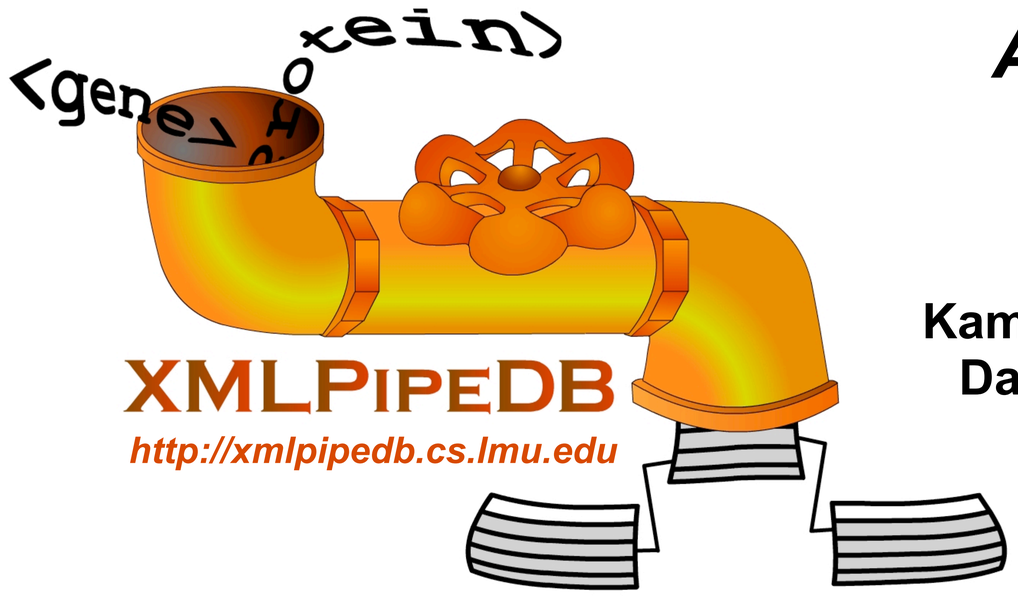


A Reusable, Open Source Tool Chain for Building Relational Databases from XML Sources

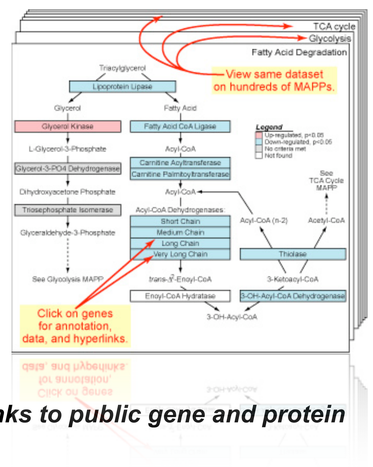
Kam D. Dahlquist[1], Joey Barrett[2], Joe Boyle[2], Adam Carasso[2], David Hoffman[2], Babak Naffas[2], Jeffrey Nicholas[2], Roberto Ruiz[2], Scott Spicer[2], John David N. Dionisio[2]

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

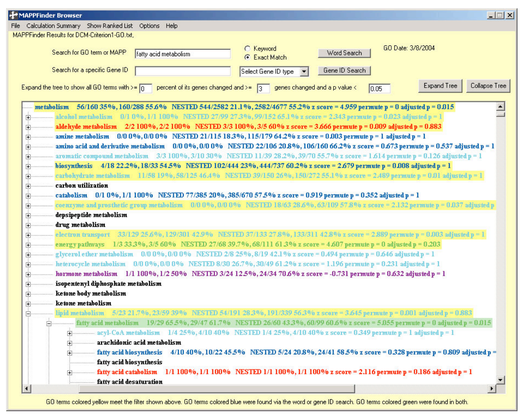
XMLPipeDB is an open source suite of Java-based tools for automatically building relational databases from an XML schema (XSD). While its applicability is fairly general, the original motivation for XMLPipeDB was to create a solution for the management of biological data from different sources that are used to create Gene Databases for GenMAPP (Gene Map Annotator and Pathway Profiler). 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, such as 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, and has been used to generate a GenMAPP Gene Database for *Escherichia coli* K12 using XML provided by UniProt and Gene Ontology.

