

Methodology

## Advancing translational research with the Semantic Web

Alan Ruttenberg<sup>1</sup>, Tim Clark<sup>2</sup>, William Bug<sup>3</sup>, Matthias Samwald<sup>4</sup>,  
Olivier Bodenreider<sup>5</sup>, Helen Chen<sup>6</sup>, Donald Doherty<sup>7</sup>, Kerstin Forsberg<sup>8</sup>,  
Yong Gao<sup>9</sup>, Vipul Kashyap<sup>10</sup>, June Kinoshita<sup>11</sup>, Joanne Luciano<sup>12</sup>, M  
Scott Marshall<sup>13</sup>, Chimezie Ogbuji<sup>14</sup>, Jonathan Rees<sup>15</sup>, Susie Stephens<sup>16</sup>,  
Gwendolyn T Wong<sup>11</sup>, Elizabeth Wu<sup>11</sup>, Davide Zaccagnini<sup>17</sup>,  
Tonya Hongsermeier<sup>10</sup>, Eric Neumann<sup>18</sup>, Ivan Herman<sup>19</sup> and Kei-  
Hoi Cheung\*<sup>20</sup>

Address: <sup>1</sup>Millennium Pharmaceuticals, Cambridge, MA, USA, <sup>2</sup>Initiative in Innovative Computing, Harvard University, Cambridge, MA, USA, <sup>3</sup>Laboratory for Bioimaging and Anatomical Informatics, Department of Neurobiology and Anatomy, Drexel University College of Medicine, Philadelphia, PA, USA, <sup>4</sup>Section on Medical Expert and Knowledge-Based Systems, Medical University of Vienna, Vienna, Austria, <sup>5</sup>National Library of Medicine, Bethesda, MD, USA, <sup>6</sup>Agfa Healthcare, Waterloo, Ontario, Canada, <sup>7</sup>Brainstage Research, Pittsburgh, PA, USA, <sup>8</sup>AstraZeneca, Mölndal, Sweden, <sup>9</sup>MassGeneral Institute for Neurodegenerative Disease, Massachusetts General Hospital, Charlestown, MA, USA, <sup>10</sup>Partners HealthCare System, Wellesley, MA, USA, <sup>11</sup>Alzheimer Research Forum, Boston, MA, USA, <sup>12</sup>Harvard Medical School, Boston, MA, USA, <sup>13</sup>Integrative Bioinformatics Unit, University of Amsterdam, Amsterdam, The Netherlands, <sup>14</sup>Cleveland Clinic Foundation, Cleveland, OH, USA, <sup>15</sup>Science Commons, Cambridge, MA, USA, <sup>16</sup>Oracle, Burlington, MA, USA, <sup>17</sup>Language & Computing, Reston, VA, USA, <sup>18</sup>Teranode Corporation, Seattle, WA, USA, <sup>19</sup>World Wide Web Consortium (W3C) and <sup>20</sup>Center for Medical Informatics, Yale University School of Medicine, New Haven, CT, USA

Email: Alan Ruttenberg - alanruttenberg@gmail.com; Tim Clark - tim\_clark@harvard.edu; William Bug - William.Bug@drexelmed.edu; Matthias Samwald - samwald@gmx.at; Olivier Bodenreider - olivier@nlm.nih.gov; Helen Chen - helen.chen@agfa.com; Donald Doherty - donald.doherty@brainstage.com; Kerstin Forsberg - kerstin.l.forsberg@astrazeneca.com; Yong Gao - ygao@partners.org; Vipul Kashyap - vkashyap1@partners.org; June Kinoshita - junekino@alzforum.org; Joanne Luciano - jluciano@cs.man.ac.uk; M Scott Marshall - marshall@science.uva.nl; Chimezie Ogbuji - ogbujic@bio.ri.ccf.org; Jonathan Rees - jar28@mumble.net; Susie Stephens - susie.stephens@gmail.com; Gwendolyn T Wong - wonglabow@verizon.net; Elizabeth Wu - ewu@alzforum.org; Davide Zaccagnini - davide@landcglobal.com; Tonya Hongsermeier - thongsermeier@partners.org; Eric Neumann - enemann@teranode.com; Ivan Herman - ivan@w3.org; Kei-Hoi Cheung\* - kei.cheung@yale.edu

\* Corresponding author

Published: 9 May 2007

BMC Bioinformatics 2007, 8(Suppl 3):S2 doi:10.1186/1471-2105-8-S3-S2

This article is available from: <http://www.biomedcentral.com/1471-2105/8/S3/S2>

© 2007 Ruttenberg et al; licensee BioMed Central Ltd.

This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

### Abstract

**Background:** A fundamental goal of the U.S. National Institute of Health (NIH) "Roadmap" is to strengthen *Translational Research*, defined as the movement of discoveries in basic research to application at the clinical level. A significant barrier to translational research is the lack of uniformly structured data across related biomedical domains. The Semantic Web is an extension of the current Web that enables navigation and meaningful use of digital resources by automatic processes. It is based on common formats that support aggregation and integration of data drawn from diverse sources. A variety of technologies have been built on this foundation that, together, support identifying, representing, and reasoning across a wide range of biomedical data. The Semantic Web Health Care and Life Sciences Interest Group (HCLSIG), set up within the framework of the World Wide Web Consortium, was launched to explore the application of these technologies in a variety of areas. Subgroups focus on making biomedical data available in RDF, working with biomedical ontologies, prototyping clinical decision support systems, working on drug safety and efficacy communication, and supporting disease researchers navigating and annotating the large amount of potentially relevant literature.

**Results:** We present a scenario that shows the value of the information environment the Semantic Web can support for aiding neuroscience researchers. We then report on several projects by members of the HCLSIG, in the process illustrating the range of Semantic Web technologies that have applications in areas of biomedicine.

**Conclusion:** Semantic Web technologies present both promise and challenges. Current tools and standards are already adequate to implement components of the bench-to-bedside vision. On the other hand, these technologies are young. Gaps in standards and implementations still exist and adoption is limited by typical problems with early technology, such as the need for a critical mass of practitioners and installed base, and growing pains as the technology is scaled up. Still, the potential of interoperable knowledge sources for biomedicine, at the scale of the World Wide Web, merits continued work.

## Background

### **Translational research and the information ecosystem**

Starting in 2002, the NIH began a process of charting a "roadmap" for medical research in the 21st century [1], identifying gaps and opportunities in biomedical research that crossed the boundaries of then extant research institutions. A key initiative that came out of this review is a move to strengthen *Translational Research*, defined as the movement of discoveries in basic research (the *Bench*) to application at the clinical level (the *Bedside*).

Much of the ability of biomedical researchers and health care practitioners to work together – exchanging ideas, information, and knowledge across organizational, governance, socio-cultural, political, and national boundaries – is mediated by the Internet and its ever-increasing digital resources. These resources include scientific literature, experimental data, summaries of knowledge of gene products, diseases, and compounds, and informal scientific discourse and commentary in a variety of forums. Together this information comprises the scientific "information ecosystem" [2]. Despite the revolution of the Web, the structure of this information, as evidenced by a large number of heterogeneous data formats, continues to reflect a high degree of idiosyncratic domain specialization, lack of schematization, and schema mismatch.

The lack of uniformly structured data affects many areas of biomedical research, including drug discovery, systems biology, and individualized medicine, all of which rely heavily on integrating and interpreting data sets produced by different experimental methods at different levels of granularity. Complicating matters is that advances in instrumentation and data acquisition technologies, such as high-throughput genotyping, DNA microarrays, protein arrays, mass spectrometry, and high-volume anonymized clinical research and patient data are resulting in an exponential growth of healthcare as well as life science data. This data has been provided in numerous disconnected databases – sometimes referred to as data silos. It has become increasingly difficult to even discover these databases, let alone characterize them.

Together, these aspects of the current information ecosystem work against the interdisciplinary knowledge transfer needed to improve the bench-to-bedside process.

### **Curing and preventing disease requires a synthesis of understanding across disciplines**

In applying research to cure and prevent diseases, an integrated understanding across subspecialties becomes essential. Consider the study of neurodegenerative diseases such as Parkinson's Disease (PD), Alzheimer's Disease (AD), Huntington's Disease (HD), Amyotrophic Lateral Sclerosis (ALS), and others. Research on these diseases spans the disciplines of psychiatry, neurology, microscopic anatomy, neuronal physiology, biochemistry, genetics, molecular biology, and bioinformatics.

As an example, AD affects four million people in the U.S. population and causes great suffering and incurs enormous healthcare costs. Yet there is still no agreement on exactly how it is caused, or where best to intervene to treat it or prevent it. The Alzheimer Research Forum records more than twenty seven significant hypotheses [3] related to aspects of the etiology of AD, most of them combining supporting data and interpretations from multiple biomedical specialist areas.

One recent hypothesis on the cause of AD [4] illustrates the typical situation. The hypothesis combines data from research in mouse genetics, cell biology, animal neuropsychology, protein biochemistry, neuropathology, and other areas. Though commensurate with the "ADDL hypothesis" of AD etiology [5], essential claims in Lesné *et al.* conflict with those in other equally well-supported hypotheses, such as the amyloid cascade [6] and alternative amyloid cascade [7].

