

Kam D. Dahlquist

Department of Biology

John David N. Dionisio

Department of Electrical Engineering
& Computer Science

Loyola Marymount University

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

BOSC
August 5, 2006

LMU|LA
Loyola Marymount
University

Outline

- **Process**
 - interdisciplinary collaboration
 - open source pedagogy
- **Motivation**
 - GenMAPP
 - Project requirements
- **XMLPipeDB Implementation**
 - XSD-to-DB
 - UniProtDB and GODB
 - XMLPipeDB Utilities
 - GenMAPP Builder
- **Future Directions**

CMSI 698: Special Studies in Bioinformatics

- **Team-taught by a biologist and a computer scientist**
- **Enrollment in Spring 2006:**
 - **eight students from Master's degree program in Computer Science**
 - **several coming from aerospace industry**
 - **none with more than college-level introductory biology**
- **Project-based class began development of XMLPipeDB**
- **XMLPipeDB development continued by four students in summer session course entitled Open Source Software Development Workshop**
- **Both courses used the open source curricular framework embraced by the Computer Science Department**

Recourse: An Open Source Culture in the Undergraduate Computer Science Curriculum

<http://recourse.cs.lmu.edu/>

Motivation: the disconnect between undergraduate computer science training and expectations/skill sets required in industry

Undergraduate Training	Industry Expectation
Work alone	Work in a team
“Toy” programs and algorithms	Large, modular project
Throwaway code	Code longevity (for better or worse)

Open Source Teaching Framework

Source Code:

- All code resides in a centralized, public repository
- As much as possible, everyone's code is visible to everyone else for code review or team fixing
- No code is thrown away, it remains available to future "generations"

Quality & Community:

- Documentation, inline and online
- Automated tests
- Constructive code review, beyond "does it work?"
- Long-term projects release early, release often
- Form collaborative communities among faculty, students, classes, and projects

“CourseForge”

A Hardware + Software Infrastructure for Supporting the Teaching Framework

- **Certain teaching elements are impractical without some degree of automation**
- **“CourseForge” is currently under development**
- **Derived from open source software, delivered as open source software — the system will interoperate with existing open source tools**
- **Our course used SourceForge.net and later added a Wiki hosted by the Computer Science Department**

XMLPipeDB Project Management: Lessons Learned

- **Students on the project had varying levels of maturity, knowledge, and skill coming into the project**
 - some naturally took on a leadership role
 - some hung back or did the minimum required to get by
- **Needed to increase communication and sense of team**
 - students preferred to interact with faculty for questions, rather than each other
 - bug trackers and developer's forum used only sporadically
 - implemented weekly reports on Wiki to increase accountability
- **[SourceForge servers were frequently down during class]**
- **6 months from conception to product**
- **Even the weakest student contributed useable code**

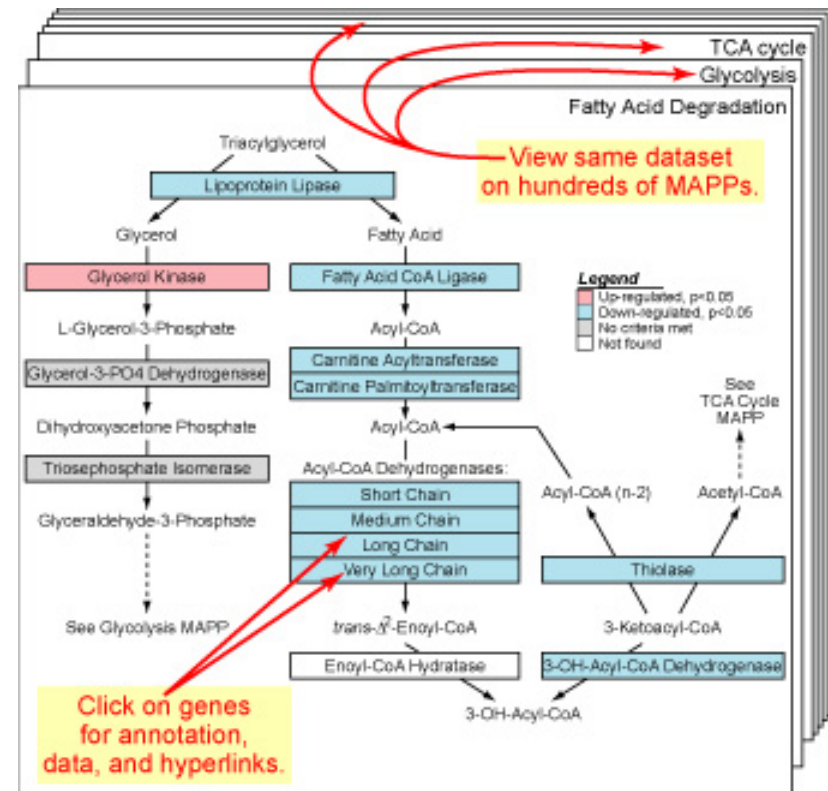
Outline

- **Process**
 - interdisciplinary collaboration
 - open source pedagogy
- **Motivation**
 - **GenMAPP**
 - **Project requirements**
- **XMLPipeDB Implementation**
 - XSD-to-DB
 - UniProtDB and GODB
 - XMLPipeDB Utilities
 - GenMAPP Builder
- **Future Directions**

How GenMAPP Works

<http://www.GenMAPP.org>

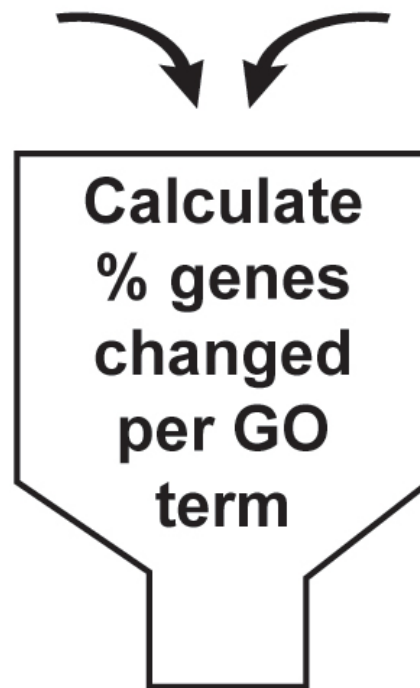
- Graphics tools make MAPPs that store gene IDs and vector coordinates for all graphical objects
- Separate Expression Dataset files store data and color-coding instructions
- Gene Databases store IDs, annotation, and hyperlinks to public gene and protein databases
- Stand-alone program implemented in Visual Basic, accessory files are Microsoft Access databases



MAPPFinder Determines Which GO Terms Are Overrepresented in a GenMAPP Expression Dataset

Hundreds of genes meeting the criterion for a meaningful gene expression change

Gene Ontology process, component, and function terms



List of Gene Ontology terms ranked 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.**
- **Gene Database design is data-driven; it tells GenMAPP what gene ID systems and relationships are present.**
- **Current GenMAPP Gene Databases are built from Ensembl as the main data source.**
 - limited to (mostly) animal species
 - sensitive to changes in flat file formats