Background & Motivation

GenMAPP is a free program for viewing and analyzing DNA microarray or other genomic and proteomic data on biological pathways



- **Standalone program implemented in Visual Basic**
  accessory files are Microsoft Access databases
- **Graphics tools for drawing MAPPs**
- **MAPPs represent biological pathways and other functional groups of genes**
  `.mapp` *files store gene IDs and vector coordinates for all graphical objects*
- **Underlying Gene Database**
  species-specific `.gdb` *files store IDs, annotation, and hyperlinks to public gene and protein databases*
- **Import expression data and set criteria to automatically color the MAPPs according to the data**
  separate Expression Dataset files (`.gex`*) store data and color-coding instructions*



MAPPFinder finds Gene Ontology terms over-represented in a GenMAPP Expression Dataset and ranks them by *p* value

Maintaining and Updating GenMAPP Gene Databases has been a Bottleneck for Development

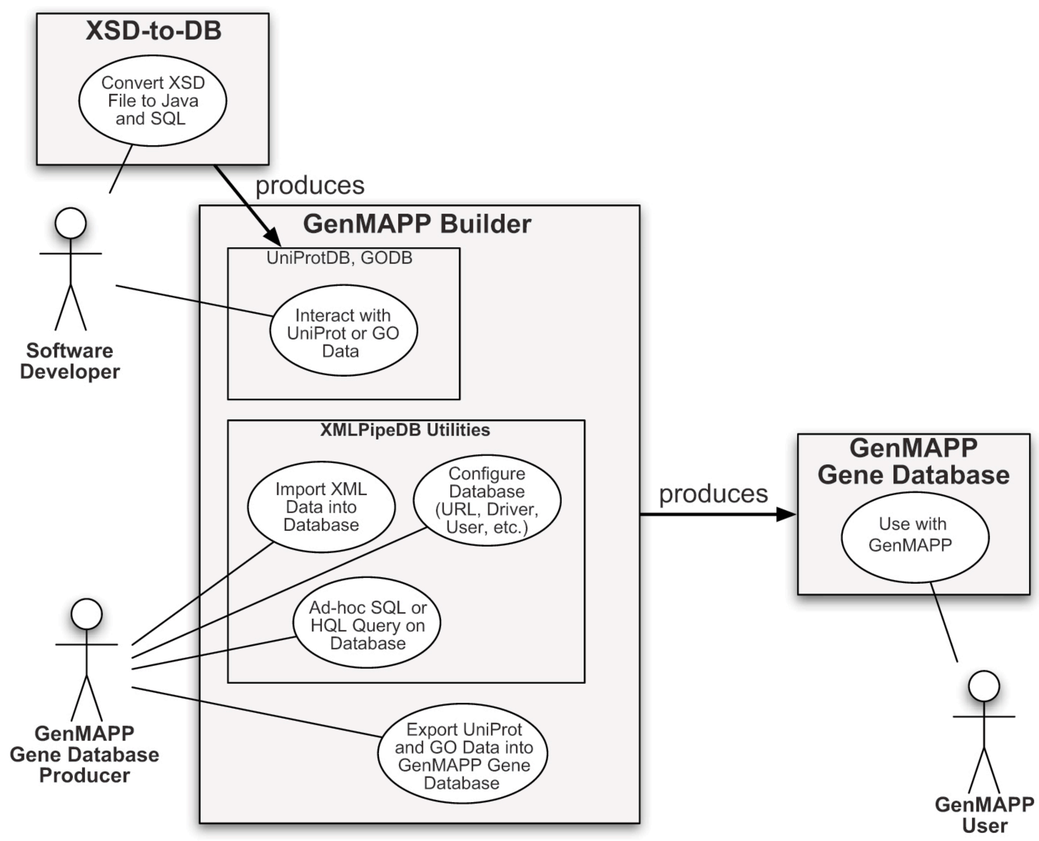
- **Microarrays use different gene ID systems for annotation; users want as much information as possible.**
- **We need to capture and reliably relate gene data from different sources and keep the data updated.**
- **Current GenMAPP Gene Databases are built from Ensembl as the main data source.**
  limited to mostly animal species (human, dog, cow, mouse, rat, chicken, zebrafish, fruit fly, worm, yeast)

  sensitive to changes in flat file formats

Thus, the motivation for XMLPipeDB was:

- **To create GenMAPP Gene Databases for other species (bacteria/plants) using UniProt as the main data source**
- **To be robust to changes in source file formats**
- **To use XML sources wherever possible**
- **To take advantage of existing open source tools**
- **To limit the manual manipulation of the data**