Consider also HD an inherited neurodegenerative disease. Although its genetic basis is relatively simple and it has been a model for autosomal dominant neurogenetic disorders for many years, [8], the mechanisms by which the disorder causes pathology are still not understood. In the case of PD, despite its having been studied for many decades, there are profound difficulties with some of the

existing treatments [9,10], and novel or modified treatments are still being developed [11,12].

Increasingly, researchers recognize that Ad, PD, and HD share various features at the clinical [13], neural [14-17], cellular [18-20], and molecular levels [21,22]. Nonetheless, it is still common for biologists in different specialties to be unaware of the key literature in one another's domain.

These observations lead us to a variety of desiderata for the information environment that can support such synthesis. It should take advantage of the Web's ability to enable dissemination of and access to vast amounts of information. Queries need to be made across experimental data regardless of the community in which it originates. Making cross-disease connections and combining knowledge from the molecular to the clinical level has to be practical in order to enable cross-disciplinary projects. Both well-structured standardized representation of data as well as linking and discovery of convergent and divergent interpretations of it must be supported in order to support activities of scientists and clinicians. Finally, the elements of this information environment should be linked to both the current and evolving scientific publication process and culture.

### **The Semantic Web**

The Semantic Web [23,24] is an extension of the current Web that enables navigation and meaningful use of digital resources by automatic processes. It is based on common formats that support aggregation and integration of data drawn from diverse sources.

Currently, links on Web pages are uncharacterized. There is no explicit information that tells a machine that the mRNA described by `<ahref="/entrez/viewer.fcgi?val=NM_000546.2">` on the Entrez page about Human TP53 gene [25] is related to TP53 in any specific way. By contrast, on the Semantic Web, the relationship between the gene and the transcribed mRNA product would be captured in a statement that identifies the two entities and the type of the relationship between them. Such statements are called "triples" because they consist of three parts – *subject*, *predicate*, and *object*. In this case we might say that the subject is *human TP53 gene*, the predicate (or relationship) is *hasGeneProduct*, and the object is *human TP53 MRNA*. Just as the subject and object – the pages describing the gene and mRNA – are identified by Uniform Resource Identifiers (URIs) [26], so, too, is the relationship, the full name of which might be `http://www.ncbi.nlm.nih.gov/entrez/hasGeneProduct`. A Web browser viewing that location might show the human readable definition of the relationship.

Since URIs can be used to describe names, all information accessible on the Web today can be part of statements in the Semantic Web. If two statements refer to identical URIs, this means that their subjects of discourse are identical. This makes it possible to merge data references. This process is the basis of data and knowledge integration on the Semantic Web.

With this as a foundation, a number of existing approaches for organizing knowledge are being adapted for use on the Semantic Web. Among these are thesauri, ontologies, rule systems, frame based representation systems, and various other forms of knowledge representation. Together, the uniform naming of elements of discourse by URIs, the shared standards and technologies around these methods of organization, and the growing set of shared practices in using those, are known as Semantic Web technologies.

The formal definition of relations among Web resources is at the basis of the Semantic Web. Resource Description Framework (RDF) [27], is one of the fundamental building blocks of the Semantic Web, and gives a formal specification for the syntax and semantics of statements (triples). Beyond RDF, a number of additional building blocks are necessary to achieve the Semantic Web vision.

- The specification of a query language, SPARQL [28], by which one can retrieve answers from a body of statements.
- Languages to define the controlled vocabularies and ontologies that aid interoperability; the RDF Schema (RDFS) [29], Simple Knowledge Organization System (SKOS) [30], and the Web Ontology Language (OWL) [31].
- Tools and strategies to extract or translate from non-RDF data sources to enable their interoperability with data organized as statements. For example, GRDDL (Gleaning Resource Descriptions from Dialects of Languages) [32] defines a way of associating XML with a transformation that turns it into RDF. There are also a variety of RDF extraction tools and interfaces to traditional databases [33].

Specifications of some of these technologies have published and are stable, while others are still under development. RDF and OWL are about three years old, a long time on the Web scale, but not such a long time for the development of good tools and general acceptance by the technical community. Other technology specifications (SKOS, GRDDL, SPARQL, etc.) will only be published as standards in the coming years – though usable implementations already exist.

Despite the youth of these technologies, active developer and scientific communities have developed around these technologies e.g. SemWebCentral [34]. Today, there are a large number of tools, programming environments, specialized databases, etc (see, e.g., [35]). These tools are offered both by the open source community and as products offered by small businesses and large corporations. Today, we are at the point at which anybody can start developing applications for the Semantic Web because the necessary development tools are now at our disposal.

#### **How can the Semantic Web help biomedical research?**

We have come to believe the judicious application of Semantic Web technologies can lead to faster movement of innovation from research laboratory to clinic or hospital. The Semantic Web approach offers an expanding mix of standards, technologies, and social practices layered on top of the most successful information dissemination and sharing apparatus in existence – the World Wide Web. Some of the elements of the technology most relevant to biomedical research include:

The **global scope of identifiers** that follow from the use of URIs offer a path out of the complexities caused by the proliferation of local identifiers for entities of biomedical interest. Too much effort has been spent developing services mapping between, for instance, the gene identifiers used by the many data sources recording information about them.

The Semantic Web schema languages, RDFS and OWL, offer the potential to simplify the management and comprehension of a complicated and rapidly evolving set of relationships that we need to record among the data describing the products of the life and medical sciences. Along with the benefits of the technologies that underlie our current data stores, there are a number of significant disadvantages that the Web schema languages remediate.

RDFS and OWL are **self-descriptive**. Scientists that integrate different types of data need to understand both what the data means at the domain level, as well as the details of its form as described in associated data schemas. Because these schemas tend to be technology and vendor specific, it is a significant burden to understand and work with them. While the need to integrate more types of data will continue, RDFS and OWL offer some relief to the burden of understanding data schemas. On the Semantic Web, classes and relationships are represented in the same way as the data. Documentation about them is uniformly discoverable due to the standardized *rdf:comment* property. In a well-designed ontology, the structure itself can often help guide users towards its correct use. Some examples of such structure are the well defined meaning of the hierarchical subclass relations, the use of properties

defined by the ontology in the construction of definitions within the ontology, and a carefully designed modularization [36].

RDFS and OWL are **flexible, extendable, and decentralized** because they are designed for use in the dynamic, global environment of the Web. RDFS and OWL support hierarchical relationships at their core, allowing for easy incorporation of subclass and subproperty relationships that are essential for managing and integrating complex data. New schemas can easily incorporate previously defined classes and properties that refer to data elsewhere on the Web without the all-too-typical copying and local warehousing of data to be built upon. When different schemas are found to have classes or properties that describe the same kinds of data or relationships, statements may be added that formally record that they should be considered the same. This allows for simpler queries that do not have to account for those equivalences.

The ability to easily extend the work of others makes worthwhile the development of ontologies that can be shared across different domains. For example, there are recent efforts to develop middle ontologies, such as EXPO [37] and the Ontology for Biomedical Investigations (OBI) [38], that are designed to model scientific experiments and investigations. Data from projects that build upon them will be easier to link together than those that use ad-hoc solutions or choose from a variety of disparate and sometimes proprietary LIMS (Laboratory Information Management Software) systems.

Reasoners for the Semantic Web schema languages introduce capabilities previously not widely available by offering the ability to do **inference, classification, and consistency checking**. Each of these capabilities has benefits across the health care and life science domains. For example, the powerful consistency checking offered by OWL reasoners can help ensure that schemas, ontologies, and data sets do not contain contradictory or malformed statements. These erroneous statements are unfortunately quite common. For example, in ongoing work merging two E. coli metabolic databases, 120 cross reference errors were found when comparing descriptions of several hundred metabolites described in both [39]. In a review of Gene Ontology (GO) term usage, up to 10% of terms used for gene annotations were obsolete [40]. When present in research data such errors can lead to missed opportunities. When present in medical records they can result in inappropriate diagnosis and treatment.

We envision the use of Semantic Web technologies will improve the productivity of research, help raise the quality of health care, and enable scientists to formulate new hypotheses inspiring research based on clinical experi-

ences. To help realize this vision, the World Wide Web Consortium (W3C) established the Semantic Web Health Care and Life Sciences Interest Group (HCLSIG) [41] which is chartered to explore and support the use of Semantic Web technologies to improve collaboration, research and development, and innovation in the information ecosystem of the health care and life science domains.

In the remainder of this paper we will describe the makeup and activities of HCLSIG, present a motivating scenario, describe efforts and issues encountered as we have explored the use of Semantic Web technologies, and discuss challenges to and prospects for the approach.

## Methods

### The HCLSIG

The HCLSIG is intended to serve as a bridge connecting the Semantic Web community's technology and expertise to the information challenges and experiences in the health care and life science communities. It pulls together scientists, medical researchers, science writers, and informaticians working on new approaches to support biomedical research. Current participants come from academia, government, non-profit organizations, as well as healthcare, pharmaceuticals, and industry vendors. The ultimate goal is that collaboration between all four groups will help facilitate the development of future standards and tools. Indeed, one objective of a Semantic Web will be to support the effective interaction between academia and industry.

