Challenges of a “What”-Oriented Framework for On-the-fly Integration of Biomedical Data

Shahram Ghandeharizadeh
Director of Database Laboratory
Computer Science Department
University of Southern California
Los Angeles, California 90089
shahram@usc.edu

Giri Narasimhan
Bioinformatics Research Group (BioRG)
School of Computing & Information Sciences
Florida International University
Miami, FL 33140
giri@cs.fiu.edu

A major informatics challenge for biomedical research is to design the framework to facilitate on-the-fly integration of scientific data. On-the-fly integration refers to scenarios where a scientist wants to integrate data from a source immediately after discovering it. The challenge is to significantly reduce the required time and skill to deal with the information technology problems and to allow the scientist to focus on the domain-specific problem. It may include both social and technical issues, see (Kötter, 2001) for a neuroscience perspective.

The ideal framework to address the identified research challenge must be “what”-oriented, empowering a scientist to specify “what” functionality they want to integrate. The framework should resolve the details of “how”. This frees the scientist to focus on their domain challenge, navigating the functionality of published data sources. The scientist may incorporate and discard operations at will. The framework would execute the compositions of a scientist quickly, empowering them to continuously change and refine their integrations. A scientist may share an instance of integration with other colleagues using electronic means. Moreover, a scientist may publish an instance of integration to be used by other members of their community. Thus a single integration can be viewed as a composition of other integrations.

Web Services (WSs) hold the potential to realize the framework envisioned above (Saxena *et al.,* 2005); for an example see [http://dblab.usc.edu/Sangam]. A Web Service (WS) is a network enabled application component with service-oriented architecture using standard interface description languages and communication protocols that facilitate easy development and deployment of data intensive applications. It is an emerging technology with the following key advantages. First, a WS and its operations represent the key functionalities supported by a data source. In essence, they represent the insights of the scientist who has analyzed that data source. This addresses a social challenge of sharing biomedical data by empowering a researcher to share the functionality they want to publish. With today’s high-throughput technology, a scientific effort may produce large volumes of data potentially requiring years to analyze. A scientist may not be willing the share all the data for one publication or may have proprietary data. Instead, s/he might be willing to share those portions of the data that they have processed and accepted for publication. The concept of WSs facilitates this mode of interaction. Second, software development environments such as Microsoft’s .NET and Sun’s Java have made publication of WS operations trivial. As a testament to this claim, note that governmental agencies such as NCI and NCBI have published their WS operations.

Why have scientists not adopted the concept of WSs widely? The answer rests with a lack of a “what”-oriented framework to minimize the required programming skills. At the time of this writing, the simple task of integrating two WSs requires programming skills, placing the concept of WSs beyond the reach of many scientists. This is because WSs are relatively new, emerging technologies. At the time of this writing, they remain relatively unexplored and untested in the context of informatics challenges in biomedical research.

One may use WSs as the building blocks of extendible visual frameworks. Such a framework would register a WS when a user specifies its URL. They facilitate integration of multiple WSs to produce composed plans. Relational algebra operators may serve as the glue to combine autonomous WSs together (Alwagait and Ghandeharizadeh, 2004). Second, they automate the composition task either fully or through interactions with a scientist. A sample scenario usage might be as follows. The scientist registers discovered WSs and specifies the desired output. The framework composes plans that produce the desired output. When it is impossible to compose a plan, the framework would identify what data functionality (WS operations) is missing. Third, the framework should learn relevant WSs based on those registered by a scientist. This can be used to discover new WSs as they are published on the Internet. The framework would bring these to the attention of the scientist, allowing them to incorporate them incrementally.

Several core concepts must be investigated. First, the framework must be fast in order to be interactive. This depends on the restrictions imposed by the participating WSs. To illustrate, NCBI requires users intending to submit numerous queries to make no more than one request every 3 seconds. The framework may respect these guidelines and provide fast response times using caching techniques. Alternatively, the database community may help service providers eliminate their restrictions by either replicating read-only WSs (Alwagait and Ghandeharizadeh, 2004) or by partitioning of update intensive WSs. Second, ontologies to address semantic heterogeneity are at the heart of the proposed frameworks. The framework learns these, requires a user to manually register them, uses a standardized ontology published by a community, or employs a hybrid of these possibilities. A scientist should be able to interrogate and enquire about the ontologies employed by the framework. The framework must include tools to allow multiple scientists to exchange their ontologies, detect and resolve conflicts. Third, the envisioned framework should complement the existing HTML infrastructure by allowing a scientist to embed their composed plans into an HTML form, thus enabling a scientist to construct complex web applications using composed plans.

Summarizing, **future biomedical research needs environments in which different sources of data can be seamlessly integrated on-the-fly and in which researchers can share and collaborate, yet protect parts of the data, and in which researchers with minimal information technology skills can creatively interrogate, navigate, and integrate scientific data from multiple sources.** WSs and Service Oriented Architectures (SOA) hold the potential to realize a "what"-oriented framework for on-the-fly integration of scientific data. If successful, the framework will revolutionize how biomedical researchers share and exchange data. The National Library of Medicine is in a unique position to nurture this relatively novel and unexplored technology to address a difficult, yet tractable informatics challenge.

References

E. Alwagait and S. Ghandeharizadeh. A Comparison of Alternative Web Service Allocation and Scheduling Policies. In IEEE International Conf. on Services Computing (SCC), Sept 2004.

M. Saxena, S. Kim, E. Alwagait, A. M. Khan, G. Burns, J. Su, A. G. Watts, and S. Ghandeharizadeh. Sangam: A Data Integration Framework for Studies of Stimulus-Circuitry-Gene Coupling in the Brain. Society of Neuroscience, Neuroscience 2005, Washington D.C., November 12-16, 2005.

R. Kötter. Neuroscience Databases: Tools for Exploring Brain Structure-Function Relationships. Phil. Trans. R. Soc. Lond. B (2001) 356, 1111-1120.