Journal of Biomedical Informatics 41 (2008) 694–705

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Methodological Review

HCLS 2.0/3.0: Health care and life sciences data mashup using Web 2.0/3.0

Kei-Hoi Cheung [a,*], Kevin Y. Yip [b], Jeffrey P. Townsend [c], Matthew Scotch [a]

[a] *Center for Medical Informatics, Yale University, 300 George Street, Suite 501, New Haven, CT 06511, USA*
[b] *Department of Computer Science, Yale University, USA*
[c] *Department of Ecology & Evolutionary Biology, Yale University, USA*

ARTICLE INFO

Article history:
Received 14 February 2008
Available online 11 April 2008

Keywords:
Web 2.0
Integration
Mashup
Semantic Web
Biomedical informatics
Bioinformatics
Life sciences
Health care
Public health

ABSTRACT

We describe the potential of current Web 2.0 technologies to achieve data mashup in the health care and life sciences (HCLS) domains, and compare that potential to the nascent trend of performing semantic mashup. After providing an overview of Web 2.0, we demonstrate two scenarios of data mashup, facilitated by the following Web 2.0 tools and sites: *Yahoo! Pipes*, *Dapper*, *Google Maps* and *GeoCommons*. In the first scenario, we exploited *Dapper* and *Yahoo! Pipes* to implement a challenging data integration task in the context of DNA microarray research. In the second scenario, we exploited *Yahoo! Pipes*, *Google Maps*, and *GeoCommons* to create a geographic information system (GIS) interface that allows visualization and integration of diverse categories of public health data, including cancer incidence and pollution prevalence data. Based on these two scenarios, we discuss the strengths and weaknesses of these Web 2.0 mashup technologies. We then describe Semantic Web, the mainstream Web 3.0 technology that enables more powerful data integration over the Web. We discuss the areas of intersection of Web 2.0 and Semantic Web, and describe the potential benefits that can be brought to HCLS research by combining these two sets of technologies.

1. Introduction

Web 2.0 refers to a second generation of Internet-based services—such as social networking sites, wikis, communication tools, and folksonomies—that emphasize online collaboration and sharing among users (http://www.paulgraham.com/web20.html). If the first generation Web has revolutionized the way people access information on the Internet, Web 2.0 has revolutionized the way people communicate across the Internet. Web 2.0 has transformed the Web into an environment that provides richer user experiences by allowing for the combination of disparate information in a variety of data formats, the facilitation of interaction between multiple parties, and the collaboration and sharing of information. Web 2.0 consists of a variety of applications implemented using diverse technologies. In general, the variety of Web 2.0 applications can be classified as follows:

- *Rich Internet applications.* These applications behave very much like desktop applications, and are easy to install and easy to use. In particular, they provide a dynamic interface with interactive features like point-and-click/drag-and-drop. These interfaces are achieved with technologies such as Ajax (Asynchronous JavaScript and XML) (http://en.wikipedia.org/wiki/AJAX), and mini plug-in programs known variously as widgets, gadgets and snippets, which create a programming environment within the browser and allow the user to easily combine information and create a variety of graphical presentations. As a result of this progress, the gap between Web programming and desktop programming has been diminishing (http://blogs.adobe.com/shebanation/2007/02/desktop_application_programmin.html).
- *Collaboration tools.* These include asynchronous collaboration tools such as wikis and blogs, to which users do not need to be simultaneously connected at any given time to collaborate. This category also includes synchronous, real-time (or near real-time) collaboration enablers, such as leading-edge instant messaging tools.
- *User-contributed content databases.* These are large-scale environments—such as YouTube, a video posting Web site, and Flickr, a photo-sharing site—in which users share content in multimedia format.
- *Integrative technologies enabling the Web as a platform.* There are abundant services and data sources scattered over the Internet. While they may be accessed independently, it has been exceedingly challenging to integrate Web-based services to create novel functionality. Web 2.0 mashup offers a solution to this problem. Mashup tools like *Yahoo! Pipes* (http://pipes.yahoo.com/pipes/) offer a graphical workflow editor that allows the user to pipe Web resources together easily. Other tools like

[*] Corresponding author. Fax: +1 203 737 5708.
E-mail address: kei.cheung@yale.edu (K.-H. Cheung).

1532-0464/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2008.04.001

Dapper (http://www.dapper.net/) provide an easy way for users to extract (or scrape) Web contents displayed in heterogeneous formats and output the extracted contents in a standard format such as tab-delimited values and XML. Data visualization tools like *Google Maps* (http://maps.google.com/) and *Google Earth* (http://earth.google.com/) offer a GIS (Geographic Information System) interface for displaying and combining geographically related data. Despite their different functionalities, these tools may interoperate. For example, the output of *Dapper* may be fed into *Yahoo! Pipes*, and *Yahoo! Pipes* in turn can be linked to *Google Map* to process and display geographical data.

