Journal of Biomedical Informatics 41 (2008) 683–686

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Guest Editorial

Semantic mashup of biomedical data

1. Introduction

As the diversity and quantity of Web-accessible data in the biomedical domain grow, there are increasing benefits in empowering end-user scientists, working on their own, to integrate the various sources of data. Traditionally, significant programming effort has been required to parse and integrate heterogeneous datasets prior to enabling scientists to answer interesting questions. The heterogeneity includes different data formats, information models, and terminologies. Recently, a new breed of Web-based data-integration tools has been developed to simplify this process. They are called "mashups." These mashup tools have been designed to empower end-users to be able to extract, format, and remix data across multiple Web sites. Examples of such tools include Dapper (http://www.dapper.net/), which allows users to extract/scrape data from Web pages visually and to produce the extracted data as feeds in formats such as Rich Site Summary (RSS) (http://web.resource.org/rss/1.0/spec); Google Maps (http://maps.google.com), which provides the ability to mashup (integrate) datasets in the Keyhole Markup Language (KML) format and to visualize the integrated results; and Yahoo! Pipes (http://pipes.yahoo.com/pipes/), which provides operators/widgets to mashup heterogeneously formatted datasets (e.g., tabular, RSS, and KML formats). In addition to accessing user-friendly mashup tools, Web programmers can directly use open Web APIs, such as those listed in ProgrammableWeb (http://www.programmableweb.com/).

Mashup tools have been designed to allow disparate data sources to be brought together to increase utility to end-users. However, even with the tools and open APIs, users must perform most of the system integration. There is a need for creating mashups that better enable computers to help people achieve more powerful and complex data integration involving semantic mappings across multiple information models, terminologies, and ontologies. The term for such machine-based integration of data is "semantic mashups." The transition to semantic mashups is made possible using Semantic Web technology (http://www.w3.org/2001/sw/), which facilitates the sharing of the meaning of data. This in turn makes it much easier to combine the stovepipe systems and to integrate data in new and unexpected ways. The key components of the Semantic Web include RDF as the basic data model, OWL for expressive ontologies, and SPARQL for query. This special issue highlights the transition from mashups to semantic mashups in the context of biomedicine.

At the American Medical Informatics Association's Annual Symposium in 1998 (AMIA98), Sir Tim Berners-Lee gave the keynote speech on the role of the Web in the information-intensive era of health care and biomedical research. In his speech, Berners-Lee envisioned the transition of the Web from being human-oriented to being increasingly machine-friendly. This burgeoning vision of the machine-friendly Web later became the Semantic Web vision. Since the seminal publication on the Semantic Web in *Scientific American* in 2001 [1], the Semantic Web has progressed from being a vision to reality [2], although we still have some way to go before reaching the most futuristic aspects of the original *Scientific American* article. Adoption of the Semantic Web has been especially evident within health care and life sciences. In part, this has been driven by the World Wide Web Consortium (W3C), which created an interest group focused on the application of the Semantic Web to this domain area (http://www.w3.org/2001/sw/hcls/). The group has been chartered to develop and support the use of Semantic Web technologies and practices to improve collaboration, research and development, and innovation adoption in health care and the life sciences. Increased adoption has been observed in the form of increasing numbers of academic papers, special issues in journals (e.g., [3]), books (e.g., [4]), and conferences (e.g., [5]). An increasing number of implementations within commercial enterprises have also been documented (http://www.w3.org/2001/sw/sweo/public/UseCases/).

The annual World Wide Web (WWW) conference is one of the world's largest meetings for Web researchers, practitioners, and developers. A workshop titled "Health Care and Life Sciences Data Integration for the Semantic Web" (http://www2007.org/workshop-W2.php) was co-located with the WWW2007 conference. While Berners-Lee's AMIA keynote speech introduced the nascent vision of the Semantic Web to the biomedical informatics community, the workshop at WWW2007 provided concrete examples of how both academic and commercial organizations are embracing the technology. A number of the papers in this special issue of JBI originated at, and are expanded from, the workshop, while other papers were selected from submissions responding to the issue's public call for papers. The aim of this special issue is to raise awareness of the benefits of using Semantic Web technology for data integration within health care and life sciences. The following section outlines the organization of this special issue and gives a brief introduction to the papers.

2. Overview and organization

This issue starts with two methodological review papers [6,7] focused on an overview of mashups and semantic mashups in the context of health care and life sciences. Next come two papers [8,9] that describe how to use RDF to support semantic mashups of

1532-0464/$ - see front matter
doi:10.1016/j.jbi.2008.08.003

biomedical data. The following paper [10] describes how to map a relational database to unique identifiers in the life sciences. Then come seven papers [11–17] that discuss the use of OWL ontologies in representing knowledge and facilitating semantic mashups in different health care and life sciences domains, including Alzheimer's disease [11], drug addiction [12], neuroimaging [13], yeast biology [14], Chinese medicine [15], Dengue [16], and blood cell modeling [17]. We close with three papers [18–20] that discuss the use of semantic Web Services within the biomedical domain. The last of these articles describes the incorporation of agent computing into Web Services.

