

AUTOMATING DATABASE CURATION WITH WORKFLOW TECHNOLOGY

Gaurav Ashokkumar Sanghi

Dr. Toni Kazic, Thesis Supervisor

ABSTRACT

Building scientific databases is extremely difficult and expensive. Costs could be reduced if the experts who curate the deposited 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. 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; procedures to deposit, review, revise, and accept data; and the volume of data is very large. We have automated the workflow used in curating several types of biochemical data. 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.