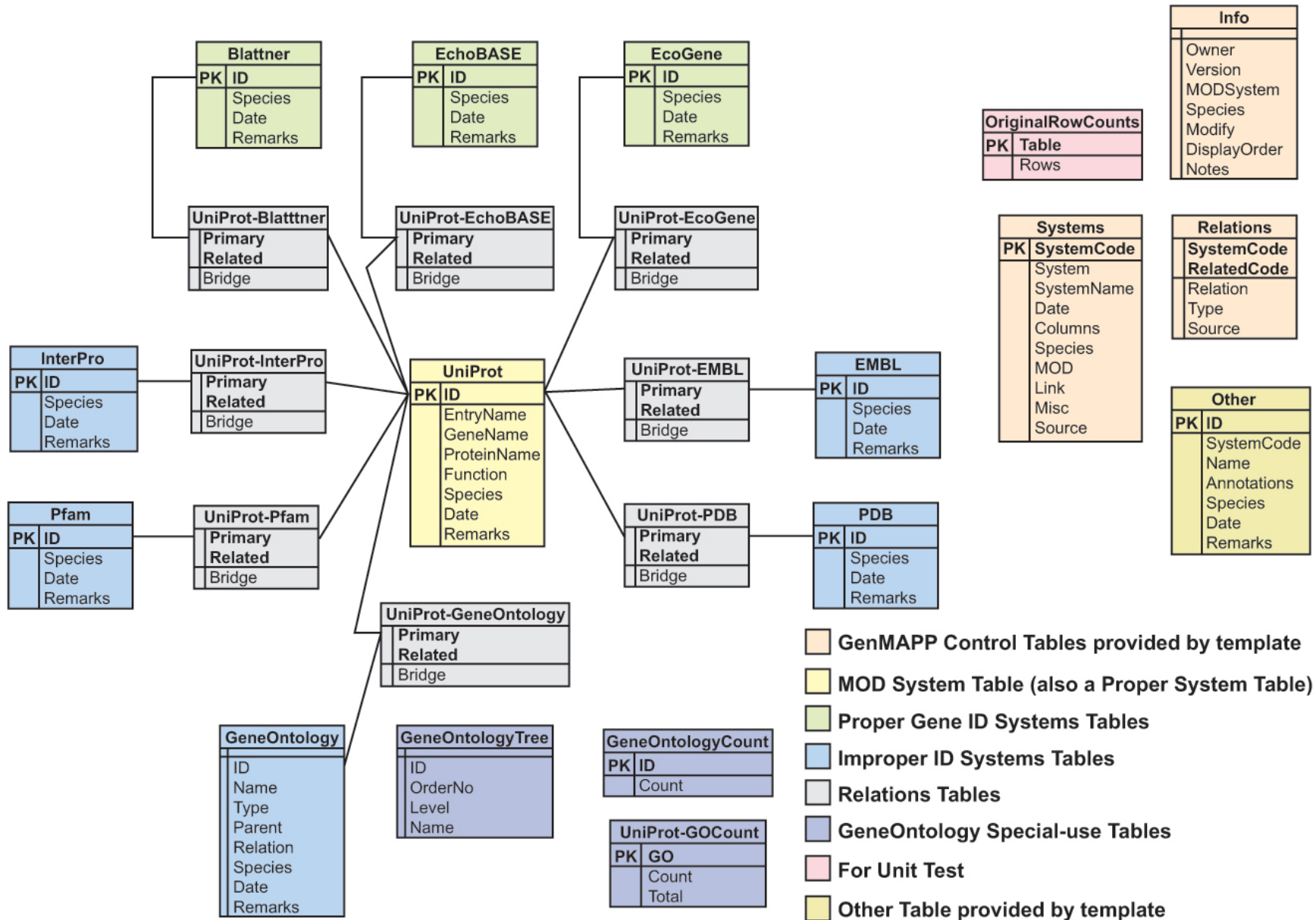
XMLPipeDB: A Reusable, Open Source Tool Chain for Building Relational Databases from XML Sources

Requirements:

- **to create 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**

First task was to build a GenMAPP Gene Database for *Escherichia coli* K12

GenMAPP Gene Database Schema for Escherichia coli K12



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).

Data Sources Required for a “Minimal” GenMAPP Gene Database

UniProt

- **UniProt complete proteome sets for many species are made available as XML downloads by the Integr8 resource**