The HCLSIG's role in the effort to create the bench-to-bed-side model is to experiment with the application of such standards-based semantic technologies in working with biomedical knowledge. A primary goal is to enable the dynamic "recombining of data", while preserving the layers of meaning contributed by all the participating research groups.

The group's scope is for two years, continuing through the end of 2007. It was chartered with three specific objectives in the domain of Health Care and Life Sciences.

- Identification of core vocabularies and ontologies to support effective access to knowledge and data.
- Development of guidelines and best practices for unambiguously identifying resources such as medical documents and biological entities.
- Development of proposals and strategies for directly and uniformly linking to the information discussed in scientific publications from within those publications – for

example the data, protocols, and algorithms used in the research.

The HCLSIG adopts a community-based approach to fostering discussions, exchanging ideas, and developing use cases. It also facilitates collaboration among individual members. In addition to using a public mailing list ([public-semweb-lifesci@w3.org](mailto:public-semweb-lifesci@w3.org)) to broadcast and exchange email messages, the HCLSIG conducts regular teleconference calls for members to participate. Wiki pages have been created [42] for describing the various activities in progress within HCLSIG, sharing data and documents produced by individual projects and writing documentation in a collaborative fashion. Face-to-face meetings took place in the United States and The Netherlands to engage the HCLSIG members in closer and more personal interactions as well as working sessions. As a result of the activities from the face-to-face meeting in January 2006, five task forces were established. Each task force plans its work within the two year overall timeframe. The task forces independently, and sometimes collectively, work on different aspects of the overall challenge. These task forces and their goals are described below.

### BioRDF

Existing biomedical data is available in different (non-Semantic-Web) formats including structured flat files, HTML, XML and relational databases. Often these formats include elements or fields, which are natural language. BioRDF has the goal of converting a number of publicly available life sciences data sources into RDF and OWL. Heterogeneous data sources have been selected so that the group can explore the use of a variety data conversion tools, thereby gaining insight into the pros and cons of different approaches.

### Ontologies

A goal of the HCLSIG is to facilitate creation, evaluation and maintenance of core vocabularies and ontologies to support cross-community data integration and collaborative efforts. Although there has been substantial effort in recent years to tackle these problems, the methodology, tools, and strategies are not widely known to biomedical researchers. The role of the ontologies task force is to work on well-defined use cases, supporting the other HCLSIG working groups. Where possible, the group works to identify ontologies that formalize and make explicit the key concepts and relationships that are central to those use cases. In cases where ontologies do not currently exist, the group works on prototyping and encouraging further development of the necessary terminology.

### Drug safety and efficacy

The development of safe and efficacious drugs rests on the proper and timely utilization of diverse information sets

and the adoption of and compliance to well-defined policies. The group works on the evaluation of Semantic Web technologies in a number of areas, focusing on the use of ontologies to aid queries against the different information sets, and rules for specification of policies. Topics include:

- Identifying and addressing challenges working with biomarkers and pharmacogenomics in coordination with U.S. Food and Drug Administration (FDA) and European Medicine Agency (EMA) guidelines.
- Detecting, examining, and classifying signals of potential drug side-effects or adverse reactions [43,44].
- Issues in clinical trial planning, management, analysis, and reporting – e.g., data security and integrity.
- Facilitating electronic submissions as per the Common Technical Document [45] specifications.

#### *Adaptable clinical pathways and protocols (ACPP)*

Evidence based clinical guidelines and protocols are recommendations for diagnostic and therapeutic tasks in a health care setting. They are increasingly perceived as an important vehicle for moving results of research and clinical trials to application in patient care. Much effort has been devoted to representing clinical guidelines and protocols in a machine-executable format [46]. This has proven to be quite a challenge. Translating the text-based guidelines to a machine-executable format is costly and thus far, solutions have required proprietary guideline execution engines, limiting widespread adoption. The slow pace of updating such guidelines limit their use in medical practices that want to quickly incorporate new clinical knowledge as it is published.

The ACPP task force explores the use of Semantic Web technologies, including RDF, OWL, logic programming, and rules to represent clinical guidelines and guide their local adaptation and execution. Guidelines encoded using these technologies can be accessed, reasoned about, and acted upon by a clinical information system. Since guidelines are Web documents, they have the potential to be more rapidly updated.

The following aspects of guideline and protocol representation and reasoning are of special interest:

- Inclusion and exclusion criteria that are used to decide whether evidence suggests the use of a particular guideline or protocol.
- Representation of temporal concepts and inference rules necessary for tracking processes and ensuring temporal constraints on treatment.

- Representation of medical intentions, goals, and outcomes.
- Use of logic programming to implement guidelines adaptable to site of care execution constraints and changes in patient condition.

#### *Scientific publishing*

Today, a large portion of biomedical knowledge production is in the form of scientific publications. Most often, on the Web, these publications are referred to either by name or by using hyperlinks. Neither the relationship of the publication to the context from which it is cited, nor the entities and relationships described by it, are explicitly represented. The scientific publishing task force is involved in several activities aimed at ameliorating this situation, attentive to the importance of social process and community engagement.

- Developing an application enabling researchers to collect publications, annotate, and interrelate the hypotheses and claims they present, and share their collections.
- Applying natural language processing techniques to scientific text to recognize and encode entities and relationships among them.
- Creating prototypes of tools and processes to enable researchers to include such information as a standard part of the scientific publication process.

#### *Neuromedicine and the semantic web*

From the outset, HCLSIG participants felt strongly that useful application of Semantic Web to biomedicine would only occur if the technology was applied to and rooted in realistic use cases, and if the various task forces were encouraged to have their work interoperate within a common domain. Although medical research and practice generally depend on data sets covering genetics to clinical outcomes, research in and therapy development for the neurodegenerative disorders is a particularly striking illustration of the need for active, ongoing, synthesis of information, data, and interpretation from many sources and subdisciplines in biomedicine. For this reason, the HCLSIG is currently exploring use cases involving neurodegenerative diseases such as PD and AD. Next, we illustrate some of the issues with a scenario of a clinical researcher attempting to develop immunotherapies for AD.

#### *Alzheimer's disease immunotherapy scenario*

A scientist working in a research hospital is pursuing immunization therapy for AD. A clinical trial of a vaccine made of synthetic Abeta1-42 ended prematurely a few years ago because 15 volunteers developed cerebral

inflammation [47]. However, the field remains enthusiastic about new immunization strategies to reduce Abeta in early Alzheimer's, believed to be the culprit of AD [48], and to study the mechanism of action of Abeta immunization [49]. Important steps would be to identify the specific form of Abeta that is toxic to neurons and/or other elements critical to proper CNS function, and the mechanism of its toxicity.

The scientist uses her local scientific knowledge management system (*sci-know*) to search the Alzheimer Research Forum Web site and finds a recently published hypothesis (*Abeta\*56 Hypothesis*) claiming a newly identified assembly of amyloid beta peptide, Abeta\*56, causes memory impairment [4]. However, the hypothesis is based on claims only supported by experimental results from a transgenic mouse model. She wonders if Abeta\*56 is found in actual AD patients, particularly in the early stages.

Based on the terms tagged to the hypothesis, that along with the original citation have been added to *sci-know*, the investigator constructs a search adding the concept *human* to the original query. The new query is run against PubMed and the hypothesis repository. Drawing on the ontology in the vicinity of the search terms to cluster the results, one research article comes to the forefront:

- i. Using a novel, attomolar detection system, Amyloid-beta Derived Diffusable Ligands (ADDL) are elevated 8-fold on average (max 70-fold) in the cerebrospinal fluid of patients with AD [50].

The Alzforum AD Hypothesis knowledgebase shows (i) is cited as supportive evidence for the *ADDL Hypothesis* claiming *ADDL causes memory impairment*. Though the Abeta\*56 hypothesis does not yet include a proposed mechanism for memory loss in the mouse model, the *ADDL hypothesis* includes a finding that ADDLs bind to human-derived cortical synaptic vesicles [51], and they inhibit hippocampal long-term potentiation (LTP) [52], a form of synaptic plasticity known to be critical for certain forms of learning and believed to be equally critical for memory storage [53,54]. Additional supporting evidence cited for this hypothesis notes Abeta alters A-type K<sup>+</sup> channels involved in learning and memory, leading to altered neuronal firing properties as a prelude to cell death in *Drosophila* cholinergic neurons [55]. This provides a possible mechanistic explanation for the demonstrated learning disabilities, memory dysfunction, and neurodegeneration in transgenic *Drosophila* expressing human Abeta [56].

Are these model organism findings relevant to patients with AD? The researcher wonders whether A-type K<sup>+</sup>

channels are plausible therapeutic targets for treating patients diagnosed with AD. She asks:

"Show me the neuron types affected by early AD."

The *sci-know* system searches the Alzforum and comes up with several instances of neuronal cell types damaged in AD. These include BDNF neurons of the nucleus basalis of Meynert [57,58] and CA1 pyramidal neurons of the hippocampus [59]. Next, the researcher asks:

"Do BDNF neurons or CA1 pyramidal neurons have A-type K<sup>+</sup> channels?"

"Are there other studies relating amyloid derived peptides to neocortical K<sup>+</sup> channels?"

