detailed description of biological and clinical investigations [98]. OBI uses the OWL language to define a set of broadly applicable terms that span biomedical and technological domains, and reuses other OBO ontologies wherever possible. Brinkman et al. demonstrate how OBI can be used to integrate different biomedical investigations in order to facilitate interpretation of the experimental process, through use cases such as neuroscience investigation, vaccine protection investigation, an automated functional genomics investigation [98].

Bridge the knowledge gap
Life scientists rely on several forms of knowledge assets, including publications, experimental data, domain-specific vocabularies and policies [99]. They need help in coping with the plethora of fast growing and scattered knowledge resources. Ideally, this knowledge should be integrated in a form that allows scientists to pose complex questions that address the properties of biological systems, independently from the origin of the knowledge [100].

SW technologies can facilitate the integration of desperate and heterogeneous knowledge resources by associating them with formal semantics. Specifically, SW technologies can be used to connect the data and the article describing the data, or to connect published experimental results with a particular biological database entry. For example, BioLit is a Web server that integrates biological literature with databases, by generating semantic annotations for the open access documents [76]. This resource aims to integrate scientific publications directly into existing biological databases, thus obviating the need for a user to search in multiple locations for information relating to a specific item of interest. In addition, ontologies can be used in combination with text mining to extract structured knowledge from textual documents, which supports effective knowledge retrieval on the semantic level [101]. For example, Sarntivijai et al. adopted SW technologies to standardize cell line names and to facilitate biomedical research. They defined a Web ontology named Cell-line ontology, extracted useful information from biomedical text, and created a cell line knowledgebase (CLKB) with a well-structured collection of names and descriptive data for cell lines cultured in vitro [102, 103].

Bridge the collaboration gap
Interdisciplinary collaborations can be problematic due to cross-disciplinary differences in interests, objectives and methodologies [104–107]. Web services have been proved very effective to support the collaboration within a distributive and multidisciplinary team [11]. Also, the SW allows scientific data and services to be explicitly described in order to eliminate misunderstandings between team members. Therefore, scientists are exploring the possibilities to integrate Web services with the SW to better to facilitate interdisciplinary collaborations in biology [108].

The W3C defines a ‘Web service’ as ‘a software system designed to support interoperable machine-to-machine interaction over a network’ [109]. Web services enable application programs to communicate in ways that are independent of specific platforms and languages, and therefore facilitate system interoperability [110]. Web services can be used to implement a cyber-infrastructure according to Service-Oriented Architecture (SOA), which is defined as ‘a set of components which can be invoked, and whose

Figure 3: The ontology design pattern to separate ‘things’ from ‘concepts of things’ (used in the bio-zen framework). The world of real things located in a certain space and time and the world of abstract concepts about things. Both worlds can only be connected through the ‘described-by’ property—otherwise, they are completely separated.
interface descriptions can be published and discovered \cite{111, 112}. A service-oriented cyber-infrastructure turns the development of e-Science applications into a pipeline of service development, service deployment, and service combination. There are two major classes of Web services: (i) RESTful Web services, in which the service is based on the representational state transfer (REST) paradigm; and (ii) arbitrary Web services, in which the service may expose an arbitrary set of operations \cite{113}.

Semantic Web Services (SWS) are at the intersection of the SW and Web services \cite{49}. SWS aims to address the challenges faced by SOA, by adding semantics to Web services standards \cite{114, 115}. SWS is characterized by the use of shared ontologies, such as OWL-S \cite{116} and Web Service Modeling Ontology (WSMO) \cite{117}, to model various aspects of Web services, including service interfaces, service messages and service structures, which enables the discovery, composition and invocation of services in an automatic and ad hoc manner \cite{118}. In particular, the SW technologies can be well integrated with RESTful web services \cite{108}. Whereas SWS technologies have several technical advantages, several real-world issues, such as authentication and authorization, must be solved before these technologies gain widespread use \cite{108}.

MyGrid is a large-scale e-Science platform designed to 'help e-Scientists get on with science and get on with scientists' \cite{119}. MyGrid has been used in domains as diverse as plant biology, proteomics and systems biology, and fill the collaboration gap between scientists worldwide effectively. MyGrid utilizes OWL ontologies to generate semantic annotations for Web services and data resources, in order for automatic service discovery, and full utilization of resources. The MyGrid team produces and uses a suite of tools: (i) BioCatalogue \cite{120}, a directory service for service annotation and registration; (ii) MyExperiment \cite{121}, a platform for sharing workflows and experimental data; (iii) Taverna \cite{27}, a workflow design, editing and execution tool. By using SWS, MyGrid provides personalized and collaborative functions for the creation of e-laboratories in biological research.