Gene Ontology

- **OBO XML format**

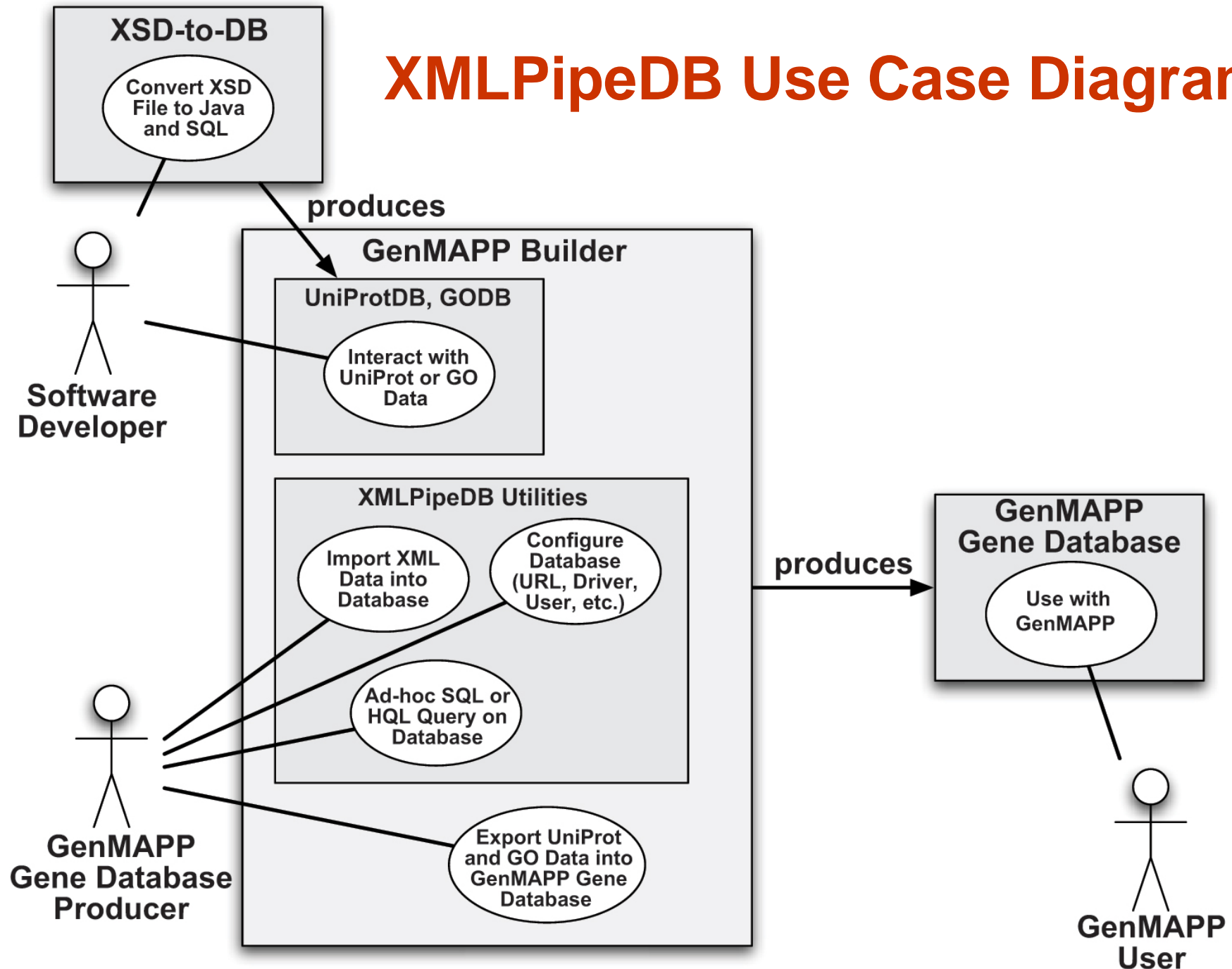
UniProt to GO associations

- **GOA downloads also available at Integr8**

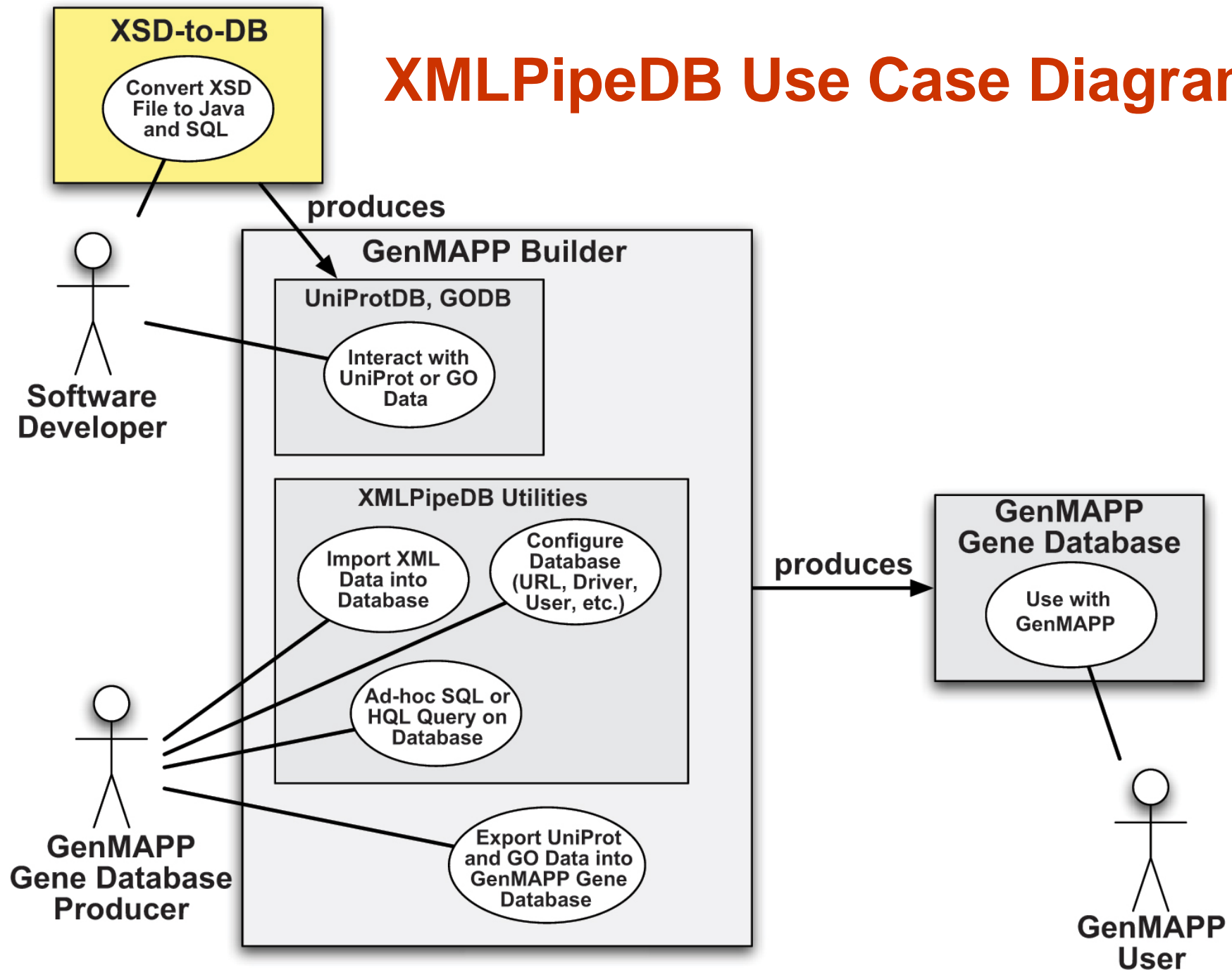
Outline

- **Process**
 - interdisciplinary collaboration
 - open source pedagogy
- **Motivation**
 - GenMAPP
 - Project requirements
- **XMLPipeDB Implementation**
 - XSD-to-DB
 - UniProtDB and GODB
 - XMLPipeDB Utilities
 - GenMAPP Builder
- **Future Directions**