The application returns results from a neuropharmacological knowledgebase, BrainPharm. [60]. BrainPharm indicates CA1 pyramidal cells have A-type potassium channels. Interestingly, this finding carries the following annotation:

"Application of beta-amyloid [Abeta] to outside-out patches reduces the A-current; leading to increased dendritic calcium influx and loss of calcium homeostasis, potentially causing synaptic failure and initiating neuronal degenerative processes." [61].

Our researcher wonders whether the 56 kD form of Abeta is responsible for this effect and is led to a series of scientific questions she would like to address in her lab. Would the Tg2576 mouse model, the one in which Abeta\*56 was reported to correlate with memory impairment, have a reduced A-current? Would blocking Abeta\*56 with an antibody restore the A-current level? Our researcher types in one more query:

"Is there an antibody to Abeta\*56 or ADDL?"

The application searches across a number of antibody resources and identifies one in another researcher's shared antibody database that even lists the e-mail address of the laboratory where she can obtain the antibody.

#### **Making data available in RDF and OWL**

In our scenario, a number of queries are posed for a variety of types of biomedical knowledge. We query for specific types of neurons, the types of their associated ion channels, for the properties of amyloid derived peptides and their molecular interactions, for hypotheses and discussions about them, and for antibody reagents. Much, but not all, of this information is available in publicly accessible data sets. However, in order for them to be used on the Semantic Web, they need to be made accessible as

RDF or OWL. The BioRDF group is exploring a number of methods for doing this. Among the data sets we have converted, and plan to make publicly available, are:

- **SenseLab.** The subset of *SenseLab* [62] that contains information about pathological mechanisms related to Alzheimer's Disease (*BrainPharm*) has been converted into RDF and the subset containing information about neuronal properties (*NeuronDB*) has been converted into OWL.
- **CoCoDat.** CoCoDat [63] is repository of quantitative experimental data on single neurons and neuronal micro-circuitry. A subset of information about ionic currents in different types of neurons has been converted into OWL.
- **Entrez Gene.** As described in [64], the Entrez Gene repository of gene-centered information was converted in its entirety to RDF.
- **PDSP Ki DB.** The *PDSP Ki Database* [65] is a repository of experimental results about receptor-ligand interactions and has a strong emphasis on neuroreceptors. It has been converted into OWL that conforms to an extended version of the established *BioPAX* [66] ontology for biomedical pathways.
- **BIND.** The *Biomolecular Interaction Network Database (BIND)* [67] is a large collection of molecular interactions, primarily protein-protein interactions. Like the PDSP KiDB, the OWL version of *BIND* is based on the *BioPAX* ontology.
- **Antibodies** – A collection of commercial antibody reagent data derived from the Alzforum Antibody Directory [68] and by crawling reagent vendor sites has been rendered in OWL.

In addition to the RDF and OWL data sets produced by the HCLSIG participants, there is a growing collection of RDF and OWL data sets that have been made available. Among these data sets are the OBO ontologies [69], Reactome [70], KEGG [71], NCI Metathesaurus [72], and UniProt [73].

Below we briefly discuss three approaches we have used to make data sets available in RDF.

#### CoCoDat

D2RQ [74] is used to provide access to CoCoDat. D2RQ is a declarative language to describe mappings between relational database schemas and either OWL or RDFS ontologies. The mappings allow RDF applications to access the contents of relational databases using Semantic Web query languages like SPARQL. Doing such a map-

ping requires us to choose how tables, columns, and values in the database map to URIs for classes, properties, instances, and data values. We illustrate some of these considerations by walking through a portion of the D2RQ document describing the mapping of CoCoDat's relational database form to RDF. In it, we see how rows in the *Neurons* table are mapped to instances, the column *ID\_BrainRegion* is mapped to a property, and the string values of that column are mapped to URIs.

**@prefix d2rq:** <http://www.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/0.1#>.

**@prefix:** <http://semweb.med.yale.edu/NeuroWeb/owl/cocodat#>.

**@prefix db1:** <http://semweb.med.yale.edu/NeuroWeb/db/cocodat#>.

The first task is to define the *namespace bindings* [75]. A namespace binding associates an abbreviation with a prefix used for a set of URIs. Following Semantic Web practice, all identifiers used in the mapping description are URIs. The mapping needs to use identifiers defined by D2RQ, identifiers we will generate for the RDF version of CoCoDat, and identifiers for parts of the relational database.

- "d2rq:" is the abbreviation for the namespace of identifiers used by D2RQ.
- "db1:" is the abbreviation for the namespace of identifiers of parts of the relational database.
- As identifiers should be globally unique, and the group undertaking the translation controls the domain 'semweb.med.yale.edu', the namespace for new identifiers in the RDF version of CoCoDat is based on that domain. This is chosen to be the *default namespace*, abbreviated as ":".

**db1:CoCoDatDB rdf:type d2rq:Database; d2rq:odb-cDSN "cocodat";**

Now the relational database where CoCoDat is stored is identified as "db1:CoCoDatDB" and defined by its connection via ODBC.

**db1:RecordingNeuronSite rdf:type d2rq:ClassMap;**

**d2rq:class :RecordingNeuronSite;**

**d2rq:uriPattern "":RecordingNeuronSite-@@Neurons.ID@@";**

### d2rq:dataStorage db1:CoCoDatDB.

Following that, each row of the database table *Neurons* is mapped to an instance of the OWL class called *:Recording-NeuronSite*. The URI of each instance is constructed using the primary key of the table, *ID*. Therefore, the instance with the primary key 1 will have the URI "<http://semweb.med.yale.edu/NeuroWeb/owl/cocodat#RecordingNeuronSite-1>", abbreviated *:RecordingNeuronSite-1*.

db1:inBrainRegion rdf:type d2rq:ObjectProperty-Bridge;

d2rq:belongsToClassMap db1:RecordingNeuronSite;

d2rq:property :inBrainRegion;

d2rq:pattern "@@Neurons.ID\_BrainRegion@@";

d2rq:translateWith db1:BrainRegionTable.

In this step, the *ID\_BrainRegion* column in the Neuron table is mapped to the property *:inBrainRegion*. The values of that column are not to be used directly, instead undergoing a translation that is defined next.

db1:BrainRegionTable rdf:type d2rq:TranslationTable;

d2rq:translation [d2rq:databaseValue "GM-Ctx\_B";  
d2rq:rdfValue :barrel-cortex;];

d2rq:translation [d2rq:databaseValue "GM-Ctx\_Gen";  
d2rq:rdfValue :general-cortex;];

d2rq:translation [d2rq:databaseValue "GM-Ctx\_SeM";  
d2rq:rdfValue :sensorimotor-cortex;];

In this last step, we see a portion of the mapping of values from the *ID\_BrainRegion* column. The string values in this column are meant to represent brain regions. Knowing that it is likely these values will need to be equated with terms from other ontologies, a decision is made to represent them as URIs. Later, one will be able to use *owl:sameAs* to equate these terms with others. With this mapping, the string "GM-Ctx\_B" is translated into the URI "<http://semweb.med.yale.edu/NeuroWeb/owl/cocodat#barrel-cortex>".

The result of this mapping specification will be the creation of statements such as `<:RecordingNeuronSite-1><:inBrainRegion><:barrel-cortex>`, assuming the ID of the first row of the *Neurons* table is 1 and the value in the *ID\_BrainRegion* column is "GM-Ctx\_B".

### Entrez Gene

The XML version of Entrez Gene was transformed to RDF using XSLT [76]. The XML source is 50 GB and the generated RDF consists of 411 million triples. The Oracle Database 10g RDF Data Model was used to store and query the data. Although it would have been expedient to use XML element names directly as RDF properties, we instead mapped the element names to property names that were more descriptive and adhered better to accepted RDF style. For example, the element *Gene-track\_geneid* was changed to the property *has\_unique\_geneid*. An authoritative URI naming scheme for NCBI resources does not exist, so the namespace "[http://www.ncbi.nlm.nih.gov/dtd/NCBI\\_Entrezgene.dtd/](http://www.ncbi.nlm.nih.gov/dtd/NCBI_Entrezgene.dtd/)" was created for use in this prototype.

**Antibodies.** The curation of information about antibody reagents is much less mature than that about genes and many other biological entities. Therefore, creation of this resource had a number of interesting problems. The most difficult challenge was how to associate antibodies with proteins. The query in our scenario depends on this association, yet the Alzforum directory and most commercial reagent vendors do not associate antibody targets with well known identifiers. Instead, they are listed by gene, protein, or molecule name. Our focus was on antibodies that react with proteins. Determining the referent of antibody names can be difficult because of the large number of gene and protein synonyms. This is further complicated because names can have variant spellings, antibodies can be non-specific, vendors can use idiosyncratic names, and protein names are often embedded in a product name. Our approach was to collect gene and protein synonyms from a variety of public databases – Entrez Gene, UniProt, OMIM [77], and Enzyme [78]. Sets of transformation rules (based on regular expressions) were applied to product listings to extract protein names, normalize common spelling variations, and recognize certain forms of lists. Finally, only unambiguous matches to names were considered reliable enough to use.

Understanding the provenance and terms of usage of data is important within science. We therefore created RDF using the FOAF [79] vocabulary to describe the Alzforum project, and used Dublin Core [80] properties to identify usage policies for the data. This RDF was linked to the newly compiled Alzforum antibody listing.