Besides MyGrid, there are a set of project that integrates SWS technologies with biological applications. BioMOBY is a large-scale attempt to integrate multiple resources using web services \cite{122}. The Moby 2.0/CardioSHARE framework is a framework for RDF-based Web Services, and aims to provide higher-level functionality for querying and reasoning against its services \cite{123}. These projects show that SWS enables scientific teams to build applications in a rapid and flexible manner, and facilitate on-demand collaboration across disciplinary boundaries.

**TYPICAL CASE STUDIES**

**SW meets systems biology**

Systems biology is an interdisciplinary domain that aims to offer a holistic view of the way in which biological systems work \cite{124, 125}. It seeks to explain biologic phenomenon, not on a gene-by-gene basis, but through the net interactions of all cellular and biochemical components within a cell or organism. To fully map biological systems and explore the cellular machinery that drives biological processes, the heterogeneous data and multi-domain knowledge needs to be brought together \cite{90}.

SW technologies can be used to design the standards of ‘omic’ data, which is of paramount importance to realize the promise of systems biology \cite{124}. The HCLS IG community predicted that systems biology would be among the earliest adopters of the SW technologies due to its highly integrative interdisciplinary nature \cite{126}. Notably, Antezana et al. proposed Semantic Systems Biology (SSB) as an approach that uses semantic description of knowledge about biological systems to facilitate integrated data analysis \cite{28, 127}. SSB would potentially evolve into a global platform for resource sharing and integration in systems biology.

A series of projects adopt SW technologies to facilitate the engineering of Web ontologies for the storage, exchange and integration of biological models \cite{128, 129}. The bio-zen ontology, which is mentioned above, allows the seamless integration of mathematical descriptions and simulation parameters into qualitative information, making a quick transition from plain data to model simulations possible. This gives bio-zen the power to act as a modeling language similar to the popular Systems Biology Markup Language (SBML \cite{130}) and CellML \cite{131}. Courtot et al. describe three ontologies created specifically to address the needs of the systems biology community, including the Systems Biology Ontology (SBO), the Kinetic Simulation Algorithm Ontology (KiSAO) and the Terminology for the Description of Dynamics (TEDDY) \cite{129}. These ontologies together provide semantic information about the model components, the simulation of systems biology models, the simulation results and general systems behavior. In addition, Visser et al. developed the BioAssay Ontology (BAO), an ontology.
that describes High-throughput Screening experiments and screening results using expressive description logic [132]. Schürer et al. utilize the BAO for the cross-analysis of diverse high-throughput screening data sets [133].

In addition, SW technologies were used to integrate knowledge in systems biology. For example, BioGateway provides a KB holding data from the diverse public sources such as the GO annotation files, the SWISS-PROT protein set and the OBO foundry candidate ontologies [100]. BioGateway provides a single entry point to query these resources through SPARQL. Knowledge resources on the SW allow the deployment of advanced computational reasoning approaches. In addition, Splendiani et al. developed a SW framework named RDFScape, which is a Cytoscape plugin that facilitate biological analysis with ontology-based reasoning capacity [134]. RDFScape demonstrates that machines can take advantage of the SW content to generate new hypotheses about the functionality of biological systems.

**SW meets integrative neuroscience**

Neuroscience is an interdisciplinary science of the nervous system, and is critical to the understanding of chronic diseases such as Parkinson’s disease and Alzheimer’s disease (AD) [135]. There has been a rich set of projects that focus on applying information technologies in neuroscience, especially for data integration. For example, the Biomedical Informatics Research Network (BIRN) applied spatial systems and ontologies for proper modeling of neuroscience data and their use in a large-scale data integration effort [136]. The Alzheimer Research Forum (Alzforum, www.alzforum.org) is an online community that is widely used by professional AD researchers for knowledge sharing and scientific discourse [137, 138]. SenseLab is a highly accessed information resource for neuroscience research on the Web [139]. These projects lead to the emergence of e-Neuroscience.