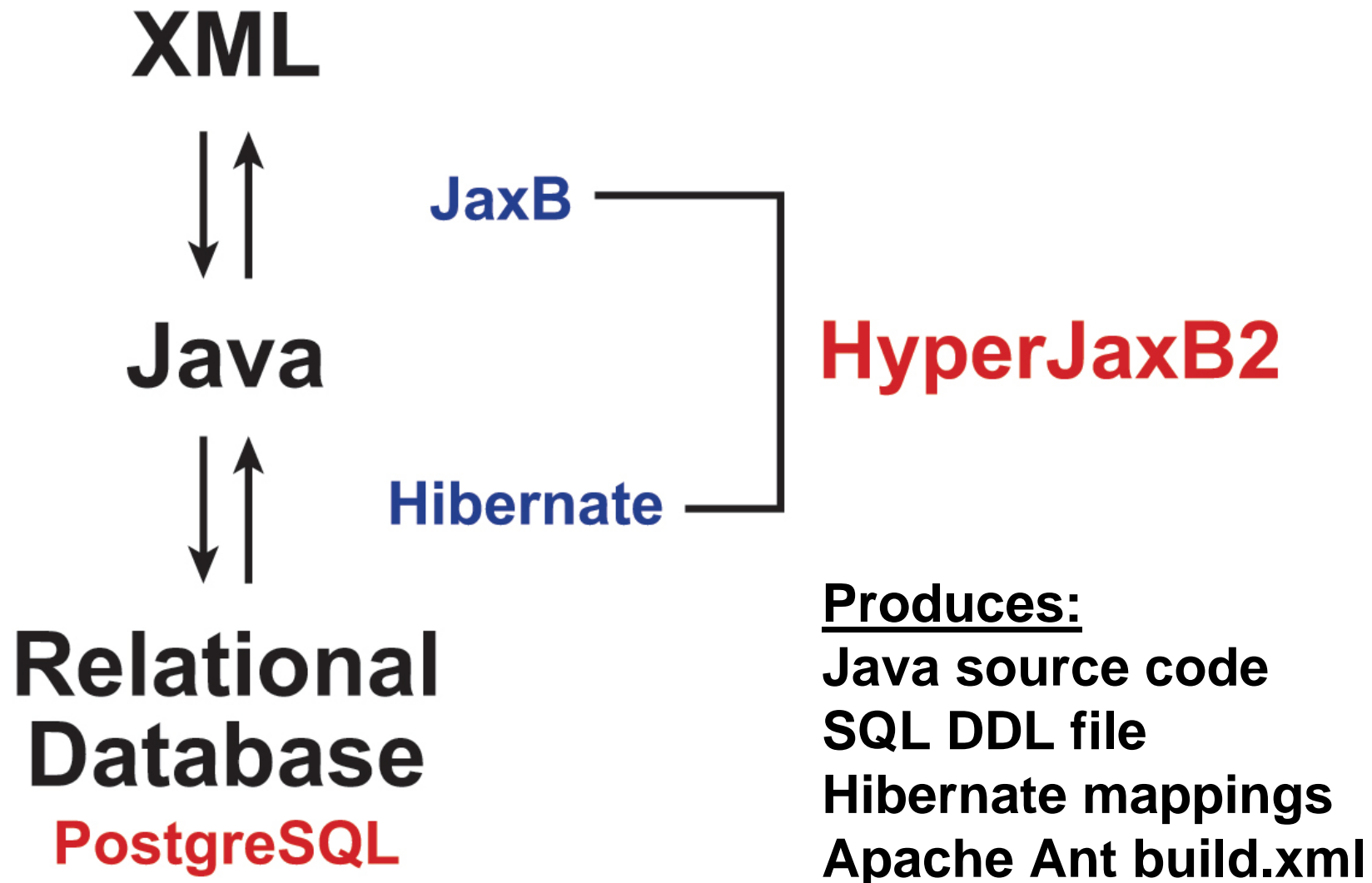
XMLPipeDB Use Case Diagram



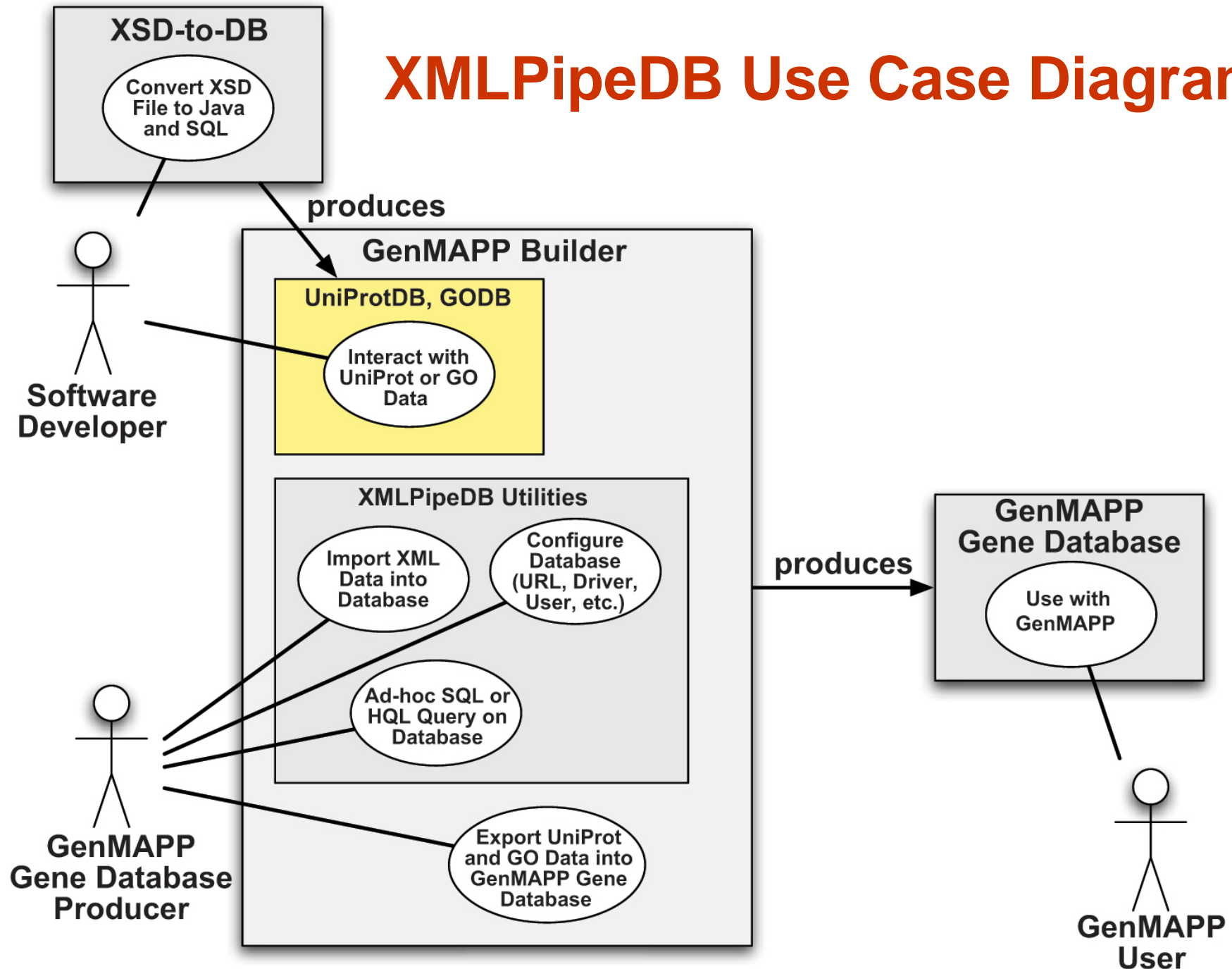
XMLPipeDB Use Case Diagram



XSD-to-DB Stands on the Shoulders of other Open Source Tools



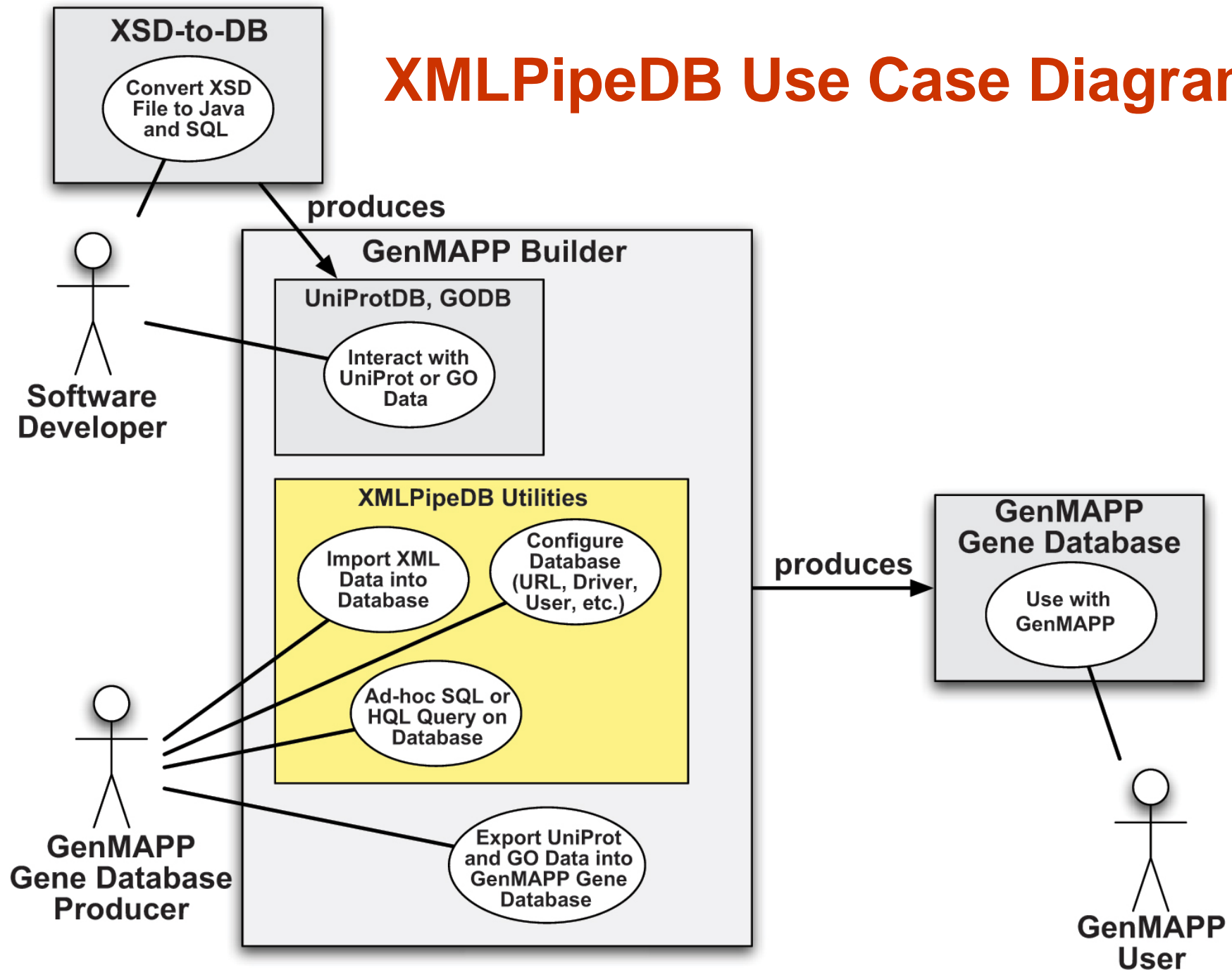
XMLPipeDB Use Case Diagram



UniProtDB and GODB Required Only 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
