XMLPipeDB: A Reusable, Open Source Tool Chain for Building Relational Databases from XML Sources

Kam D. Dahlquist[1], Alexandrea Alphonso[1], Derek Smith[2], Chad Villaflores[1],
John David N. Dionisio[2]

[1]Department of Biology, [2]Department of Electrical Engineering and Computer Science, Loyola Marymount University, 1 LMU Drive, Los Angeles, California, 90045 USA

Project Site: http://xmlpipedb.cs.lmu.edu
Source Code Availability: http://sourceforge.net/projects/xmlpipedb

XMLPipeDB is an open source tool chain for building relational databases from XML sources with minimal manual processing of the data. While its applicability is intended to be general, the original motivation for XMLPipeDB was to create a solution for the management of biological data from different sources that are used to create Gene Databases for GenMAPP (Gene Map Annotator and Pathway Profiler), software for viewing and analyzing DNA microarray and other genomic and proteomic data on biological pathways. XMLPipeDB has a modular architecture with three components that may be used separately or together. XSD-to-DB reads an XSD (XML Schema Definition) and automatically generates an SQL schema, Java classes, and Hibernate mappings. XMLPipeDB Utilities provides functionality for configuring the database, importing data, and performing queries. GenMAPP Builder is based on the XMLPipeDB Utilities and exports GenMAPP-compatible Gene Databases based on data from UniProt and Gene Ontology (GO). We have previously used GenMAPP Builder to create Gene Databases for *Escherichia coli* K12 and *Arabidopsis thaliana*. We have extended GenMAPP Builder by automating the process of creating new Gene Databases for any additional species for which UniProt data are available. We have thus recently created Gene Databases for *Plasmodium falciparum* and *Vibrio cholerae*. We have also added functionality to GenMAPP Builder that automatically checks for data integrity from the original XML source, the intermediary PostgreSQL relational database, and the exported GenMAPP Gene Database. GenMAPP Builder has proved to be robust to changes in the XSDs from UniProt and GO, both of which have changed several times throughout the course of this project. We also report on the compatibility of other common bioinformatics XML formats with the XMLPipeDB suite.