The HCLS IG community has taken a series of efforts to integrate SW technologies into e-Neuroscience, demonstrating the feasibility and benefits of using SW technologies in Neuroscience [140]. For example, The SWAN project (SW Applications in Neuromedicine) aims to develop a practical, common, semantically-structured, framework for scientific discourse [141]. SWAN has built on Alzforum’s successful ten-year history as a scientific web community and strong social network, and helps scientists to organize, manage, share and compare their knowledge related to AD [142]. The SWAN enables scientists to organize their personal knowledge as a web of assertions whose relationships to each other and to their supporting evidence is well-characterized, giving rise to a semantically-structured network of hypotheses, claims, dialogue, publications and digital repositories. Users can carry out the following tasks with the help of the system: (i) Understand what kind of problems can be solved by one’s research results; (ii) Understand what kind of hypotheses can be proved or falsified by one’s discovered evidences; (iii) Understand the relationships between different hypotheses and evidences (Does Hypothesis A and B coincide with each other? Can evidence C support Hypothesis A?); (iv) Identify open problems that needs to be solved in one’s domain; and (v) integrate knowledge units into research and clinical plans.

In addition of SWAN, SW technologies are frequently used in the context of neuroscience. Ruttenberg et al. developed a KB called Neurocommons, and demonstrate the utility of the KB through a few use cases in which one interact with the KB by posing precise queries [143]. Sanwald et al. semi-automatically translated several of SenseLab suites of neuroscience databases into OWL ontologies with manual addition of semantic enrichment. Entrez Neuron is a web portal that enables neuroscience database annotation and integration based on the ontologies, and provides an easily accessible and intuitive web user interface for neuroscientists. These works have demonstrated the effectiveness of SW technologies, which will continuously contribute to e-neuroscience.

**SW meets integrative bio-pharmaceutics**

The life sciences ‘omics’ revolution has the potential of boosting the drug discovery process. In particular, the progress of systems biology enables biopharmaceutical research paradigm to be shifted towards more comprehensive systems-based understandings of drug action [144]. The major obstacle of the ‘genes-to-drugs’ translation is dealing with the volume and diversity of data generated [145, 146].

We illustrate the SW’s advantages through a case study in pharmacogenomics [145], which is an emerging field that aims to translate functional genomics into rational therapeutics, and support individualized medicine [147, 148].There is a set of valuable resources for pharmacogenomics, including PharmGKB [149], DrugBank [150], PubChem [151], UniProt [68]. A series of projects aimed to develop ontology-based integration solutions that integrate these resources to support knowledge discovery. Coulet et al. developed
a formal ontology called SO-Pharm, which provides a comprehensive and integrated representation of domain knowledge in pharmacogenomics [152]. To achieve this goal, SO-Pharm articulates ontologies from subdomains of pharmacogenomics (i.e. genotype, phenotype, drug, trial representations), and enables the representation of knowledge about pharmacogenomics hypothesis, case study and investigations in pharmacogenomics. The SO-Pharm effort offers a first step towards representing and integrating pharmacogenomics (and related) knowledge with OWL. As a simpler alternation to SO-Pharm, the Pharmacogenomics Ontology (PO) also provides effective knowledge representation for pharmacogenomics knowledge [145]. PO identifies 40 core concepts spanning drugs, genotypes, phenotypes and drug treatments. Dumontier et al. created a KB by populating the PO using PharmGKB web services and demonstrated its utility in answering sophisticated questions about pharmacogenomics knowledge.

In addition, the LOD cloud can be used to integrate data more effectively across all drug discoveries and development business units. The LODD project is a project conducted by the HCLS IG [23]. LODD has brought together the pharmaceutical companies Eli Lilly, AstraZeneca and Johnson & Johnson, in a cooperative effort to interlink openly licensed data about drugs and clinical trials, in order to aid drug-discovery. LODD has published a series of data resources in compliance with the linked data principles, and established their links to other parts of the LOD cloud.

A critical question is whether the integration capabilities provided by the SW provides tangible benefits to drug discovery. The HCLS IG has developed BioDash, a prototype of a drug development dashboard that demonstrates the principles of and advantages of SW [153]. Multiple forms of knowledge, including genomic, pathway, disease and single nucleotide polymorphism (SNP) data, can be brought together into useful, aggregated displays through SW approaches to support the discovery process. BioDash's topic view visualizes the discovery efforts underway regarding a specific gene, and BioDash's pathway view can be used to navigate pathways in which a gene participates. In addition, Stephens et al. demonstrates the usability of SW technologies in drug safety determination [154].

