Building an effective Semantic Web for Health Care and the Life Sciences

Editor(s): Krzysztof Janowicz, Pennsylvania State University, USA; Pascal Hitzler, Wright State University, USA
Solicited review(s): Rinke Hoekstra, Universiteit van Amsterdam, The Netherlands; Kunal Verma, Accenture, USA

Michel Dumontier*
Department of Biology, Institute of Biochemistry, School of Computer Science, Carleton University, 1125 Colonel By Drive, Ottawa, Ontario, Canada, K1S5B6

Abstract. Health Care and the Life Sciences (HCLS) are at the leading edge of applying advanced information technologies for the purpose of knowledge management and knowledge discovery. To realize the promise of the Semantic Web as a framework for large-scale, distributed knowledge management for biomedical informatics, substantial investments must be made in technological innovation and social agreement. Building an effective Biomedical Semantic Web will be a long, hard and tedious process. First, domain requirements are still driving new technology development, particularly to address issues of scalability in light of demands for increased expressive capability in increasingly massive and distributed knowledge bases. Second, significant challenges remain in the development and adoption of a well founded, intuitive and coherent knowledge representation for general use. Support for semantic interoperability across a large number of sub-domains (from molecular to medical) requires that rich, machine-understandable descriptions are consistently represented by well formulated vocabularies drawn from formal ontology, and that they can be easily composed and published by domain experts. While current focus has been on data, the provisioning of semantic web services, such that they may be automatically discovered to answer a question, will be an essential component of deploying Semantic Web technologies as part of academic or commercial cyberinfrastructure.

Keywords: semantic web, health care, life sciences, digital libraries, cyberinfrastructure, ontology

1. Introduction

The vision of the Semantic Web (SW) outlines that common standards for all aspects of knowledge management will facilitate the development of an interoperable ecosystem of data and services so that it becomes easier to publish, find, and re-use information in ways that go beyond their original design (Berners-Lee, Hendler, & Lassila, 2001). As a major consumer of information technologies, the Health Care and Life Sciences (HCLS) has traditionally placed demanding requirements to support activities related to knowledge management and knowledge discovery. While HCLS data is highly heterogeneous and growing at an unprecedented rate, SW technologies offer a salient solution to accurately publish this diverse knowledge in so that it becomes a major resource for research and development. In fact, the W3C Semantic Web HCLS Interest Group is specifically chartered to develop, advocate and support SW technologies for HCLS communities (HCLS, 2005). Our experience maintains that in order to build an effective Semantic Web for the HCLS, significant efforts still have to be made towards the coordinated development of high quality vocabularies, well thought out protocols for data sharing and publication, and scalable, cohesive cyberinfrastructure.

Coordinated efforts by a wide range of communities to promote a coherent representation of data will foster commoditization of information and create entirely new commercial opportunities and public-good efforts devoted to provisioning data, in-depth analysis and effective visualization. There is little doubt that by making biomedical data available through the Semantic Web, we will dramatically improve overall productivity, increase investment re-

* Corresponding author. E-mail: michel_dumontier@carleton.ca

turns, decrease the cost of research, create new economic activity and augment the outcomes of basic and applied research. The challenge then is to assess the vision for the Semantic Web with respect to the state-of-the art in knowledge representation and technology.

2. State of the Art

The SW positions itself as a platform for information exchange between intelligent agents. Interoperability is achieved by ensuring that the information is consistently encoded (syntax) and uses symbols that have a formally defined meaning such that they can be consistently interpreted (semantics). An effective Semantic Web will ensure interoperability between cyberinfrastructure components including i) capacity to capture knowledge, ii) infrastructure to publish and share information, iii) efficient middleware for question answering and knowledge discovery.

2.1. RDF and Linked Data

The Resource Description Framework (RDF) is a core SW language that offers a lightweight mechanism to describe entities in term of their types, attributes and relations to other entities. Entities are identified by International Resource Identifiers (IRIs) which includes web based identifiers (HTTP URIs) that can be resolved on the Web. Statements about these entities captured as subject-predicate-object "triples", and are described using vocabularies from domain-specific ontologies. RDF Schema (RDFS) makes it possible to specify simple type and relation hierarchies using the "is a" relation. RDF can be queried using the SPARQL query language.