The rest of the paper is structured as follows. Section 2 gives an overview of data integration in health care and life sciences domains. Section 3 describes two scenarios demonstrating the use of a number of Web 2.0 tools/sites in achieving health care and life science data mashups. Section 4 discusses the strengths and weaknesses based on our experience with these Web 2.0 tools/sites. Section 5 introduces Web 3.0 with a main focus on Semantic Web and its potential application in health care and life sciences data mashup (semantic mashup). Section 6 discusses how Web 2.0 and Semantic Web can be combined to reap a greater benefit. Section 7 gives a conclusion. Finally, a glossary table is provided for defining/describing the terms related to Web 2.0/3.0 with examples.

2. Health care and life sciences data integration

The popularity of the Web [1] and the success of the Human Genome Project (HGP) [2] have led to an abundance and diversity of biomedical data available via the Web. Fig. 1 indicates the rate of growth in the number of Web-accessible biological databases that were published in the annual Database Issue of Nucleic Acids Research (NAR) between 1999 and 2005. These databases (which only represent a small portion of all biomedical databases in existence today) play an indispensable role in modern Health Care and Life Sciences (HCLS) research. They facilitate data mining and knowledge discovery [3]. The benefits for integrating these databases include the following:

- HCLS data are more meaningful in context, while no single database supplies a complete context for a given HCLS research study.
- New hypotheses are derived by generalizing across a multitude of examples from different databases.
- Integration of related data enables validation and ensures consistency.

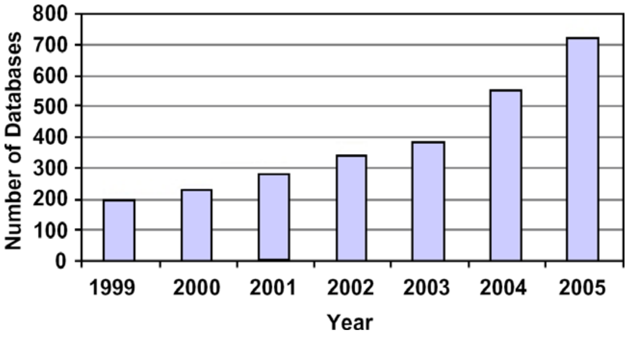

Fig. 1. Number of databases published in the NAR Database Issues between 1999 and 2005.

Via a Web browser, an HCLS researcher may easily access diverse information including DNA sequences, biochemical pathways, protein interactions, functional domains and annotations, gene expression data, disease information, and public health data. Integrating such data from diverse sources, however, remains challenging. Researchers wishing to analyze their own experimental data in combination with publicly available data face the cumbersome tasks of data preprocessing and cleaning [4], which includes scraping Web pages, converting file formats, reconciling incompatible schemas, and mapping between inconsistent naming systems. Even experienced programmers find such data integration tasks daunting and tedious.

A variety of approaches, including data warehousing [5,6], database federation [6,7], and Web services [8,9], have been developed to facilitate data integration in the context of HCLS. One problem with these approaches is that they require their developers to have significant database/programming expertise. Moreover, these systems may not be able to anticipate or offer the flexibility needed by the end users (who may themselves not be well versed programmers). Furthermore, it is difficult if not impossible for these systems to keep up with the growth of Web data sources. There are very few such systems that allow the user to add new external data sources easily, especially ones that do not conform to standard data formats.

To address these problems, Web 2.0 mashups have emerged. A mashup is a Web application that combines multiple third-party services over the Web. Numerous mashup examples are available from www.programmableWeb.com. Most of the current mashups are for non-scientific use. The potential of data mashup in the HCLS domains has only recently been demonstrated by using *Google Earth* to geographically integrate and visualize different types of data, including epidemiological and public health data, to help track the global spread of avian influenza [10]. However, more HCLS use cases are needed to demonstrate the need and value of Web 2.0 mashups in the HCLS domains.

3. Mashup scenarios

We provide two scenarios that illustrate the use of several Web 2.0 mashup tools and sites to implement data integration in the HCLS domains. The first scenario, within a life sciences context, shows how to use *Dapper* and *Yahoo! Pipes* to integrate diverse data such as microarray measurements and gene annotation data. The second scenario, within a public health context, demonstrates how to geographically correlate cancer data with environmental data using *Yahoo! Pipes*, *Google Maps*, and *GeoCommons* (http://www.geocommons.com/).

3.1. Life sciences scenario

Fig. 2 shows the workflow of a typical research study featuring the use of a spotted microarray, one kind of microarray technology. As shown in the figure, two biological samples (normal vs. disease), which consist of quantitatively distinct distributions of mRNA sequences, are labeled with fluorescent dyes. Sequences transcribed from the disease sample mRNA are labeled with the red fluorescent dye and sequences transcribed from the normal sample mRNA are labeled with the green fluorescent dye. Next, the two labeled samples are mixed in equal total amount, and that mixture is allowed to ‘‘hybridize” (bind) to the affixed reference sequences that have been deposited on the surface of a chemically-treated microscopic glass slide. Each spot on the slide contains many strands of the DNA sequence corresponding to one specific gene. A large number of spots, and therefore many gene sequences, may be featured on a given slide.