### Curating and navigating disease hypotheses, claims, and evidence

In our scenario, an essential part of the navigation that leads the scientist from desired therapy to molecular mechanism is based on relationships between hypotheses. Although much of what we represent in biomedical databases are experimental measurements or observa-

tions, the act of creating and consuming knowledge occurs in a complex web of activities and relationships. From this perspective, one way to view biomedical knowledge is as an incomplete network whose "growing edges" contain unresolved contradictions, i.e. varying interpretations of experimental data in relation to hypotheses.

A natural science focused ontology of AD might contain the relationship  $\langle \text{NeurofibrillaryTangle} \rangle \langle \text{locatedIn} \rangle \langle \text{Neuron} \rangle$ , asserting a known fact. However, for active researchers in a field, many times the most interesting relationships are those that are just emerging, i.e. they cannot yet be considered validated, and are often the subject of scientific controversy. Perhaps more than anywhere, the collection of these hypotheses, claims, and disputes characterizes the world of science and provides the raw material propelling experiments, grants, and publications. How, then, can we assist scientists in taking advantage of this class of knowledge?

SWAN (Semantic Web Applications in Neuromedicine), developed in part by members of the HCLSIG, is an application that focuses on enabling AD researchers to curate, organize, annotate, and relate scientific hypotheses, claims and evidence about the disease. The ultimate goal of this project is to create tools and resources to manage the evolving universe of data and information about AD, in such a way that researchers can easily comprehend their larger context ("what hypothesis does this support or contradict?"), compare and contrast hypotheses ("where do these two hypotheses agree and disagree?"), identify unanswered questions, and synthesize concepts and data into more comprehensive and useful hypotheses and treatment targets for this disease.

The application is oriented towards use by both the individual researchers and within the community. Therefore the application supports both secure personal workspaces as well as shared, public workspaces.

The 2005 pilot application was developed as a proof of concept for hypothesis management [81]. In SWAN, personal and public knowledgebases are structured as RDF triple stores manipulated by the Jena framework [82]. Content can be exported and shared peer-to-peer or via public knowledge servers. Neuroscientists and scientific editors have used the system. Knowledge in the workspaces has been integrated with data from SenseLab and other data sets using the Oracle RDF Data Model [83,84]. Development continues and initial deployment will be as part of the Alzheimer Research Forum Web site [85].

#### **Working with clinical guidelines**

Much effort has been devoted to representing clinical guidelines and protocols in a machine-executable format

[46]. The high cost of creating these frameworks and the specialized software needed to use them has hindered wide adoption of such systems. One challenge is that the encoded guidelines are not generally interoperable between systems, diluting what could be a combined effort to build this valuable resource. We observe that much of the technology needed to represent and execute such guidelines is available as part of the Semantic Web stack. Thus, we are experimenting with using Semantic Web technologies to implement such guidelines in order to show their effectiveness and to give feedback to developers on where additional capabilities are needed. Working within the Semantic Web would benefit this field for at least two reasons. First, the open standards for the technologies on which such systems can be built would encourage researchers and vendors to build systems that can interoperate. Second, it would speed development of such systems by making it easier for them to incorporate essential and current biomedical knowledge created by others, saving the cost of encoding that knowledge in each system that uses it.

Adaptability to changing conditions is an important requirement for making clinical recommendations. These changes take the form of a patient's condition progressing in potentially unpredictable ways, and new medical research and clinical trials that should be considered in addition to established guidelines.

Within ACPP we have modeled guidelines as directed graphs using RDF and OWL [86]. Within such a network, each node is a task. Depending on the granularity desired by clinical practices using the guideline, the task might be a process or a set of processes. Each process is designed to accomplish a clinical goal, such as acquiring knowledge via a diagnostic test and is associated with its expected outcome and a desired timeframe for that outcome. OWL is used to represent the ontology of clinical goals and outcomes following [87].

Each task has a *context* describing a set of sufficient conditions that make the process worthy of recommendation and safe to carry out. The context describes a mix of the patient's clinical and physical conditions, treatment status, and care setting. For example, it can make reference to states of prior or parallel processes, such as whether they were completed or aborted, and clinical settings such as a long term care centre, or an emergency room. These conditions are organized into *inclusion* and *exclusion* criteria. Inclusion criteria may be weighted and a minimum sum of weights of satisfied criteria is specified as a threshold above which a task can be recommended.

As an example, consider the treatment of dementia in AD patients. Prescription of cholinesterase inhibitors such as

donepezil, rivastigmine, and galantamine are recommended based on evidence from clinical trials [88]. In our model, using OWL, *prescribingCholinesteraseInhibitors* is an instance of the *Process* class. An inclusion criterion would be a diagnosis of either AD Dementia, PD or Lewy Body Dementia (DLB). These diagnoses are represented as classes, and so the inclusion criterion can be represented as an OWL union of the classes. Exclusion criteria would be vomiting or other severe gastrointestinal disorders.

A clinical decision support system can recommend the next task in a patient-specific pathway based on rules. Although we have used OWL for evaluating rules using instance classification, the current standard is not expressive enough to use the weights and thresholds we assign to criteria in class definitions. To implement the following, we use Notation 3 [89] rules. All tasks are evaluated in the following way to see which are candidates for recommendation.

- Query the healthcare information network for all past and present patient conditions mentioned in the inclusion or exclusion criteria.
- If any exclusion criteria hold then discard the task.
- Collect the satisfied inclusion criteria.
- Add the weights assigned to each satisfied inclusion criteria.
- If the sum exceeds the threshold, the task may be recommended.

Regular re-evaluation during periods of patient stability and upon any change in medical condition allow us to adapt the treatment plan to the current medical situation.

This approach to representing guidelines is also well suited to the incorporation of new knowledge. Each guideline would be available as an individual RDF or OWL document uniquely identified by its URI. Trusted sources would be identified that maintain up-to-date guidelines and protocols. Analogous to the *contexts* of tasks, each guideline or clinical trial would be associated with its own inclusion and exclusion criteria that would qualify the whole body of knowledge, i.e. all tasks described in the guideline. With this approach, the same form of rules used to identify relevant tasks would be used to identify relevant guidelines [90]. The tasks from all relevant guidelines and protocols would then be evaluated to determine the set of recommendations. By applying this method, if a patient has multiple clinical conditions, all relevant guidelines can be utilized to ensure doctors

have appropriate information to ensure the best possible treatment for their patients.

## Discussion

### Data integration

There is a tacit assumption within the Semantic Web community that every data set and ontology will interoperate. The reality is that different conceptualizations and representations of the same data can exist. While the architecture and basic tools of the Semantic Web remove a set of previous roadblocks to data integration, positive progress towards it requires study, experimentation, and at-scale efforts that exercise proposed solutions.

To date, we have primarily focused on building prototypes that have functioned independently. Much of the RDF and OWL that has been generated mirrors the structure of the original data sets. Such translations are more syntactic than semantic. Even so, the common syntax enables an easier creation of cross-domain queries. As an example, in [83] the RDF translation of BrainPharm and SWAN's publication, data in RDF format were loaded into a single RDF store. Having both data sets available simultaneously allowed interesting new queries. For example, one could retrieve commentary by Alzforum members on articles that discussed drugs for which BrainPharm had models about cellular mechanism of action. This type of query succeeds because the two data sets being integrated do not, for the most part, discuss the same type of entity.

In order to integrate data sets, one of two things must happen: either terms for entities and relationships must be shared between the data sets (the data sets must be built using a shared ontology) or concordances must be available that relate terms in one data set to those in another.

Even when the ontology is shared, there is no guarantee that integration will be successful. Consider the BioPAX exchange format, an OWL-based ontology that provides a common framework for the many data sources that are repositories of information on cellular pathways. Despite the common ontology, it remains difficult to query an aggregation of different sources of BioPAX formatted data, e.g., for interactions related to the glucose metabolism pathway. This is because the terms shared among the data sources (the ones defined in the BioPAX standard) do not cover the scientific domain adequately to support such a query.

Building such ontologies is hard. The ontologies task force has therefore started focusing on identifying available knowledge resources (e.g., thesauri, terminologies, ontologies) that cover the basic biomedical entities and relations required to formally represent well defined scenarios like the one we present above.

While concepts in evolving areas of research may be incomplete, unclear, in transition or under dispute, there are many important entities and relations upon which most biomedical researchers and clinicians will agree. Mitochondria are found inside viable eukaryotic cells, voluntary movement in humans requires functional innervation of skeletal muscles, etc.

Our first goal is to construct a skeleton ontology specifying the required high-level biomedical domains, and, then to determine which public resources provide the required domain entities and relations along with clear prose definitions of them. These textual definitions are essential to guide curators and translators of data sets towards consistent usage of terms. Where definitions that we need do not exist in public resources, we will attempt to define the terms and work with others in the biomedical ontology community to refine and formalize them.

For example, an important term in our scenario is *Ion Channel*. In order to pose a query about ion channels and retrieve information about *A-type K<sup>+</sup> channels*, we need to ensure that the definition is clear enough that competent informaticians who are not necessarily domain experts have enough hints to gather sufficient information to realize that a *K<sup>+</sup> channel* is an ion channel.