A number of life science projects are using RDF as their core language of representation and publishing the information so that information about the entities can be queried and visualized. Bio2RDF[1] is at the forefront of generating and provisioning ~40 billion triples of linked life science data from over 40 high profile databases. Bio2RDF normalizes the data IRIs so as to facilitate linking of datasets (Belleau, Nolin, Tourigny, Rigault, & Morissette, 2008). Each dataset is deployed as its own SPARQL endpoint, which allows original data provider to actively participate in the network while decentralization of resource offerings provides web-scalability. Bio2RDF offers specialized federated query services across its global mirrors (Ottawa, Quebec City, Guelph and Brisbane). The Linking Open Drug Data (LODD) [2] and Chem2Bio2RDF[3] projects are generating linked data to support chemical-based investigations including drug discovery. These projects provision RDF data from relational databases using D2R. LinkedLifeData[4] consists of a diverse array of life science datasets provisioned through cluster-based data warehouse solution using the commercial BigOWLIM engine. Yet all of these projects largely involve information retrieval in the most basic sense, without making full use of the background knowledge provided by ontologies.

2.2. Ontologies

Initially driven by the need to query gene and gene product annotation across a number of model organisms, the Gene Ontology (GO) has emerged as an vast controlled vocabulary of biological processes, molecular functions and cellular components (GO Consortium, 2008). Since its inception, GO strives to more accurately describe their 20,000+ terms principally organized via an "is a" axis, but also augmented with other relations (e.g. parthood). Following GO, there are now over 150 Open Biomedical Ontologies (OBO) listed at the National Center for Bio-Ontology (NCBO) BioPortal, which now spans molecular, anatomical, physiological, organismal, health, experimental information. Yet significant overlap exists between ontologies, as a search yielding 20 different terms for "protein" will attest. Towards developing a set of orthogonal ontologies, the OBO Foundry (Smith et al., 2007) promotes development over basic categories drawn from the Basic Formal Ontology (BFO) and encourages the use of reuse basic, domain-independent relations from the Relational Ontology (RO). Well defined relations should make it clear when the relations are to be used, and what inferences, if any, may be drawn from them.

2.3. OWL and Linked Knowledge

Drawing from the well understood area of Description Logics, the Web Ontology Language (OWL) provides a substantially more expressive vocabulary to axiomatically describe entities for enhanced reasoning. Building these kinds of ontologies not only

[1] http://bio2rdf.org

[2] http://esw.w3.org/HCLSIG/LODD

[3] http://chem2bio2rdf.org

[4] http://linkedlifedata.org

requires domain expertise to properly define describe the entities, but also requires a keen understanding of formal knowledge representation so that knowledge is properly captured and becomes intuitive to query using an information system.

Several projects have now demonstrated the use of OWL-based information systems. The HCLS knowledge base contains a collection of instantiated ontologies used to identify interesting molecular agents in the treatment of Alzheimer's (Ruttenberg et al., 2007). With consideration of how genetics plays a role in effective drug treatment, the Pharmacogenomics Knowledge Base (PGKB) offers depression-related pharmacogenomic information that facilitates additional knowledge curation beyond the PharmGKB database (Dumontier & Villanueva-Rosales, 2009). Thus, ontologies can play an important role both in semantic data integration as well as guide curation activities with well established use cases towards populating a specialized knowledge base.

2.4. Semantic Web Services

Web services define application programming interfaces by structuring messages and content with the Web Services Description Language (WSDL). HCLS web services may be registered and annotated using the Web 2.0 inspired BioCatalogue (Goble, Stevens, Hull, Wolstencroft, & Lopez, 2008). Workflow application tools like Taverna facilitate chaining of services, to obtain and logically consume content (Oinn et al., 2004). Yet, the pairing of services still remains rather difficult because the inputs are generally datatypes as opposed to semantic types that can be reasoned about. SADI, a new semantic web services framework project, uses OWL ontologies to formally describe services, in which the Semantic Health And Research Environment (SHARE) query system undertakes service matchmaking and invocation through a SPARQL query (Vandervalk, McCarthy, & Wilkinson, 2009). This has been put to use in CardioSHARE, a system that integrates patient data with analytical services so as to identify *bone fide* cardiovascular health indicators.

