Semantic Web for Health Care and Life Sciences: a review of the state of the art

Biomedical researchers need to be able to ask questions that span many heterogeneous data sources in order to make well-informed decisions that may lead to important scientific breakthroughs. For this to be achieved, diverse types of data about drugs, patients, diseases, proteins, cells, pathways and so on must be effectively integrated. Yet, linking disparate biomedical data continues to be a challenge due to inconsistency in naming and heterogeneity in data models and formats.

Many organizations are now exploring the use of Semantic Web technologies in the hope of easing the cost of data integration [1]. The benefits promised by the Semantic Web include integration of heterogeneous data using explicit semantics, simplified annotation and sharing of findings, rich explicit models for data representation, aggregation and search, easier re-use of data in unanticipated ways, and the application of logic to infer additional information [2].

The World Wide Web Consortium (W3C) (http://www.w3.org/) has established the Semantic Web for Health Care and Life Sciences Interest Group (HCLS IG) (http://www.w3.org/2001/sw/hcls/) to help organizations in their adoption of the Semantic Web. The HCLS IG is chartered to develop and support the use of Semantic Web technologies to improve collaboration, research and development, innovation, and adoption in the domains of Health Care and Life Sciences. As a part of realizing this vision, a workshop on the Semantic Web for Health Care and Life Sciences was organized in conjunction with WWW2008 (http://esw.w3.org/topic/HCLS/WWW2008) [3]. The workshop provided a review of the latest positions and research in this domain.

Five of the seven papers within this issue originated from the HCLS/WWW2008 workshop and review a range of Semantic Web technologies/approaches employed in different biomedical domains.

Vandervalk *et al*. describe 'The State of the Union' for the adoption of Semantic Web standards by key institutes in bioinformatics. The paper explores the nature and connectivity of several community-driven semantic warehousing projects. It reports on the progress with the CardioSHARE/Moby-2 project, which aims to make the resources of the 'Deep Web' transparently accessible through SPARQL queries. It points out that the warehouse approach is limited, in that queries are confined to the resources that have been selected for inclusion. It also discusses a related problem that the majority of bioinformatics data exist in the 'Deep Web', that is, the data does not exist until an application or analytical tool is invoked, and therefore does not have a predictable Web address. It also highlights that the inability to utilize Uniform Resource Identifiers (URIs) to address bioinformatics data is a barrier to its accessibility in the Semantic Web.

Das *et al*. discuss the use of ontologies to bridge diverse Web-based communities. The paper introduces the Science Collaboration Framework (SCF) as a reusable platform for advanced online collaboration in biomedical research. SCF supports structured Web 2.0 community discourse amongst researchers, makes heterogeneous data resources available to collaborating scientists, captures the semantics of the relationships between resources, and structures discourse around the resources. The first instance of the SCF framework is being used to create an open-access online community for stem cell research—StemBook (http://www.stembook.org). The SCF framework has been applied to interdisciplinary areas such as neurodegenerative disease and neuro-repair research, but has broad utility across the natural sciences.

Zhao *et al*. describe various design patterns for representing and querying provenance information relating to mapping links between heterogeneous data from sources in the domain of functional genomics. The paper illustrates the use of named RDF graphs at different levels of granularity to make provenance assertions about linked data. It also demonstrates that these assertions are sufficient to support requirements including data currency, integrity, evidential support and historical queries.

Dumontier *et al*. discuss a number of approaches for capturing pharmacogenomic data and other related information to facilitate data sharing and knowledge discovery. The paper describes how recent advances in Semantic Web technologies have presented exciting new opportunities for knowledge discovery related to pharmacogenomics by representing information with machine-understandable semantics. It illustrates progress in this area with respect to a personalized medicine project which aims to facilitate pharmacogenomics knowledge discovery through intuitive knowledge capture and sophisticated question answering using automated reasoning over expressive ontologies.

Manning *et al*. review several data integration approaches that involve extracting data from a wide variety of public and private data repositories, each of which is associated with a unique vocabulary and schema. The paper presents an implemented data architecture that leverages semantic mapping of experimental metadata to support the rapid development of scientific discovery applications. This achieves the twin goals of reducing architectural complexity while leveraging Semantic Web technologies to provide flexibility, efficiency and more fully characterized data relationships. The architecture consists of a metadata ontology, a metadata repository and an interface that allows access to the repository. The paper describes how this approach allows scientists to discover and link relevant data across diverse data sources. It provides a platform for development of integrative informatics applications.