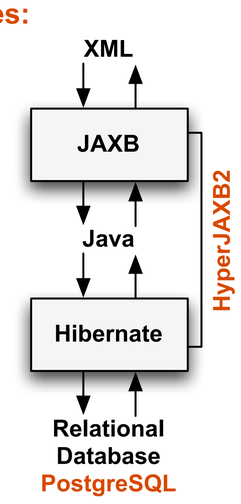
XMLPipeDB Use Case Diagram



XMLPipeDB Implementation

1. XSD-to-DB produces:

- Java source code
- SQL DDL file
- Hibernate mappings
- Apache Ant build.xml

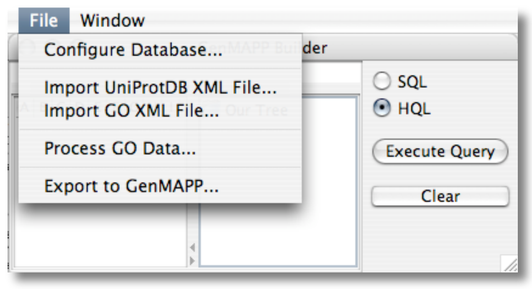


2. UniProtDB and GODB Required Nominal Post-Processing

- **Naming: XSD or DTD definitions might use names that are SQL reserved words and thus cannot be used as table or attribute names, for example:**
  - ➡ *In UniProtDB, "end" was renamed to "endPosition"*
  - ➡ *In GO, "to" was renamed to "to_"*
- **Datatypes: Some XSD datatypes are not easily supported in SQL, for example:**
  - ➡ *In UniProtDB, the definition for citationType was changed from month/year to string*
  - ➡ *Some definitions were changed from SQL varchar(255) to varchar (unspecified length)*

3. GenMAPP Builder Uses the UniProtDB, GODB, and XMLPipeDB Utilities Libraries

- **External to GenMAPP Builder, the user must install the database server and load the GenMAPP Builder schema file generated by XSD-to-DB**
- **Within GenMAPP Builder, perform each function in the order shown by the *File* menu**



Availability

Gene Database for *Escherichia coli* K12
http://xmlpipedb.cs.lmu.edu

XMLPipeDB Programs and Source Code
http://sourceforge.net/projects/xmlpipedb

XMLPipeDB is available under the GNU Lesser General Public License.

Data Sources

XML UniProt Proteome Sets and GOA files from the Integr8 resource
http://www.ebi.ac.uk/integr8/

Gene Ontology OBO XML
http://www.godatabase.org/dev/database/

Contact Us

Kam D. Dahlquist
http://myweb.lmu.edu/kdahqui
kdahlquist@lmu.edu

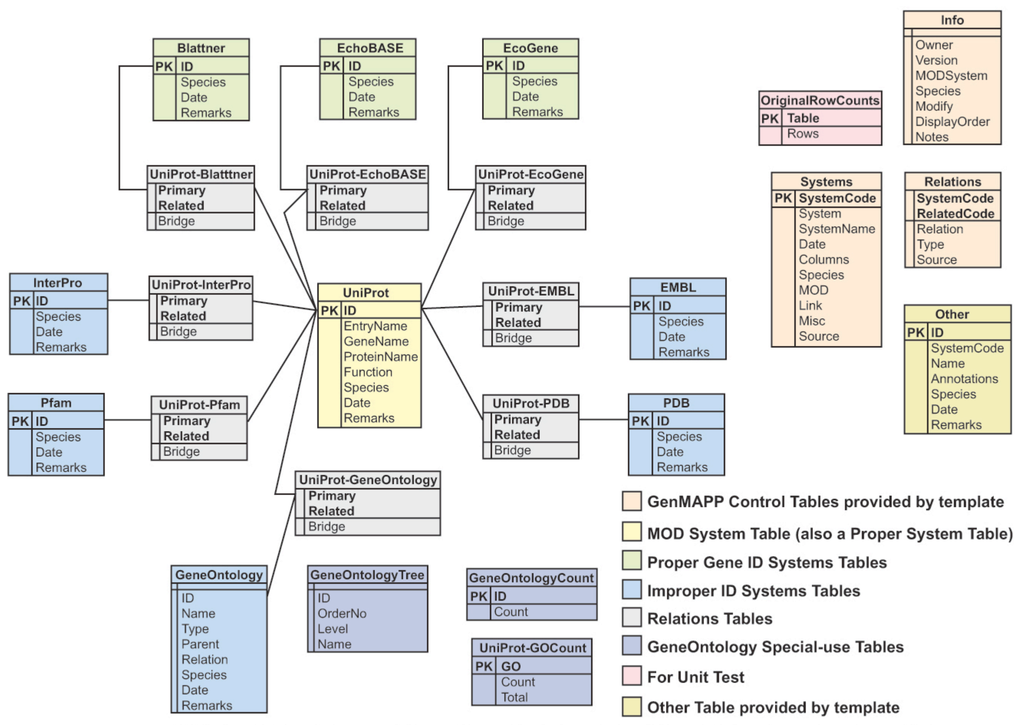
John David N. Dionisio
http://myweb.lmu.edu/dondi
dondi@lmu.edu

Acknowledgments

GenMAPP.org: especially Bruce R. Conklin, Scott W. Doniger, Kristina Hanspers, Steven C. Lawlor, and Alexander Pico

At LMU: Wesley T. Citti, Caskey L. Dickson, Ryan Nakamoto

Schema Diagram of Escherichia coli K12 Gene Database Created by GenMAPP Builder



NOTE: Some Relations tables are not shown. All possible pairwise Relations tables exist between Proper ID systems and between Proper and Improper ID systems, but not between Improper ID systems (i.e., Proper-Proper, Proper-Improper, but NOT Improper-Improper).