

InterMine - open source data warehouse and query interface

Richard Smith^{1,2}, Sergio Contrino¹, Hilde Janssens¹, Jakub Kulaviak¹, Rachel Lyne¹, Kim Rutherford¹, Julie Sullivan¹, Dan Tomlinson¹, Matthew Wakeling¹, Xavier Watkins¹ and Gos Micklem¹.

¹Systems Biology Centre, University of Cambridge, UK. info@flymine.org

²corresponding author: richard@flymine.org

InterMine (www.intermine.org) is an open-source system for building query-optimised data warehouses. It supports data integration from standard biological formats and makes it easy to add your own data. A sophisticated web application provides flexible query access for any data model. Queries can be run via RESTful web services with results returned as HTML to embed in other web pages – promoting data re-use and mash-ups.

InterMine was developed to enable FlyMine (www.flymine.org) and is now used in other projects, including the \$57m modENCODE project. InterMine is written in Java, all code is freely available under the LGPL license (see www.intermine.org/wiki/SVNCheckout).

Data Integration

InterMine makes it easy to integrate multiple data sources into a central data warehouse. It has a core data model based on the sequence ontology and supports many biological data formats. The object-based data model is defined as XML from which a database schema and Java classes are automatically generated. It is simple to extend the data model and integrate your own data. Supported formats include GFF3, FASTA, Chado, GO gene association, UniProt XML, PSI XML (protein interactions), PDB XML and Ensembl.

Web Application

A web application provides flexible query access to the data warehouse. The interface allows users to create custom queries, use template queries - web forms to run 'canned' queries, upload and operate on lists of data and analyse lists with interactive graphical or statistical 'widgets'. Queries and lists can be saved in a MyMine account.

The web application works 'out of the box' with any data model and can be heavily customised. Much of the presentation can be controlled by non-programmers: an admin user can publish new template queries and lists at any time and can change report pages by tagging.

Web Services

Custom queries, template queries and lists are available via RESTful web services. Results can be returned as XML, tab or comma separated formats. It is also possible to fetch HTML to include features of InterMine in other web sites.

All template queries have an 'embed this query' link which provides a URL to include results in report pages of another web interface. This feature is designed to promote the 'mash-up' model – other sites can fetch information related to the data they present without the need for local data integration. Results will change as the InterMine database is updated, avoiding maintenance overhead or the inclusion of stale data.

InterMine and FlyMine are actively developed by a team of seven people at the University of Cambridge, UK. The project is funded by the Wellcome Trust.