Chen *et al*. survey the feasibility and state of the art for using Semantic Web technology to represent, integrate and analyze knowledge in a range of biomedical networks. The paper introduces a conceptual framework to enable researchers to integrate graph mining with ontology reasoning in network data analysis. Four case studies are used to demonstrate how semantic graph mining can be applied to the analysis of disease-causal genes, Gene Ontology (GO) category cross-talks, drug efficacy analysis and herb–drug interaction analysis.

Ruttenberg *et al*. review the use of Semantic Web technologies for assembling and querying biomedical knowledge from multiple sources and disciplines. The paper presents the Neurocommons prototype knowledge base, a demonstration intended to show the feasibility and benefits of using Semantic Web technologies. The prototype allows one to explore the scalability of current Semantic Web tools and methods for creating such a resource, and to reveal issues that will need to be addressed in order to further expand its scope and use. The paper demonstrates the utility of the knowledge base by reviewing a few example queries that provide answers to precise questions relevant to the understanding of the disease.

LOOKING TO THE FUTURE

There has been a considerable increase in the adoption of Semantic Web technologies in the life sciences and health care over the last 5 years. Much of the adoption has resulted from a strong need to be able to integrate and analyze data across databases, applications and communities. It has been fascinating to witness the breadth of solutions that have been implemented to meet these needs.

Some applications have focused on demonstrating the scalability of extremely large triple stores, running on platforms as diverse as clusters, PCs and iPhones. Another group of users is primarily focused on using the latest capabilities in OWL to further knowledge discovery through inference. Others have focused on Linked Data, which allows people to use data browsers to surf across silos of data that have been converted into linked data using approaches like relational to RDF mapping.

Recently, there has been much interest in Web 2.0, developments in social networking and mashups (Web applications that combine data from more than one source into a single integrated tool). However, many researchers are now exploring the additional capabilities that come with the Semantic Web (or Web 3.0) as it provides a machine-readable framework as to how people can say things about data. Using these technologies in concert transforms social networking sites from being fun pastimes for teenagers, to being serious research tools for sharing knowledge across communities. The incorporation of the Semantic Web into Wikis greatly enhances their usability through improved categorization and search of data. Scientific publishing has the potential to be transformed through support for community annotations, interconnected citations and services for semantically tagged key concepts and statements within papers.

Going forwards, we are expecting to see increased adoption of Semantic Web technologies within both industry and academic settings. It is looking likely that many implementations will focus on

light-weight approaches that enable the linking of data across silos, while a few large-scale efforts will rely on the use of heavy-weight ontologies as a top-down approach to data integration. We are also expecting a continued convergence of the Semantic Web and social networking, thereby enabling a more collaborative approach to science.

Kei-Hoi Cheung[1], Eric Prud'hommeaux[2], Yimin Wang[3] and Susie Stephens[4]
[1]Center for Medical Informatics, Yale University School of Medicine, New Haven, CT, USA
[2]World Wide Web Consortium, Cambridge, MA, USA
[3]Lilly Singapore Centre for Drug Discovery, Singapore and
[4]Discovery IT, Eli Lilly, Indianapolis, IN, USA

Acknowledgement
We thank the many reviewers who contributed their time and expertise to evaluation and revision of papers in this issue.

FUNDING
National Institutes of Health Grants (P01 DC04732 and U24 NS051869 to K.-H.C., in part).

References

1. Feigenbaum L, Herman I, Hongsermeier T, *et al*. The Semantic web in action. *Sci Am* 2007;**297**:64–71.
2. Baker CJO, Cheung KH, (eds). *Semantic Web: Revolutionizing Knowledge Discovery In The Life Sciences*. New York: Springer, 2007.
3. Chen H, Cheung KH, Dumontier M, *et al*. Report on Semantic Web for Health Care and Life Sciences. In: Proceedings of WWW2008, Beijing, China, pp. 1273–4.