Poster B-36

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

Authors:
Kam D. Dahlquist (*Loyola Marymount University*)
Joey Barrett (*Loyola Marymount University*)
Joe Boyle (*Loyola Marymount University*)
Adam Carasso (*Loyola Marymount University*)
David Hoffman (*Loyola Marymount University*)
Babak Naffas (*Loyola Marymount University*)
Jeffrey Nicholas (*Loyola Marymount University*)
Roberto Ruiz (*Loyola Marymount University*)
Scott Spicer (*Loyola Marymount University*)
John David N. Dionisio (*Loyola Marymount University*)

Short Abstract: XMLPipeDB is an open source suite of Java-based tools for automatically building relational databases from an XML schema (XSD). XMLPipeDB can be used for the management of biological data from different sources. We have used it to generate GenMAPP Gene Databases containing UniProt and Gene Ontology data for Escherichia coli and other bacterial species.

Long Abstract:

XMLPipeDB is an open source suite of Java-based tools for automatically building relational databases from an XML schema (XSD). XMLPipeDB provides functionality for managing, querying, importing, and exporting information to and from XML data with minimum manual processing of the data. While its applicability is fairly 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. The GenMAPP Gene Database is used to relate gene IDs on MAPPs (representations of pathways and other functional groupings of genes) to data in Expression Datasets (e.g., DNA microarray or other high-throughput data). GenMAPP is a stand-alone application that requires the Gene Database, MAPPs, and Expression Dataset files to be stored on the user's computer.

The creation of Gene Databases for GenMAPP has been difficult because there are a number of different gene ID systems in common usage such as those provided by NCBI, UniProt, Ensembl, and the various model organism databases. Most of the genes and proteins in these systems overlap, but they have different identifiers and different types of annotation associated with them. Since different platforms for high-throughput data often use different gene identifiers to annotate their results, we have the problem of relating one set of gene identifiers to the other.

The GenMAPP Gene Databases are designed to accommodate these different gene ID systems, relating them to one another. In the past, the GenMAPP.org project team shouldered the full responsibility of merging the various ID systems using a complicated

process of building relationships from independent tables collected from greater than twenty different sources. This led to serious concerns over data integrity and transparency. Furthermore, the process of updating the database was onerous because changes made by the data sources to the format and/or content of the data required substantial manual modifications to the database updating process. To ameliorate these problems, the GenMAPP Gene Databases currently use the integrated data source from Ensembl, which is dedicated to relating gene ID systems and annotations. However, this limits the number of species that can be represented in GenMAPP to the mostly animal species supported by Ensembl.

Here we report that we have used the XMLPipeDB software tool chain to create relational databases for UniProt and Gene Ontology. In turn, we have used these databases to generate UniProt-centric GenMAPP Gene Databases for Escherichia coli and other bacterial species, extending the functionality of GenMAPP to species not currently supported by the GenMAPP.org project team. Moreover, since XMLPipeDB can create the relational databases based solely on the XSD and XML files, it will be more robust to changes in the source files made by the data providers.

XMLPipeDB has the following tools for developers and database designers: the XSD-to-DB application takes a well-formed XSD or DTD file and converts it into a collection of Java source code and Hibernate mapping files that allows XML files based on that definition file to be read into a relational database. XSD-to-DB's conversion functions are based on the open source Hyperjaxb2 project, which adds Hibernate functionality to Sun Microsystems' JAXB library. The XMLPipeDB Utilities library is a suite of Java classes that provide functions needed by many XMLPipeDB database applications. Specifically, the library includes reusable classes for: importing XML files into Java objects, saving these XML-derived Java objects to a relational database, querying the relational database using either HQL (Hibernate Query Language) or SQL, and configuring a client application to communicate with a relational database. Finally, GenMAPP Builder is an application for creating the GenMAPP Gene Database files. GenMAPP Builder's UniProt and Gene Ontology database libraries were generated with XSD-to-DB, and the application itself uses the XMLPipeDB Utilities library. The application works by first importing UniProt and Gene Ontology XML files as well as a tab-delimited UniProt-to-GO associations file into a relational database. The database can then be queried by organism in order to produce a GenMAPP Gene Database.

GenMAPP Builder has been tested for use with the open source PostgreSQL relational database, but can be used with any other relational database management system for which a JDBC driver is available. JDBC-to-ODBC connectivity is used to transfer data from this relational database to a Microsoft Access MDB file, which is the format expected by the GenMAPP application.

XMLPipeDB is available under the GNU Library or Lesser General Public License (LGPL) at http://sourceforge.net/projects/xmlpipedb.