 - In UniProtDB, “end” was renamed to “endPosition”
 - In GODB, “to” was renamed to “to_”
- **Datatypes:** Some XSD datatypes are not easily supported in SQL
 - 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)

XMLPipeDB Use Case Diagram

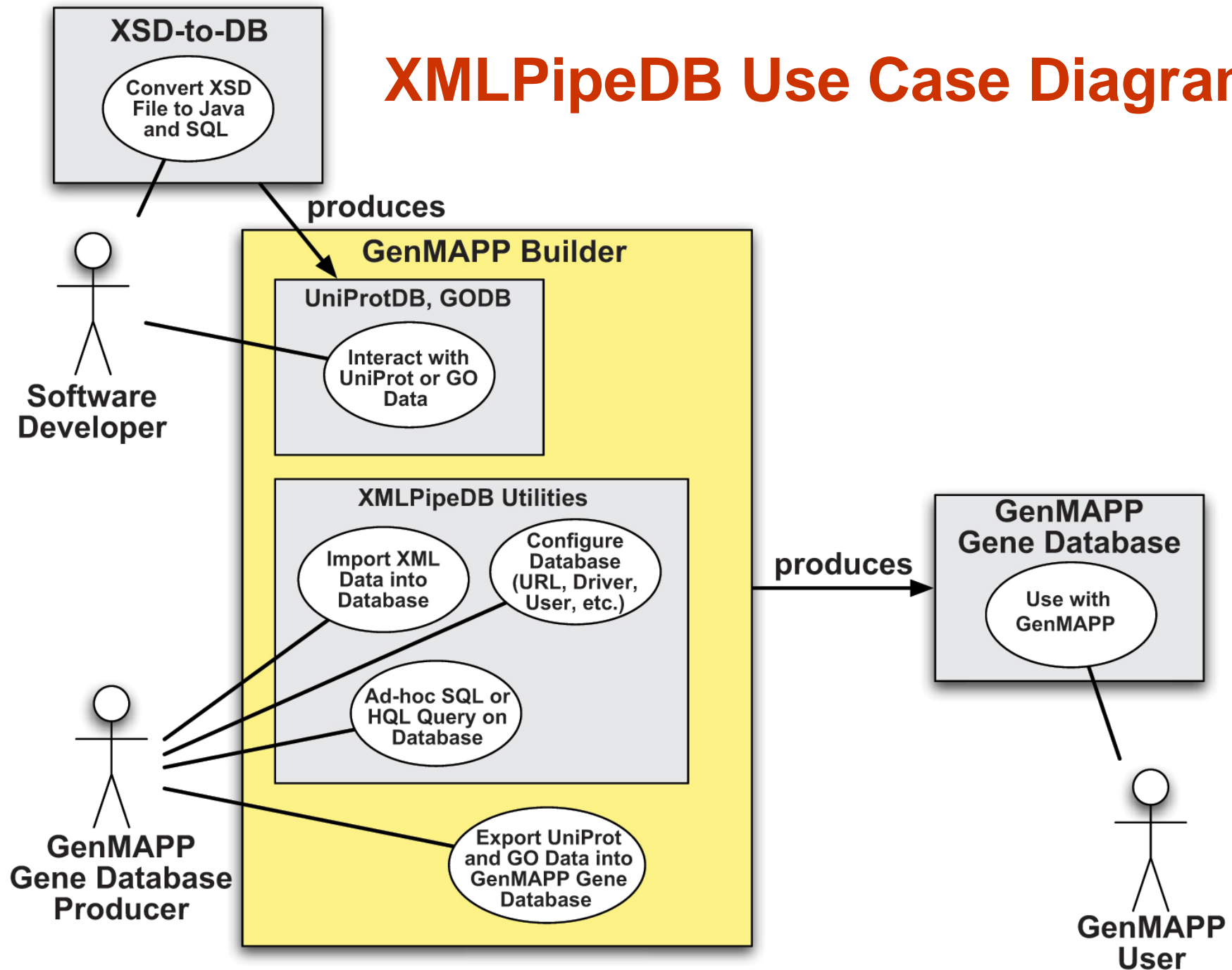


“Rule of Three”

