Public Abstract
Gaurav Ashokkumar Sanghi
M.S.
Computer Science
Automating Database Curation with Workflow Technology
Advisor: Dr. Toni Kazic
Graduation Term: Winter 2005

PUBLIC ABSTRACT

Building scientific databases is extremely difficult and expensive in terms of time and money. Expert staffs at various levels are needed to curate and maintain data, and each database has its own design and domain model specific to its requirements. Costs could be reduced if the experts who curate the data are provided with data that are reviewed by other experts at lower levels for accuracy and consistency. Since expertise is distributed around the world, a common platform that implements a well-accepted work process is needed to support such community curation. To improve the productivity of curation and to help people capture more data in databases, I have automated the workflow used in curating several types of biochemical data. The workflow is complicated because there are many different types of biochemical data and the relationships among the data are complex; different data types need different kinds of checks; different groups of curators and experts use different procedures to deposit, review, revise, and accept data; and the volume of data is very large.

Using The *Agora* as an example, I have automated the procedures involved in data curation, minimizing human intervention and eventually reducing errors in the database. This model is flexible enough to accommodate additional processes idiosyncratic to particular groups of curators, such as those for enzymatic reactions, biochemical terms, and molecular structures. This work demonstrates the application of workflow technology to intellectually complex, geographically distributed, multidisciplinary scientific processes.