**SW meets translational research**

In 2002, US National Institute of Health (NIH) proposed a roadmap to strengthen translational research, defined as the movement of discoveries in basic research (the Bench) to application at the clinical level (the Bedside) [155]. Translational research is a driving force for personalized medicine, in which research institutions, hospitals and pharmaceutical companies would be gradually integrated into a boundaryless virtual organization that delivers personalized healthcare services to patients. Translational medicine requires the integration of knowledge using heterogeneous data from health care to the life sciences [126].

The HCLS IG community recognized the SW as a promising approach to eliminate the boundaries imposed by the traditional disciplinary structure, and to accelerate the translation of the findings in basic research into medical practice and meaningful health outcomes [26, 126]. Therefore, it has established the task force of Translational Medicine, which aims to demonstrate how information-based translational medicine activities can be made easier and more effective using SW technologies. The major works of this task force include Translational Medicine Ontology (TMO) and Translational Medicine Knowledge Base (TMKB) [156, 157].

TMO aims to drive personalized medicine by bridging the language gap from bedside to bench [157]. TMO provides terminology that bridges diverse areas of translational medicine including hypothesis management, discovery research, drug development and formulation, clinical research and clinical practice. TMO provides a foundation upon which chemical, genomic and proteomic data may be linked to disease, treatments and electronic health records.

TMKB is a prototypical KB capable of answering questions relating to clinical practice and pharmaceutical drug discovery [156]. TMKB uses SW technologies to integrate patients’ data with biomedical knowledge based on the TMO ontology. TMKB can aid physicians in providing tailored patient care, and facilitates the recruitment of non-responsive patients into active clinical trials. Thus, patients, physicians and researchers may explore the KB to better understand therapeutic options, efficacy and mechanisms of action. The TMKB project demonstrates the use of SW technologies to facilitate integration of relevant external sources.

In addition to TMO and TMKB, scientists have done various other works to lay the foundation for interdisciplinary collaboration in translational research [26]. For example, Holford et al. developed a SW framework to integrate cancer omics data with
biological knowledge, which allows us to pose significant translational medicine questions [158]. Splendidiani et al. established the DC-THERA directory, which is a web portal designed to address the collaborative and sharing needs of the DC-THERA community [19]. These works demonstrate the feasibility of using the SW to model and share adaptable clinical pathways and protocols, which serve to translate results of research and clinical trials to application in patient care.

**PERSPECTIVES, CHALLENGES AND VISIONS**

In 21 century, we confront some important and significant problems concerning our living conditions and fundamental interests, such as environmental crisis, unhealthy life styles and chronic diseases. All of these problems are related to biology and cannot be answered by any single discipline alone. Members of different disciplines must engage in meaningful dialogues and collaborations in order to achieve a clear and common understanding of these problems. We must utilize the knowledge and tools from other relevant scientific disciplines, in order to explore new insights into biological processes and systems, and translate these insights into practical solutions. Therefore, interdisciplinary integration, including the integration among scientific disciplines, and also between science and technology, becomes increasingly important.

**The SW for interdisciplinary integration**

The SW is an extension to the current Web, in which information is given well-defined meaning, better enabling computers and people to work in cooperation [13]. The major goal of the SW is to maximize the ‘interoperability’ of the internal resources with external resources, so as to maximize its usefulness and visibility, beyond the boundaries of the specific research network that was initially served [26]. As SW technologies are maturing, it is important to analyze how the requirements of interdisciplinary integration can be met by the SW.

First, the SW can facilitate IB through Web-scale data integration. The SW principles and practices have been adopted by an increasing number of research organizations, resulting in the creation of a global data space on the Web containing billions of RDF triples which reflect the biological reality. The LOD cloud is entering the threshold of exponential growth, and we expect to see the size of LOD cloud doubles every year in the near future. Therefore, the LOD cloud might become the portal for data analysis and mining in 5–10 years, just as the Web has been the portal for scientific papers.

Second, the SW can facilitate ontology-based knowledge integration [159] and the integration of bio-ontologies themselves. SW technologies prove to be well suited for the creation, integration, maintenance and querying of biological knowledge. The current SW can potentially evolve into a fine-grained global knowledge network that connects semantic facts, hypothesis, evidences, rules and experimental data, which will become an irreplaceable utility for IB [99].