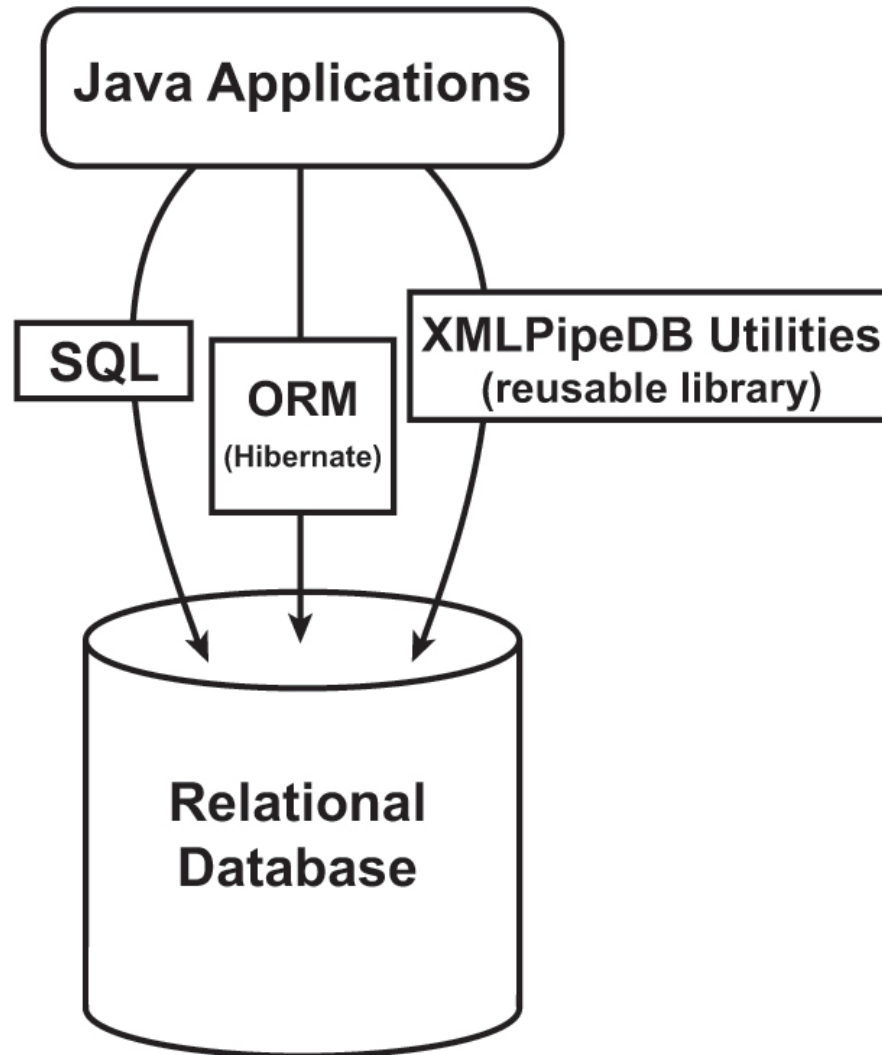
XMLPipeDB Utilities Library is a Suite of Java Classes that Provide Functions Common to Most XMLPipeDB Database Applications

- **Loading of XML files into Java objects**
- **Saving XML-derived Java objects to a relational database**
- **Rudimentary query and retrieval of Java objects from the relational database**
 - HQL (Hibernate Query Language), SQL query
 - object browser that shows results of query
- **Configuring a client application to communicate with a relational database**