3. Challenges

3.1. Scalable Semantic Web Technologies

Requirements of Semantic Web technologies have been drawn from extensive analysis of domain requirements, technical feasibility and vendor capabilities. While these including HCLS centric concerns, they do not reflect the enormous amounts of data (trillions of facts), nor the widespread and decentralized nature of databases (thousands of indirectly connected databases) that would have to be accommodated. Current stand-alone solutions appear to scale up into hundreds of millions of triples, while cluster-based solutions (Virtuoso Cluster Edition; BigOWLIM; BigData) appear to scale into the tens or hundreds of billions of statements, but with highly restricted capability to reason about OWL data. New and sustained efforts into large-scale reasoning and possibly incomplete reasoning may be required, as recently demonstrated (Urbani, Kotoulas, Maaseen, van Harmelen, & Bal, 2010).

3.2. From Linked Data to Linked Knowledge

RDF linked data efforts currently employ a simple model for representing knowledge: entities are either related to other entities or related to valued attributes through a single relation. Model 1 (Figure 1) exemplifies a typical linked data model for representing the volume of a protein using a decimal datatype. Such a model does not express the unit of measure, and no statements can be made as to how or under what conditions the value was obtained. In contrast, Model 2 overcomes these limitations by explicitly representing the entity, quality, measurement value, and the unit as distinct entities. However, moving from 2 triples in Model 1 to the 8 triples required in Model 2 translates to a 4x increase in the storage requirements and requires more sophisticated query to retrieve all the relevant information. The benefit increasing our capacity to make meaningful statements about any one of these entities, which cannot (easily) be done in Model 1, is nevertheless substantial.

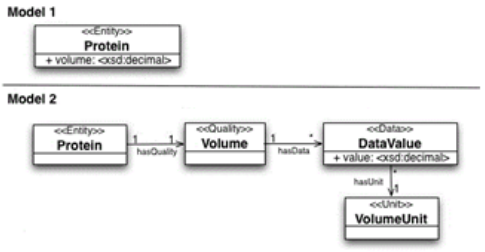

Figure 1 Two models for representing a physical attribute.

3.3. Consistent Knowledge Representation

If Model 2 is deemed desirable, then the challenge lies not only in getting scalable systems to accommodate this influx of triples (possibly by devising customizable indexes), but also in getting users to learn about and deploy standard patterns which they can apply to their own data. The patterns should be coherent, intuitive and well specified such that non-experts can read, understand and apply the guidelines found therein. Importantly, these patterns should specify the relations that hold between instances, and for this reason having a coherent, well founded set of types and basic relations supported by formal ontology is of critical value. While BFO+RO combination provides guidance for instantiable types, it lacks the capacity to handle all elements of scientific discourse (Dumontier & Hoehndorf, 2010), specifically with types that may be hypothesized (putative agents of disease), predicted (genes and proteins from genomic sequences), or simply do not occur (perpetual motion). This necessitates significantly more effort in developing a foundational ontology (types + relations) to represent a more diverse array of knowledge, including that which is *already* found in linked data.

Recent work by the W3C HCLS subgroup on translational medicine has produced a knowledge base composed of the Translational Medicine Ontology, which provides 75 core classes mapped to 223 classes from 40 ontologies, and acts as a global schema over a set of fake patient data and linking open data (LOD) resources (Dumontier et al., 2010). They featured queries that span bedside to bench by not only matching patients to clinical trials, but also in finding trials for which their drugs had different mechanisms of action so as to potentially avoid common side effects. Here, the integration of electronic health records with public data provides new avenues for clinical research and improved health care. With increased interest in building smarter health care systems using electronic health records, Semantic Web technologies can play a pivotal role in incentivizing interoperability between health care providers by linking valuable to public data.

3.4. The need for axiomatic description of classes

Until recently, OBO ontologies have been largely crafted using the OBO language, an ad-hoc language with its own (non-XML) syntax and lacking formal semantics. OBO ontologies differ enormously in terms of their development status, expressivity, and overall quality. While the standard transformation to OWL involves fixed semantics, more recent work demonstrates how more flexible semantics can be assigned as patterns associated with well defined relations such as the RO (Hoehndorf et al., 2010). Axiomatic description of classes should improve ontology quality by forcing ontology designers to be explicit about the necessary conditions for class membership, as opposed to relying on potentially vague descriptions using natural language. Such formalization can make use of automated reasoners to find errors and provide explanations for unexpected inferences.

3.5. Provenance and Attribution

Provenance and the corresponding attribution of knowledge is normal practice in science. Several approaches (Research Objects, Provenance Ontology, Provenir Ontology, SWAN-SIOC provenance) have now been articulated, and must now be unified. Importantly, contributions to community-based ontologies need to be acknowledged. Further, the wholesale provenance of data need to be specified, and while RDF reification or OWL axiom annotations supports this, they generate significantly higher overhead (4 triples per statement). In contrast, TriX/TRiG/RDF Named Graphs may be more effective and needs to go down the path of standardization.