It is important that the same attention that is given to identifying and defining classes is also given to defining relationships (properties) [91]. There are fewer definitions for such relationships, in public resources, than for classes. For example, in order to record details of the hypotheses in our scenario, we need to define the relationship between Abeta and development of symptoms of AD. Therefore we might define "isAPeptideContributing-CauseOf" to be "a potentially causal relationship between peptides such as Abeta1-42, Abeta\*56 and a disease such as AD or a clinical condition such as Memory Impairment". The definition notes the type of subject (peptide) and object (clinical condition or disease) of the property that will formally link, as domain and range of the property, and then to classes in our ontology. This definition will serve as our input to other communities working in this domain – for example when we participate in an upcoming workshop on clinical trial ontologies organized by the National Center for Biomedical Ontology (NCBO) [92].

#### **Current technical limitations of semantic web**

Semantic Web technologies are young. Gaps in standards and implementations still exist and adoption is limited by typical problems with early technology, such as the need for a critical mass of practitioners and installed base, and growing pains as the technology is scaled up. Some issues that have affected the work of the HCLSIG are:

#### **Scarcity of semantically annotated information sources**

Although we have listed a number of public sources of data that are available in RDF, most common sources of data for bioinformatics are not currently in a RDF or OWL. However, mapping tools such as D2RQ should lower the barrier to making these data sets available.

#### **Performance and scalability**

RDF and OWL stores are slower than optimized relational databases, but are improving steadily [93]. However, logical reasoning over large or complex ontologies remains a problem.

#### **Representation of evidence and data provenance**

It is often important to know where knowledge has come from and how it has been processed. It is also useful to know who believes something and why. However, there is no standard way of expressing such information about a statement or collection of RDF statements. *Named graphs* [94] may solve many of these problems and are already being employed in projects such as myGrid [95] to trace data provenance. However, they are not a standard and, therefore, are not widely supported by Semantic Web tools.

#### **Lack of a standard rule language**

Although there are technologies that enable the use of rules, there is no standard rule language. This makes it impossible to write sets of rules that can be used in different implementations, limiting the reach of the ACP group's vision of distributed clinical guidelines encoded as rules. We note, however, that the W3C Rule Interchange Format Working Group [96] is currently working to solve this problem.

#### **Cross-community interactions**

There is an emerging consensus in the bioinformatics community at large for the need to formalize and share data annotation semantics. This is championed by such institutions as the UK e-Science project myGrid [97], the Bio-Health Informatics Group [98] at the University of Manchester, U.K., the NIH-funded National Center for Biomedical Ontology [42,99], and the growing Open Biomedical Ontologies (OBO) Foundry [100].

The Semantic Web and biomedical communities need to further coordinate efforts in areas critical to translational research, namely:

- Formalizing the semantics of the elements of health care information systems, such as medical records, as well as clinical decision making, such as disease and symptoms.
- Making scientific publishing more effective at supporting research communities by finding ways to systemati-

cally capture research results and make them available on the Semantic Web.

- Engaging systems biology researchers as "early adopters" of Semantic Web technologies, and as a resource for driving use cases.
- Working with natural language processing researchers to enhance their algorithms with biomedical ontologies, and to target their output to use terms from established ontologies.
- Working with the U.S. National Library of Medicine (NLM) to find appropriate ways to translate their extensive vocabularies and knowledge resources into RDF for effective use on the Semantic Web.

As discussed in [101], tensions have occurred between the Semantic Web communities and other communities like the XML and database communities, as some people believe that the technologies being advocated by these communities cannot coexist with each other. One way to ease such tensions is for the Semantic Web community to develop a complementary rather than competitive relationship with these communities. The Semantic Web should be perceived as a complement instead of a replacement to existing technologies. For example, RDF/OWL can be serialized as XML, and can be used to provide a richer semantic layer for use with other XML technologies. The developers of triple stores and RDF query languages have been greatly inspired by the theoretical and practical work done by the database community. Providers of valuable knowledge such as curators of biological pathways would be more willing to make their data accessible to the Semantic Web community if they did not need to abandon their own formats. For example, converters can be provided for translating BioPAX into other pathway data formats so that tools that were built based on these formats can still be used. At the same time, additional tools can be developed to exploit the new features (e.g., reasoning) enabled by representing BioPAX in OWL.

#### **Education and incentives**

The vision of a Semantic Web accelerating biomedical research crucially depends on the holder of scientific and clinical data making that data available in a reusable form. Often the effort that goes into preparing and serving this data will not directly benefit the provider. Instead, researchers are measured for producing scientific discoveries and writing about them, doctors for helping sick patients, and pharmaceutical companies for producing safe, effective drugs. There are also privacy risks involved with sharing personal information. Valuable patient data can only be acquired with appropriate consent and with sensitivity to those privacy issues. It is an open question of

how to structure incentives to make these holders of valuable information consider the effort to be in their best interest.

If the research community decided today that it was motivated to publish data semantically, we do not yet have adequate numbers of skilled knowledge workers. Data modelling even without the intention of interoperating is a hard-learned skill, and the challenge is substantially magnified when the intention is to share information for unforeseen uses. We need to establish and populate a new discipline, a mix of interdisciplinary skills that include solid understanding of biomedicine, computer science, philosophy and the social anthropology of science and computing.

#### **Conclusion**

We have discussed the potential of the Semantic Web to facilitate translational research. Although Semantic Web technologies are still evolving, there are already existing standards, technologies, and tools that can be practically applied to a wide range of biomedical use cases. There are challenges to the widespread adoption of the Semantic Web in the health care and life sciences industries. Some parts of the technology are still in development and are untested at large scales. Informaticians need training and support to be able to understand and work with these new technologies. Incentives need to be provided to encourage appropriate representation of important research results on the Web.

By grounding the development and application of this technology in real concerns and use cases of the biomedical community, and enabling close interaction between informaticians, researchers, and clinicians, and the W3C standards development community, the W3C HCLSIG is providing a rich collaborative environment within which to start resolving these issues. The potential of interoperable knowledge sources for biomedicine, at the scale of the World Wide Web, certainly merits continued attention.

#### **Competing interests**

The authors declare that they have no competing interests.

#### **Authors' contributions**

KC initiated and orchestrated the effort of writing the paper. All authors have contributed to the manuscript and participated in the discussions at the face-to-face meetings, teleconferences and on e-mail. IH, EN, and TH helped facilitate forums for discussing the paper. JK, TC, EW, GW, and WB developed the AD immunotherapy scenario. AR edited the manuscript, with help from TC, WB, KC, JR, SS, and SM.

## Acknowledgements

KC was partly supported by NSF grant DBI-0135442 and NIH grant P01 DC04732. JL was supported by NSF grant IIS-0542041. BB receives support from NIH grants P20 MH62009 (MBL) and RR043050-S2 (Mouse BIRN). The SWAN project is partly supported by a grant from the Ellison Medical Foundation. A significant portion of this work was performed within the framework of the Health Care and Life Sciences Interest Group of the World Wide Web Consortium. The authors appreciate the forum and the resources given by this Interest Group. Thanks to SM and IH for hosting the HCLSIG Amsterdam face-to-face meeting discussions during which seeds of the paper were planted. TC and JK are principal investigators for the SWAN project. EN and TH are the co-chairs of the HCLSIG and IH is its liaison to the W3C. SS, VK and HC coordinate the task forces. The authors would also like to acknowledge Bo. H. Andersson, Dirk Colaert, Jeorg Hakenberg, and Ray Hookway, who were participants of the Amsterdam face-to-face meeting, for their participation in the discussion about the paper. We would like to thank Alzheimer Research Forum and Brainstage Research, Inc for contributing to part of the publication costs.

This article has been published as part of *BMC Bioinformatics* Volume 8 Supplement 3, 2007: Semantic e-Science in Biomedicine. The full contents of the supplement are available online at <http://www.biomedcentral.com/1471-2105/8?issue=S3>.

