LifeDB: An Autonomous Semantic Data Integration System for Life Sciences*

Anupam Bhattacharjee, Aminul Islam, Mohammad Shafkat Amin, Shahriyar Hossain, Shazzad Hosain, and Hasan Jamil
Department of Computer Science, Wayne State University, USA
{anupam,aminul,shafkat,shah_h,shazzad,hmjamil}@wayne.edu

Abstract

Data intensive applications in Life Sciences extensively use the Hidden Web as a platform for information sharing. Access to Hidden Web resources is limited through the use of predefined web forms and interactive interfaces that users must navigate manually. Hence, the effective use of these resources rely on users' rational interpretation of the associated schema and the presented information. Since the computational model for an application is usually in the user's mind, the user is responsible for reconciling schema heterogeneity, mediating missing information, extracting information and piping, transforming format and so on in order to implement the desired query sequences or scientific workflows. While simple and relatively modest applications can be implemented and executed this way, large scale scientific investigations are often hard to capture, reuse, and maintain without the support of a set of sophisticated tools.

The traditional solution to this problem was to download needed data from different sources to a local machine and then develop customized applications to implement the workflow in mind by manually reconciling the schema and format heterogeneity. Although this alternative is efficient, it lacks currency and invites view materialization related complications. A second alternative is to write glue codes, say in Perl, or Java, to connect the remote sites, download data, and run queries . Although the advantage here is increased currency and less maintenance, the disadvantage is the increased cost of needed programming.

In LifeDB, we offer a third alternative that combines the advantages of the previous two approaches – currency and reconciliation of schema heterogeneity, in one single platform through a declarative query language called BioFlow. In our approach, schema heterogeneity is resolved at run time by treating the hidden web resources as a virtual warehouse, and by supporting a set of primitives for data integration on the fly to extract information and pipe to other resources, and to manipulate data in a way similar to traditional database systems in order to meet application demands. We use a state of the art schema matching system called OntoMatch, a wrapper generation system called FastWrap, and the latest internet computing tools to build LifeDB and to design a query processing engine for BioFlow. We offer several language constructs to support mixed-mode queries involving XML and relational data, application design using stored workflows, structured programming using process definition and reuse, and workflow design using ordered process graphs. We demonstrate the salient features of our system using a substantial set of online examples in real time. Finally, we show that a graphical tool can be used by a novice user to design applications in BioFlow without actually knowing anything about the language. Readers may refer to the lab home page at http://integra.cs.wayne.edu/ for further information on BioFlow and LifeDB.

* Research supported in part by National Science Foundation grants CNS 0521454 and IIS 0612203.