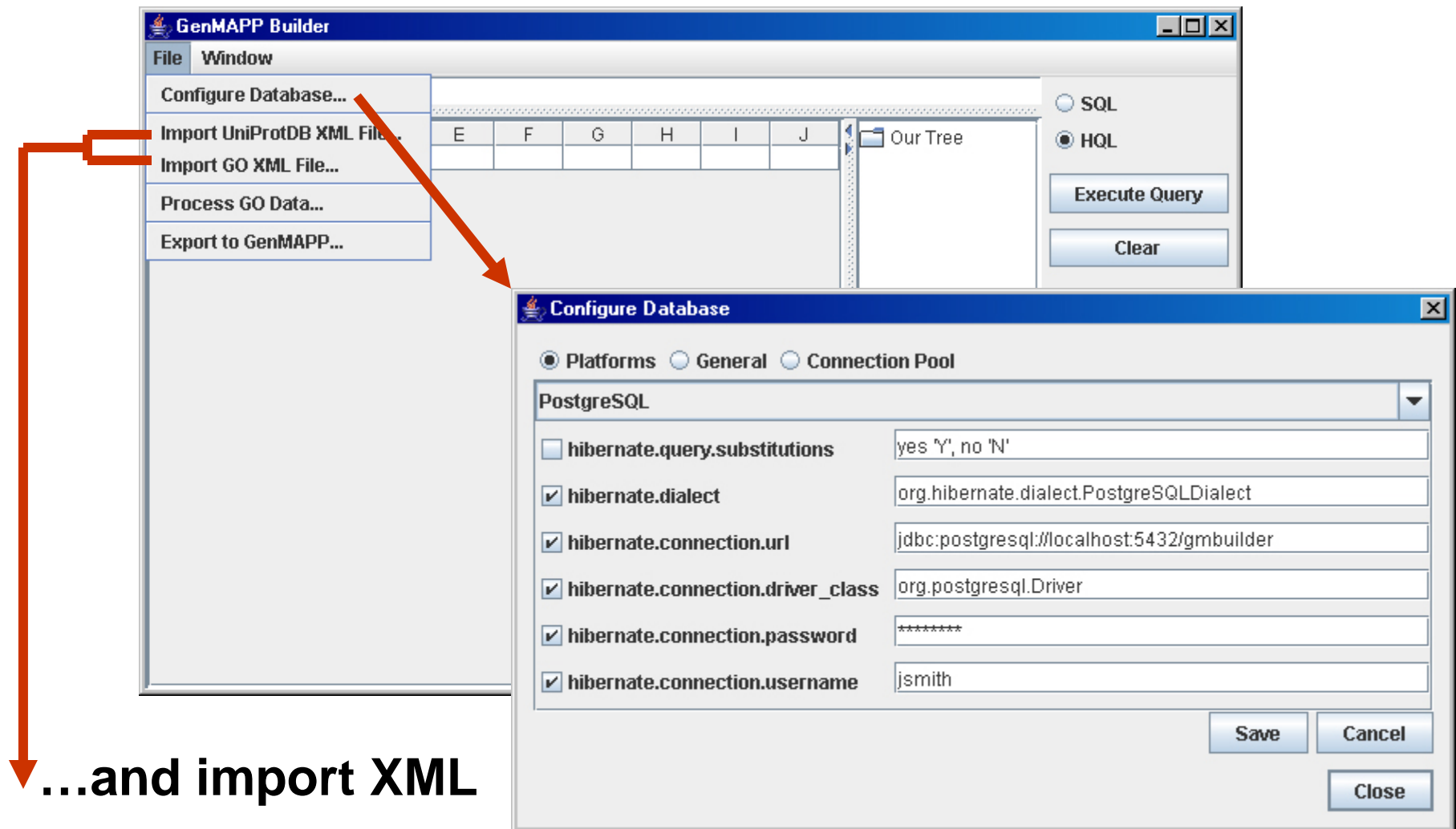
XMLPipeDB Use Case Diagram



GenMAPP Builder Interacts with PostgreSQL in Three Ways



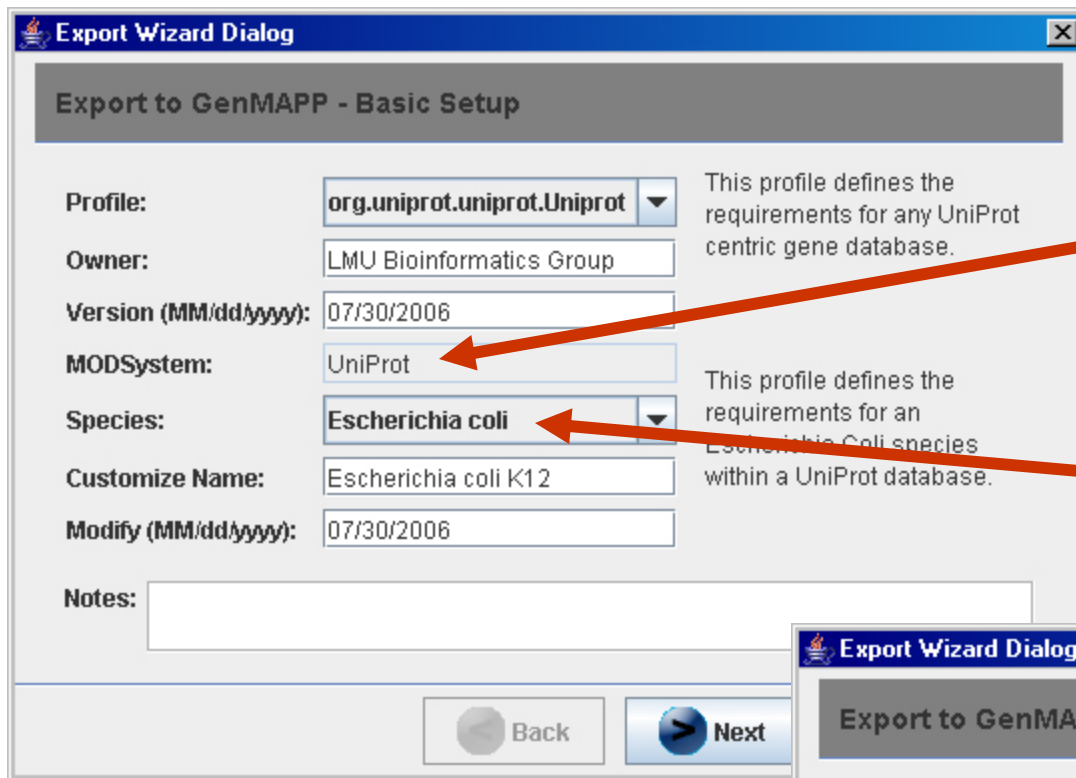
GenMAPP Builder Uses the XMLPipeDB Utilities Library to Configure the PostgreSQL Database



GenMAPP Builder Has Customized Profiles

-- for each Primary Data Source (e.g. UniProt)

-- for each species (e.g. *Escherichia coli*)



Export Wizard Dialog

Export to GenMAPP - Basic Setup

Profile: **org.uniprot.uniprot.Uniprot** This profile defines the requirements for any UniProt centric gene database.

Owner: LMU Bioinformatics Group

Version (MM/dd/yyyy): 07/30/2006

MODSystem: **UniProt** This profile defines the requirements for an Escherichia Coli species within a UniProt database.

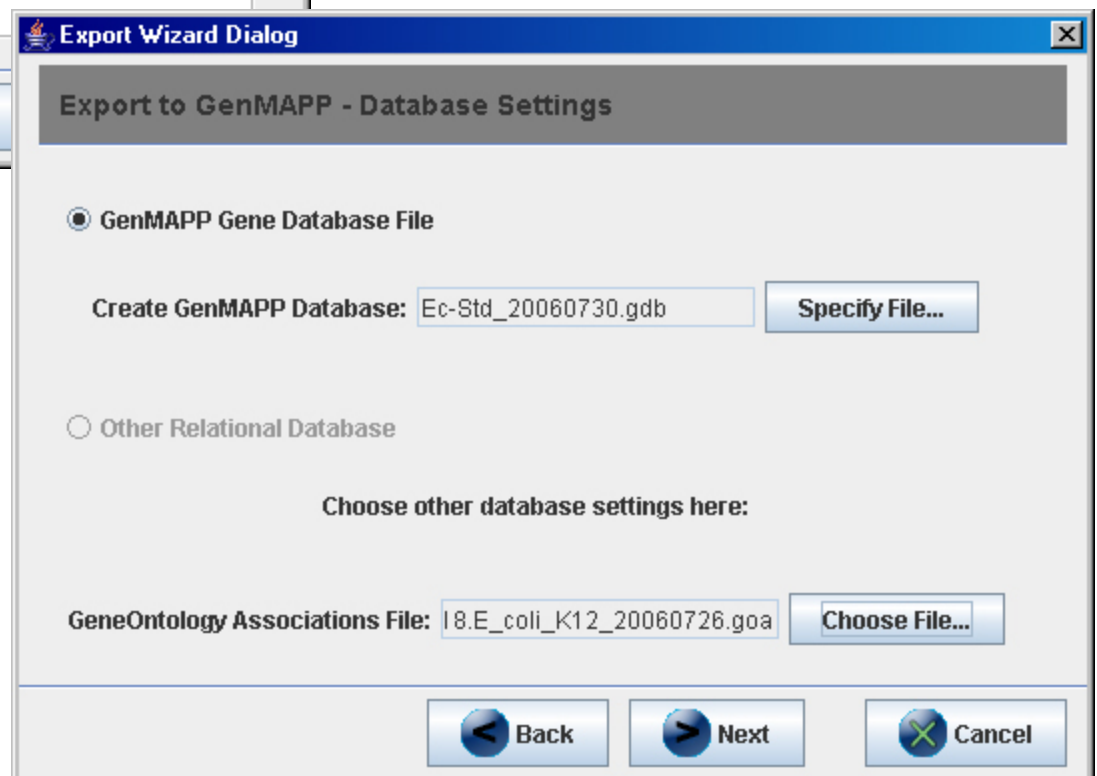
Species: **Escherichia coli**

Customize Name: Escherichia coli K12

Modify (MM/dd/yyyy): 07/30/2006

Notes:

Back Next



Export Wizard Dialog

Export to GenMAPP - Database Settings

☒ GenMAPP Gene Database File

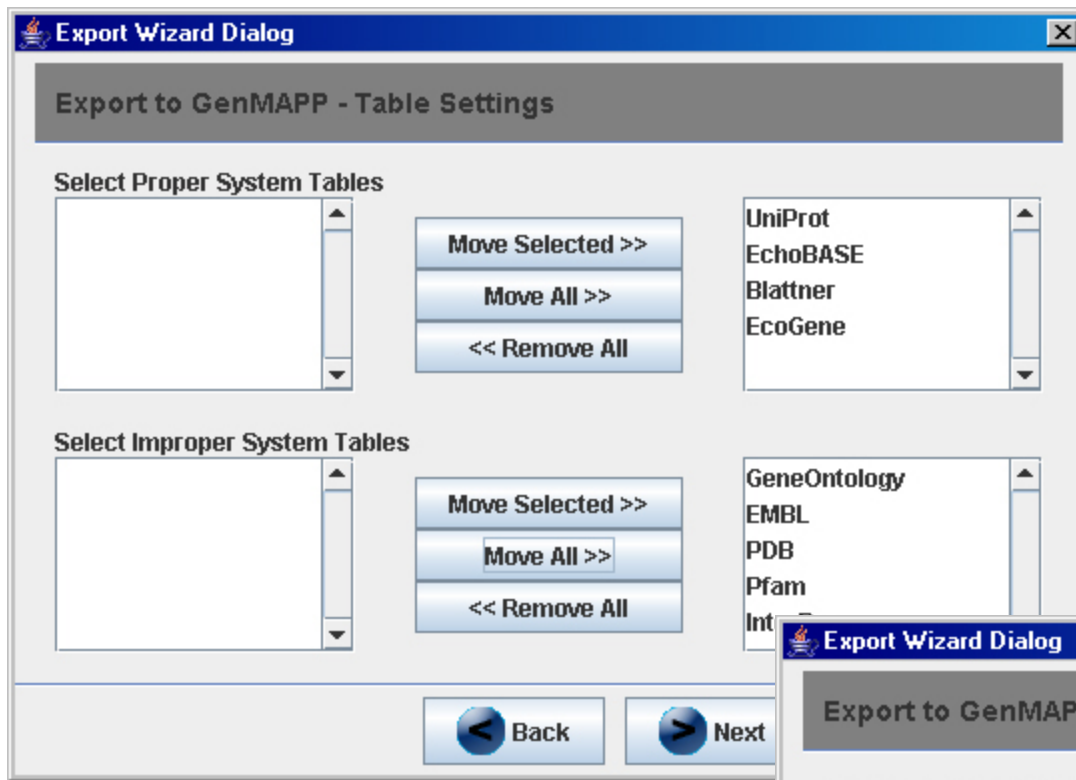
Create GenMAPP Database: Ec-Std_20060730.gdb **Specify File...**

☐ Other Relational Database

Choose other database settings here:

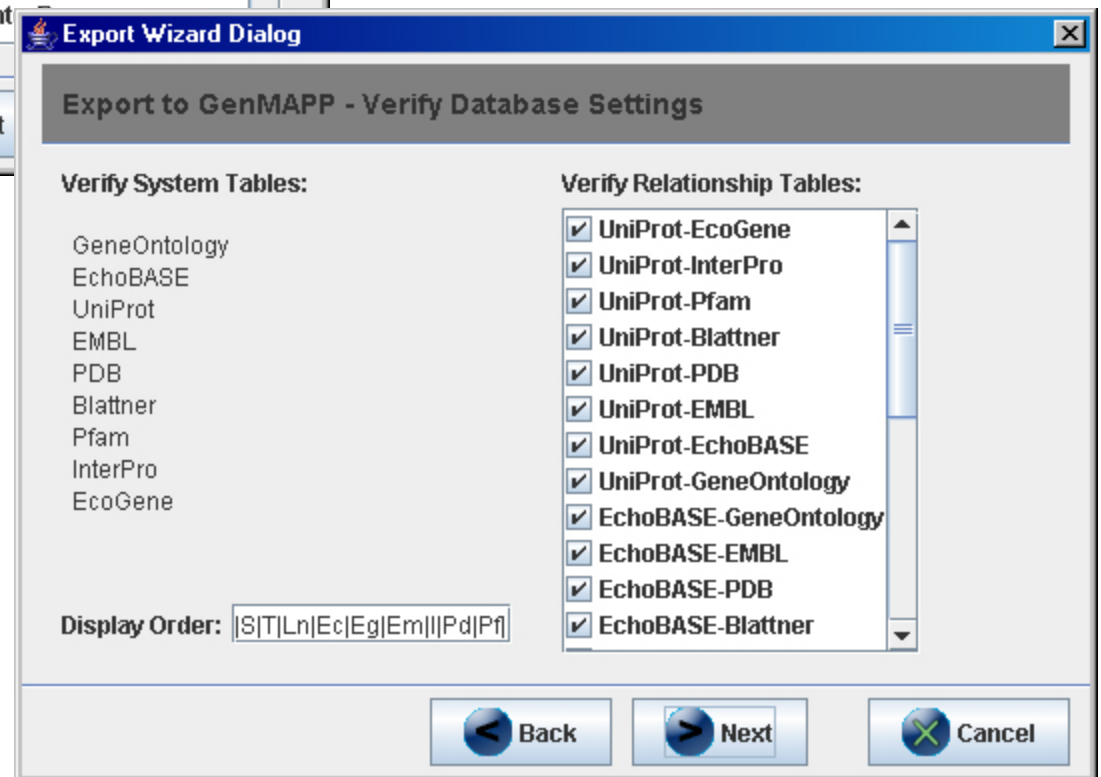
GeneOntology Associations File: I8.E_coli_K12_20060726.goa **Choose File...**

Back Next Cancel



The User Chooses
Which Gene ID Systems
and Relations to Export
to the Gene Database

Building the *E. coli* K12
Gene Database takes
~2 hours



Take-home Messages

- **Used an Open Source paradigm for Master's level course, resulting in useful bioinformatics software**
 - software is NOT perfect, but acceptable for now
 - students will flow in and out of the project
- **GenMAPP Builder can make Gene Databases for any species represented in UniProt**
 - produced a Gene Database for *Escherichia coli* K12
- **XMLPipeDB is a general set of tools that can be **re-used** for other bioinformatics and non-bioinformatics applications**
 - LGPL license
 - *we have not experienced a change to an XSD yet*

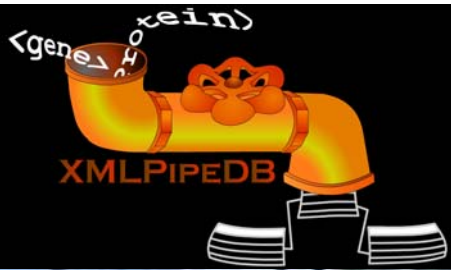
Future Directions for XMLPipeDB

Near Term:

- Clean-up internal design and GUI for GenMAPP Builder
- Produce Gene Databases for additional species
- Add data sources (TIGR CMR, NCBI Gene, Affymetrix)
- Further automate building databases and data integrity checks

Longer Term:

- Use XML sources to build MAPPs for GenMAPP e.g., KEGG-ML, BioPAX
- Applications that we haven't imagined yet



LMU Bioinformatics Group

<http://xmlpipedb.cs.lmu.edu>

LMU|LA
Loyola Marymount
University

Kam D. Dahlquist

<http://myweb.lmu.edu/kdahqui>
kdahlquist@lmu.edu

John David N. Dionisio

<http://myweb.lmu.edu/dondi>
dondi@lmu.edu

ISMB Poster B-36

Special Thanks

GenMAPP.org Development Group

Caskey L. Dickson, Wesley T. Citti

NSF CCLI Program (<http://recourse.cs.lmu.edu>)

XSD-to-DB

Adam Carasso

Jeffrey Nicholas

Scott Spicer

XMLPipeDBUtils

David Hoffman

Babak Naffas

Jeffrey Nicholas

Ryan Nakamoto

UniProtDB

Joe Boyle

Joey Barrett

GODB

Scott Spicer

Roberto Ruiz

GenMAPP Builder

Joey Barrett

Jeffrey Nicholas

Scott Spicer