## References

- Zerhouni E: **Medicine. The NIH Roadmap.** *Science* 2003, **302**:63-72.
- Davenport T, Prusak L: *Information Ecology: Mastering the Information and Knowledge Environment* 1st edition. Oxford University Press; 1997.
- Current Hypotheses** [<http://www.alzforum.org/res/adh/cur/default.asp>]
- Lesne S, Koh MT, Kotilinek L, Kaye R, Glabe CG, Yang A, Gallagher M, Ashe KH: **A specific amyloid-beta protein assembly in the brain impairs memory.** *Nature* 2006, **440**:352-357.
- Catalano SM, Dodson EC, Henze DA, Joyce JG, Krafft GA, Kinney GG: **The role of amyloid-beta derived diffusible ligands (ADDLs) in Alzheimer's disease.** *Curr Top Med Chem* 2006, **6**:597-608.
- Selkoe DJ: **Alzheimer's disease: genes, proteins, and therapy.** *Physiol Rev* 2001, **81**:741-766.
- Marchesi VT: **An alternative interpretation of the amyloid Abeta hypothesis with regard to the pathogenesis of Alzheimer's disease.** *Proc Natl Acad Sci USA* 2005, **102**:9093-9098.
- Gusella JF, MacDonald ME, Ambrose CM, Duyao MP: **Molecular genetics of Huntington's disease.** *Arch Neurol* 1993, **50**:1157-1163.
- Castro-Garcia A, Sesar-Ignacio A, Ares-Pensado B, Relova-Quinteiro JL, Gelabert-Gonzalez M, Rumbo RM, Noya-Garcia M: **Psychiatric and cognitive complications arising from subthalamic stimulation in Parkinson's disease.** *Rev Neurol* 2006, **43**:218-222.
- Hely MA, Morris JG, Reid WG, Trafficante R: **Sydney Multicenter Study of Parkinson's disease: non-L-dopa-responsive problems dominate at 15 years.** *Mov Disord* 2005, **20**:190-199.
- Castro A, Valdeoriola F, Linazasoro G, Rodriguez-Oroz MC, Stochi F, Marin C, Rodriguez M, Vaamonde J, Jenner P, Alvarez L, et al.: **[Optimization of use of levodopa in Parkinson's disease: role of levodopa-carbidopa-entacapone combination].** *Neurologia* 2005, **20**:180-188.
- Lindvall O, Bjorklund A: **Cell therapy in Parkinson's disease.** *NeuroRx* 2004, **1**:382-393.
- Royall DR, Lauterbach EC, Cummings JL, Reeve A, Rummans TA, Kaufer DI, LaFrance WC Jr, Coffey CE: **Executive control function: a review of its promise and challenges for clinical research. A report from the Committee on Research of the American Neuropsychiatric Association.** *J Neuropsychiatry Clin Neurosci* 2002, **14**:377-405.
- Planells-Cases R, Lerma J, Ferrer-Montiel A: **Pharmacological intervention at ionotropic glutamate receptor complexes.** *Curr Pharm Des* 2006, **12**:3583-3596.
- Levy YS, Gilgun-Sherki Y, Melamed E, Offen D: **Therapeutic potential of neurotrophic factors in neurodegenerative diseases.** *BioDrugs* 2005, **19**:97-127.
- Hawkes C: **Olfaction in neurodegenerative disorder.** *Adv Otorhinolaryngol* 2006, **63**:133-151.
- Zadikoff C, Lang AE: **Apraxia in movement disorders.** *Brain* 2005, **128**:1480-1497.
- Sauer SW, Okun JG, Schwab MA, Crnic LR, Hoffmann GF, Goodman SI, Koeller DM, Koller S: **Bioenergetics in glutaryl-coenzyme A dehydrogenase deficiency: a role for glutaryl-coenzyme A.** *J Biol Chem* 2005, **280**:21830-21836.
- Bossy-Wetzel E, Schwarzenbacher R, Lipton SA: **Molecular pathways to neurodegeneration.** *Nat Med* 2004, **10**(Suppl):S2-9.
- Bursch W, Ellinger A: **Autophagy – a basic mechanism and a potential role for neurodegeneration.** *Folia Neuropathol* 2005, **43**:297-310.
- Bertram L, Tanzi RE: **The genetic epidemiology of neurodegenerative disease.** *J Clin Invest* 2005, **115**:1449-1457.
- Miklossy J, Arai T, Guo JP, Klegeris A, Yu S, McGeer EG, McGeer PL: **LRRK2 expression in normal and pathologic human brain and in human cell lines.** *J Neuropathol Exp Neurol* 2006, **65**:953-963.
- Antoniou G, Van Harmelen F: *A Semantic Web Primer Cambridge, MA, USA: The MIT Press; 2004.*
- Berners-Lee T, Hendler J, Lassila O: **The Semantic Web.** *Scientific American* 2001, **May**.
- TP53 Human** [[http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene&cmd=Retrieve&dopt=full\\_report&list\\_uids=7157](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene&cmd=Retrieve&dopt=full_report&list_uids=7157)]
- RFC 3986 Uniform Resource Identifier (URI): Generic Syntax** 2005 [<http://www.rfc-editor.org/rfc/rfc3986.txt>].
- RDF Primer** 2004 [<http://www.w3.org/TR/rdf-primer/>]. W3C
- SPARQL Query Language for RDF** 2006 [<http://www.w3.org/TR/rdf-sparql-query/>]. W3C
- RDF Vocabulary Description Language – RDF Schema 1.0, 2004** 2004 [<http://www.w3.org/TR/rdf-schema/>]. W3C
- SKOS Core Guide** 2005 [<http://www.w3.org/TR/swbp-skos-core-guide/>]. W3C
- OWL Web Ontology Language, 2004** 2004 [<http://www.w3.org/TR/owl-guide/>]. W3C
- Gleaning Resource Descriptions from Dialects of Languages (GRDDL), 2006** 2006 [<http://www.w3.org/TR/grddl/>]. W3C
- Erling O, Mikhailov I: **Mapping Relational Data to RDF in Virtuoso.** 2006 [<http://virtuoso.openlinksw.com/wiki/main/Main/VOSSQLRDE>].
- Semweb Central Developer Site** [<http://www.semwebcentral.org/>]
- Semantic Web Tools, 2006** 2006 [<http://esw.w3.org/topic/SemanticWebTools>]. W3C
- Rector AL: **Modularisation of domain ontologies implemented in description logics and related formalisms including OWL.** *Proceedings of the international conference on Knowledge capture* 2003:121-128.
- Soldatova LN, King RD: **An ontology of scientific experiments.** *Journal of the Royal Society, Interface/the Royal Society* 2006, **3**:795-803.
- Whetzel PL, Parkinson H, Causton HC, Fan L, Fostel J, Fragoso G, Game L, Heiskanen M, Morrison N, Rocca-Serra P, et al.: **The MGED Ontology: a resource for semantics-based description of microarray experiments.** *Bioinformatics (Oxford, England)* 2006, **22**:866-873.
- Zucker J, Rutenberg A: **Debugging the Bug.** 2006 [[http://bio.free-logy.org/wiki/Debugging\\_the\\_bug](http://bio.free-logy.org/wiki/Debugging_the_bug)].
- Park YR, Park CH, Kim JH: **GOChase: correcting errors from Gene Ontology-based annotations for gene products.** *Bioinformatics (Oxford, England)* 2005, **21**:829-831.
- Semantic Web Health Care and Life Sciences Interest Group** [<http://www.w3.org/2001/sw/hcls/>]
- HCLSIG Wiki** [<http://esw.w3.org/topic/SemanticWebForLifeSciences>]
- Stephens S, Morales A, Quinlan M: **Applying Semantic Web Technologies to Drug Safety Determination.** *Intelligent Systems, IEEE [see also IEEE Intelligent Systems and Their Applications]* 2006, **21**:82-86.
- Neumann EK, Quan D: **Biodash: a Semantic Web dashboard for drug development.** *Pac Symp Biocomput* 2006:176-187.