3.6. User Interfaces

Despite a decade of research and development around Semantic Web technologies, significant gaps still remain in tools that facilitate data management and knowledge discovery. User interfaces are still developed "close to the metal", forcing a model that is not meant for human consumption. New innovative approaches need to consider FreeBase's Parallax[5], but for the Semantic Web. Impressively, the sig.ma[6] Mashup tool uses the Sindice Semantic Web Search engine to provide an enhanced view of indexed RDF triples, including those provided by Bio2RDF and DBpedia. For OWL knowledge bases, SMART (Battista, Villanueva-Rosales, Palenychka, & Dumontier, 2007) offers a way to craft queries as class expressions using the Manchester OWL syntax. Significantly more research in human-computer interaction is required to identify effective ways to work with

[5] http://www.freebase.com/labs/parallax/
[6] http://sig.ma/

with hyper-dimensional data from multiple (and possibly untrustworthy) sources.

4. Conclusion

Building an effective Semantic Web for HCLS is clearly a long term effort that needs coherent representations along with simple tools to create, publish, query and visualize generic semantic web data. With hundreds of bioinformatics web services, thousands of biological databases and millions of unrecorded facts in waiting, significant effort will also have to be placed in training the next generation of application developers to correctly use Semantic Web technologies. HCLS communities can then be served by custom portals, and ultimately act as a key component of cyberinfrastructure for both textual and semantically annotated data and services.

5. References

[1] Battista, A. D., Villanueva-Rosales, N., Palenychka, M., & Dumontier, M. (2007). *SMART: A Web-Based, Ontology-Driven, Semantic Web Query Answering Application*. Busan, South Korea: International Semantic Web Conference.

[2] Belleau, F., Nolin, M. A., Tourigny, N., Rigault, P., & Morissette, J. (2008). Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Semantics, 41*(5), 706-716.

[3] Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American, 284*, 34-43.

[4] Dumontier, M., & Hoehndorf, R. (2010). Realism for Scientific Ontologies. In *6th International Conference on Formal Ontology in Information Systems* (p. 12). Toronto, Canada.

[5] Dumontier, M., & Villanueva-Rosales, N. (2009). Towards pharmacogenomics knowledge discovery with the semantic web. *Briefings in Bioinformatics, 10*(2), 153-163.

[6] Dumontier, M., Andersson, B., Batchelor, C., Denney, C., Domarew, C., Jentzsch, A., et al. (2010). The Translational Medicine Ontology: Driving personalized medicine by bridging the gap from bedside to bench. In *Proceedings of Bio-Ontologies 2010: Semantic Applications in Life Sciences* (p. 4). Boston.

[7] GO Consortium. (2008). The Gene Ontology project in 2008. *Nucleic Acids Research, 36*(Database issue), D440-4.

[8] Goble, C., Stevens, R., Hull, D., Wolstencroft, K., & Lopez, R. (2008). Data curation + process curation=data integration + science. *Briefings in Bioinformatics, 9*(6), 506-17.

[9] HCLS. (2005). Semantic Web Health Care and Life Sciences Interest Group. Retrieved from http://www.w3.org/blog/hcls.

[10] Hoehndorf, R., Oellrich, A., Dumontier, M., Kelso, J., Herre, H., Rebholz-Schuhmann, D., et al. (2010). OWLDEF: Integrating OBO And OWL. In *Proceedings of the 7th International OWL: Experiences and Directions*. San Francisco.

[11] Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., et al. (2004). Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics (Oxford, England), 20*(17), 3045-54.

[12] Ruttenberg, A., Clark, T., Bug, W., Samwald, M., Bodenreider, O., Chen, H., et al. (2007). Advancing translational research with the Semantic Web. *BMC bioinformatics, 8 Suppl 3*, S2.

[13] Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology, 25*(11), 1251-1255.

[14] Urbani, J., Kotoulas, S., Maaseen, J., van Harmelen, F., & Bal, H. (2010). OWL reasoning with WebPIE: calculating the closure of 100 billion triples. *Proceedings of the 2010 Extended Semantic Web Conference*.

[15] Vandervalk, B. P., McCarthy, E. L., & Wilkinson, M. D. (2009). Moby and Moby 2: creatures of the deep (web). *Briefings in bioinformatics, 10*(2), 114-28.