Third, the SWS can potentially support Web-scale collaboration. SOA represents a promising approach to integrating data and software across different institutional and disciplinary sources, thus facilitating Web-scale collaboration while avoiding the need to convert different data and software to common formats [160]. Based on SOA, workflow tools facilitate scientific experiments by accelerating service discovery, composition and orchestration tasks. SWS enhance the existing Web services and Workflow tools with modeling and reasoning capabilities, thus better satisfying the requirements of interdisciplinary collaboration.

**Limitations of SW technologies**

As a young technology, the SW has limitations to satisfy the requirements of IB. First, the LOD cloud still needs to provide enough incentives for biological organizations to publishing their valuable data resources, instead of locking them in organizational data warehouses. The integration of clinical data (e.g. electronic health records) with publicly accessible knowledge creates new opportunities for integrative studies and personalized medical care, but only limited amounts of clinical data are available for research purposes, and even the available data are under-utilization due to the use of natural language text and local coding schemes. We will discuss some issues that can be united into the major theme: making LOD a healthy platform so that it is worthwhile to publish biological data on the platform.

First, IB requires the generation of semantic links among data sets. As a mechanism of semantic link discovery and maintenance is the key to hold the biological data space together as a giant cluster, and to keep it from scattered into a multitude of data islands. However, to maintain semantic links in such a dynamic and decentralized environment is a difficult problem.
Second, integrative studies typically involve the sharing of sensitive data, and pose a great demand for access control mechanisms, which is currently not specified in the SW community. Access control is a key technical mechanism to ensure data security in a collaborative environment. Notably, Deus et al. proposed S3QL, a distributed domain-specific language for controlled semantic integration of life sciences data [161]. S3QL supports a permission control mechanism that allows users to specify contextual minutia such as provenance and access control on the semantic level. The effectiveness of S3QL was illustrated through use cases of IB, such as genomic characterization of cancer and molecular epidemiology of infectious diseases. We expect S3QL or its variations to be accepted as the standard access control mechanism by the SW community.

Another concern of IB is that the global open access of data would risk medical privacy. In personalized medicine, caregiver networks provide support to the patient in the community, through a personal health record (PHR). One critical problem is how to generate a virtual PHR through the integration of multiple data sources, while preserving patient privacy. Fox et al. attempt to use Semantic Mashup technologies to solve this problem [162]. They provide a mashup maker called Sqwelch, which enables trusted collaboration between a caregiver network’s members through a virtual, distributed PHR. Sqwelch provides an intuitive means for the caregiver network to create personalized mashups, while the patient retains privacy control through trust specifications. Sqwelch demonstrates the ability of SW to protect medical privacy.

CONCLUSIONS

Although meaningful molecular level models of human cell and tissue function are a distant goal, systems biology efforts are already influencing healthcare and drug discovery. The ultimate goal of the SW is to create a single, ‘crawbable’ and ‘queryable’ web of biological data and knowledge, similar to the existing WWW [163]. With the efforts of the HCLS IG community, the SW has made rapid progress towards this goal in the recent years, and will evolve into the platform for interdisciplinary integration in biology. Intelligent agents will be able to work on top of the global data space, and facilitate biological research and decision-making. This vision, when realized, will dramatically improve our ability to conduct integrative studies using the vast and growing stores of web-accessible resources.

Key Points

- IB focuses on bridging the gap between different disciplines and the wedding of technological advances to biological insight, and typically involves the integration of the data, knowledge and capabilities across disciplinary boundaries in order to solve complex problems.
- The SW enables people to share content beyond the boundaries of applications and websites, resulting into a web of data that is meaningful and understandable to any computers.
- SW technologies can be used to build open, standardized and interoperable solutions for interdisciplinary integration on a global basis with typical applications in system biology, integrative neuroscience, bio-pharmaceutics and translational medicine.
- The merging of the SW and IB will remove a number of domain-specific gaps such as the language gap, the knowledge gap and the methodological gap.

FUNDING

The work of the authors is funded by China National Science Foundation (NSFC61070156), and China national 863 program China Cloud Initiative.

REFERENCES


64. Hendler J. Where are all the intelligent agents? IEEE Intel Syst 2007;11:2–3.