The methodological review paper by Goble and Stevens [6] provides a summary of the data integration problems in bioinformatics and describes different approaches to overcoming the challenges. The authors describe the characteristics of mashups and semantic mashups, as well as their differences. The former support a very lightweight approach to data integration, while the latter support a heavier but more standard approach to data integration. These human-friendly and machine-friendly data integration approaches help to build the bioinformatics nation [21].

The methodological review by Cheung et al. [7] provides a review of Web 2.0 and Web 3.0 (Semantic Web) approaches to integrating biomedical data. The paper describes use cases for testing the mashup capability of Web 2.0 tools, including Dapper and Yahoo! Pipes. In addition, the authors demonstrate how the Semantic Web can be used to annotate Web content for enabling semantic mashup. The paper discusses the potential benefits of combining Web 2.0 and Web 3.0 technologies to create more powerful tools for biomedical data integration.

The paper by Nolin and co-workers [8] highlights the advantages of using RDF as the standard format for building an infrastructure (Bio2RDF) for mashing up biomedical data. It also proposes a standard namespace for identifying data objects. The authors demonstrate how their system can be used to integrate a variety of public biomedical databases containing different but related types of data, including pathways, proteins, genes, and ligands, in the context of a Parkinson's disease use case.

The article by Gudivada et al. [9] describes how to use RDF to represent a semantic network of genomic and phenomic data. On the basis of this network representation, casual relationships are inferred. To approach the problem of inferring likely causality roles, the authors generate Semantic Web methods-based network data structures and perform centrality analyses to rank genes according to model-driven semantic relationships. This is tested by prioritizing genes that are involved in cardiovascular system diseases.

Bafna et al. [10] describe the implementation of semantic mashup through the mapping of relational databases to life sciences identifiers. A SQL-like language is defined for generating these identifiers. As a demonstration, this approach is applied to a relational database containing information necessary for constructing large-scale phylogenetic trees involving many different biological species.

The paper by Clark and co-workers [11] describes the SWAN project and its ontological framework for biomedical discourse. This framework has been developed in the context of building applications for biomedical researchers (e.g., Alzheimer's disease researchers). The paper describes the design approach of the SWAN ontology, explains its main classes/relationships and their applications, and shows its relationship to other ongoing activities in biomedicine.

Sahoo et al. [12] discuss how Semantic Web technologies can support information integration and simplify the creation of semantic mashups. This is demonstrated in the context of understanding the genetic basis of nicotine dependence. In this paper, gene and pathway information sources are integrated, and several complex scientific queries are answered using the integrated knowledge base. Also introduced in the paper is the Entrez Knowledge Model (EKoM), which is an information model in OWL for gene resources that is integrated with the BioPAX ontology for pathways.

The work by Temal et al. [13] describes a generic approach to building an application ontology. This approach is based on the reuse of a foundational ontology (DOLCE) and core components of domain-specific ontologies. It is applied to the neuroimaging area, involving both the objective nature of image data and the subjective nature of image content, through annotations based on interests expressed by both human users and computer programs.

The paper by Villanueva-Rosales and Dumontier [14] describes an OWL knowledge base for performing semantic data integration in the context of yeast biology. The authors discuss the challenges encountered during the construction of the knowledge base and how they are addressing these challenges. For example, the knowledge base makes use of ontologies to integrate identical resources from different data providers and overcomes the problem of data integration when heterogeneous identifiers have been used.

Mao et al. [15] address both scalability and evolvability of large ontologies. Their study is evaluated in the context of traditional Chinese medicine. Their approach involves caching context-specific subontologies for boosting performance. In addition, a genetic algorithm is used to optimize the quality of subontologies for dynamic knowledge reuse.

The paper by Rajapakse et al. [16] presents a literature-driven, ontology-centric navigation infrastructure comprising a content acquisition engine, a domain ontology, and an ontology instantiation pipeline delivering sentences related to Dengue that have been derived by text mining. Also included in the infrastructure is a visual query tool for OWL querying and reasoning. This informatics infrastructure is tested with the literature relating to Dengue disease. It demonstrates how such an infrastructure can simplify searching and knowledge discovery for Dengue, with implications for other, similar application domains.

Novacek [17] discuss a dynamic ontology lifecycle scenario (DINO) involving ontology creation, versioning, evaluation, and negotiation. Their work also incorporates the notion of ontology learning into ontology integration. Particularly, the semi-automatic integration of ontology learning results into a manually created ontology is developed. This approach involves using methods of automatic negotiation of agreed ontology alignments, inconsistency resolution, and natural language generation. It is demonstrated in the context of extending an ontology fragment related to blood cells.

Dang et al. [18] explore the combination of Semantic Web and Service technologies including Business Process Management and Service Oriented Architecture to build an adaptive medical workflow system. An ontology is designed for capturing knowledge for a complex personalized health care scenario. The ontology also allows users to create and manage context-aware medical workflows and to execute them dynamically.

The article by DiBernardo et al. [19] demonstrates how a Semantic Web framework can help to manage and assemble a large number of existing bioinformatics Web Services such as those registered in BioMoby. This paper tackles the problem of automated service composition by annotating services and their interfaces with semantic information. It also features reasoning over services based on their composite types. A prototype workflow assembly client is implemented to help users to select and rank services of their interest. In addition, an evaluation is performed to show the effectiveness of the approach in terms of assisting the user to find their desired services quickly during the assembly process.