45. **International Conference on Harmonisation; guidance on electronic common technical document specification; availability. Notice.** *Federal register* 2003, **68**:16060-16061.
46. Peleg M, Tu S, Bury J, Ciccarese P, Fox J, Greenes RA, Hall R, Johnson P, Jones N, Kumar A, et al.: **Comparing models of decision and action for guideline-based decision support: a case-study approach: Stanford University.** 2002. [Part 1 – <http://smi.stanford.edu/smi-web/research/details.jsp?PubId=922>; Part 2 – <http://smi.stanford.edu/smi-web/research/details.jsp?PubId=923>]
47. Gilman S, Koller M, Black RS, Jenkins L, Griffith SG, Fox NC, Eisner L, Kirby L, Rovira MB, Forette F, Orgogozo JM: **Clinical effects of Abeta immunization (AN1792) in patients with AD in an interrupted trial.** *Neurology* 2005, **64**:1553-1562.
48. Vasilevko V, Cribbs DH: **Novel approaches for immunotherapeutic intervention in Alzheimer's disease.** *Neurochemistry international* 2006, **49**:113-126.
49. Levites Y, Smithson LA, Price RW, Dakin RS, Yuan B, Sierks MR, Kim J, McGowan E, Reed DK, Rosenberry TL: **Insights into the mechanisms of action of anti-A {beta} antibodies in Alzheimer's disease mouse models.** *The FASEB Journal* 2006.
50. Georganopoulou DG, Chang L, Nam JM, Thaxton CS, Mufson EJ, Klein WL, Mirkin CA: **Nanoparticle-based detection in cerebral spinal fluid of a soluble pathogenic biomarker for Alzheimer's disease.** *Proc Natl Acad Sci USA* 2005, **102**:2273-2276.
51. Deshpande A, Mina E, Glabe C, Busciglio J: **Different conformations of amyloid beta induce neurotoxicity by distinct mechanisms in human cortical neurons.** *J Neurosci* 2006, **26**:6011-6018.
52. Walsh DM, Klyubin I, Fadeeva JV, Cullen WK, Anwyl R, Wolfe MS, Rowan MJ, Selkoe DJ: **Naturally secreted oligomers of amyloid beta protein potently inhibit hippocampal long-term potentiation in vivo.** *Nature* 2002, **416**:535-539.
53. Morris RG: **Long-term potentiation and memory.** *Philos Trans R Soc Lond B Biol Sci* 2003, **358**:643-647.
54. Lynch MA: **Long-term potentiation and memory.** *Physiol Rev* 2004, **84**:87-136.
55. Kidd JF, Brown LA, Sattelle DB: **Effects of amyloid peptides on A-type K<sup>+</sup> currents of Drosophila larval cholinergic neurons.** *J Neurobiol* 2006, **66**:476-487.
56. Iijima K, Liu HP, Chiang AS, Hearn SA, Konsolaki M, Zhong Y: **Dissecting the pathological effects of human Abeta40 and Abeta42 in Drosophila: a potential model for Alzheimer's disease.** *Proc Natl Acad Sci USA* 2004, **101**:6623-6628.
57. Siegel GJ, Chauhan NB: **Neurotrophic factors in Alzheimer's and Parkinson's disease brain.** *Brain Res Brain Res Rev* 2000, **33**:199-227.
58. Mufson EJ, Ginsberg SD, Ikonovic MD, DeKosky ST: **Human cholinergic basal forebrain: chemoanatomy and neurologic dysfunction.** *J Chem Neuroanat* 2003, **26**:233-242.
59. Selkoe DJ: **Biochemistry of altered brain proteins in Alzheimer's disease.** *Annu Rev Neurosci* 1989, **12**:463-490.
60. Marenco L, Tosches N, Crasto C, Shepherd G, Miller PL, Nadkarni PM: **Achieving evolvable Web-database bioscience applications using the EAV/CR framework: recent advances.** *J Am Med Inform Assoc* 2003, **10**:444-453.
61. Chen C: **beta-Amyloid increases dendritic Ca<sup>2+</sup> influx by inhibiting the A-type K<sup>+</sup> current in hippocampal CA1 pyramidal neurons.** *Biochem Biophys Res Commun* 2005, **338**:1913-1919.
62. Skoufos E, Mirsky JS, Healy MS, Singer MS, Hines ML, Nadkarni PM, Miller PL, Shepherd GM: **Acquisition, storing and retrieving diverse biomedical data using the World-Wide-Web: The Senselab Paradigm.** *AMIA'98 Annual Symposium* 1998.
63. Dyhrfeld-Johnsen J, Maier J, Schubert D, Staiger J, Luhmann HJ, Stephan KE, Kotter R: **CoCoDat: a database system for organizing and selecting quantitative data on single neurons and neuronal microcircuitry.** *Journal of neuroscience methods* 2005, **141**:291-308.
64. Sahoo SS: **Converting biological information to the W3C Resource Description Framework (RDF): Experience with Entrez Gene.** 2006 [<http://lsdis.cs.uga.edu/~satya/BioRDF/Report-Satya-S-Sahoo.pdf>]. *Lister Hill National Center for Biomedical Communications (NLM/NIH)*
65. Roth B, Kroeze W, Patel S, Lopez E: **The Multiplicity of Serotonin Receptors: Uselessly diverse molecules or an embarrassment of riches?** *The Neuroscientist* 2000, **6**:252-262.
66. **BioPAX** [<http://biopaxwiki.org>]
67. Bader GD, Betel D, Hogue CW: **BIND: the Biomolecular Interaction Network Database.** *Nucleic acids research* 2003, **31**:248-250.
68. **Alzforum Antibody Directory** [<http://www.alzforum.org/res/com/ant/default.asp>]
69. Bada M, Hunter L: **Enrichment of OBO Ontologies.** *J Biomed Inform* 2006.
70. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, et al.: **Reactome: a knowledgebase of biological pathways.** *Nucleic acids research* 2005, **33**:D428-432.
71. Kanehisa M: **The KEGG database.** *Novartis Foundation symposium* 2002, **247**:91-101. discussion 101-103, 119-128, 244-152
72. Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu WL, Wright LW: **NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information.** *Journal of biomedical informatics* 2007, **40**:30-43.
73. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, et al.: **The Universal Protein Resource (UniProt): an expanding universe of protein information.** *Nucleic acids research* 2006, **34**:D187-191.
74. **D2RQ** [<http://sites.wiwiw.fu-berlin.de/suhl/bizer/D2RQ/>]
75. **Using Qualified Names (QNames) as Identifiers in XML Content** 2004 [<http://www.w3.org/2001/tag/doc/qnameids.html>]. W3C
76. **XSL Transformations (XSLT)** 1999 [<http://www.w3.org/TR/xslt>]. W3C
77. **Online Mendelian Inheritance in Man, OMIM (TM)** 2006 [<http://www.ncbi.nlm.nih.gov/omim/>]. *McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD)*
78. Bairoch A: **The ENZYME database in 2000.** *Nucleic acids research* 2000, **28**:304-305.
79. Brickley D, Miller L: **Friend of a Friend (FOAF).** 2005 [<http://xmllns.com/foaf/0.1/>].
80. Beckett D, Miller E, Brickley D: **Expressing Simple Dublin Core in RDF/XML.** *Institute for Learning and Research Technology (ILRT) University of Bristol*; 2002.
81. Gao Y, Kinoshita J, Wu E, Miller E, Lee R, Seaborne A, Cayzer S, Clark T: **SWAN: A Distributed Knowledge Infrastructure for Alzheimer Disease Research.** *Journal of Web Semantics* 2006, **4**:8.
82. Carroll JJ, Dickinson I, Dollin C, Reynolds D, Seaborne A, Wilkinson K: **Jena: Implementing the Semantic Web Recommendations.** *Bristol, England, UK: Digital Media Systems Laboratory HP Laboratories*; 2003.
83. Lam Y, Marenco L, Clark T, Gao Y, Kinoshita J, Shepherd G, Miller P, Wu E, Wong G, Liu N, et al.: **Semantic Web Meets e-Neuroscience: An RDF Use Case.** *In Proceedings of International Workshop on Semantic e-Science, ASWC 2006; Beijing, China Jilin University Press*; 2006:158-170.
84. Cheung K, Lam Y, Marenco L, Clark T, Gao Y, Kinoshita J, Shepherd G, Miller P, Wu E, Wong G, et al.: **AlzPharm: A Light-Weight RDF Warehouse for Integrating Neurodegenerative Data.** *5th Annual International Semantic Web Conference (ISWC); Athens, GA, USA* 2006.
85. Kinoshita J, Strobel G: **Alzheimer Research Forum: A Knowledge Base and e-Community for AD Research.** *In Alzheimer: 100 Years and Beyond* Edited by: Jucker M, Beyreuther K, Haass C, Nitsch RM, Christen Y. *Berlin Heidelberg: Springer-Verlag*; 2006:457-464. *Research and Perspectives in Alzheimer's Disease*
86. Zaccagnini D: **Design of a goal ontology for medical decision-support.** *In Masters of Science Massachusetts Institute of Technology, Harvard University – MIT Division of Health Sciences and Technology*; 2005.
87. Fox J, Alabassi A, Blank E, Hurt C, Rose T: **Modelling Clinical Goals: a Corpus of Examples and a Tentative Ontology.** *Symposium on Computerized Guidelines and Protocols (CGP-2004)* 2004.
88. Takeda A, Loveman E, Clegg A, Kirby J, Picot J, Payne E, Green C: **A systematic review of the clinical effectiveness of donepezil, rivastigmine and galantamine on cognition, quality of life and adverse events in Alzheimer's disease.** *International journal of geriatric psychiatry* 2006, **21**:17-28.
89. **Notation 3** 2006 [<http://www.w3.org/DesignIssues/Notation3.html>]. W3C

90. **ACPP N3 Logic Example** [<http://esw.w3.org/topic/HCLS/ACPPTaskForce/LogicFramework>]
91. Smith B, Rosse C: **The role of foundational relations in the alignment of biomedical ontologies.** *Medinfo* 2004, **11**:444-448.
92. **National Center for Biomedical Ontology Workshop on the Ontology of Clinical Trials** [[http://www.bioontology.org/wiki/index.php/Workshop\\_on\\_Clinical\\_Trial\\_Ontology](http://www.bioontology.org/wiki/index.php/Workshop_on_Clinical_Trial_Ontology)]
93. Marshall MS, Post L, Roos M, Breit TM: **Using semantic web tools to integrate experimental measurement data on our own terms.** In *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops* Edited by: Meersman R, Tari Z, Herrero P. Montpellier, France: Springer; 2006:679-688.
94. Carroll JJ, Bizer C, Hayes P, Stickler P: **Named Graphs.** *Journal of Web Semantics* 2005, **3**:32.
95. Stevens RD, Robinson AJ, Goble CA: **myGrid: personalised bioinformatics on the information grid.** *Bioinformatics* 2003, **19**(Suppl 1):302-304.
96. **W3C Rule Interchange Format Working Group** [<http://www.w3.org/2005/rules/wg/>]
97. Stevens RD, Robinson AJ, Goble CA: **myGrid: personalised bioinformatics on the information grid.** *Bioinformatics (Oxford, England)* 2003, **19**(Suppl 1):302-304.
98. **Bio-Health Informatics Group** [<http://www.cs.manchester.ac.uk/bhig/>]
99. **The National Center for Biomedical Ontology** [<http://www.bioontology.org/>]
100. **The OBO Foundry** [<http://obofoundry.org/>]
101. Good BM, Wilkinson MD: **The Life Sciences Semantic Web is full of creeps!** *Briefings in bioinformatics* 2006, **7**:275-286.

Publish with **BioMed Central** and every scientist can read your work free of charge

*"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."*

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:  
[http://www.biomedcentral.com/info/publishing\\_adv.asp](http://www.biomedcentral.com/info/publishing_adv.asp)

