# InterMine

## Open source data warehouse and query interface

www.intermine.org
info@intermine.org

*Richard Smith, Sergio Contrino, Hilde Janssens, Jakub Kulaviak, Rachel Lyne, Kim Rutherford, Julie Sullivan, Dan Tomlinson, Matthew Wakeling, Xavier Watkins and Gos Micklem.*
*University of Cambridge, UK.*

## Introduction

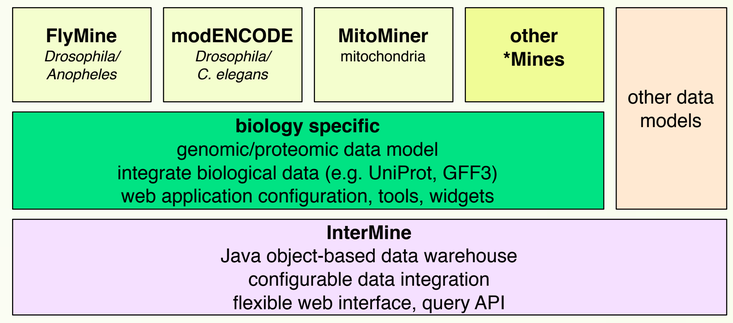
InterMine (www.intermine.org) is an open source system for building query-optimised data warehouses. It supports data integration from standard biological formats and makes it easy to add your own data. A sophisticated web application provides flexible query access for any data model. Queries can also be built with an API and executed via web services.

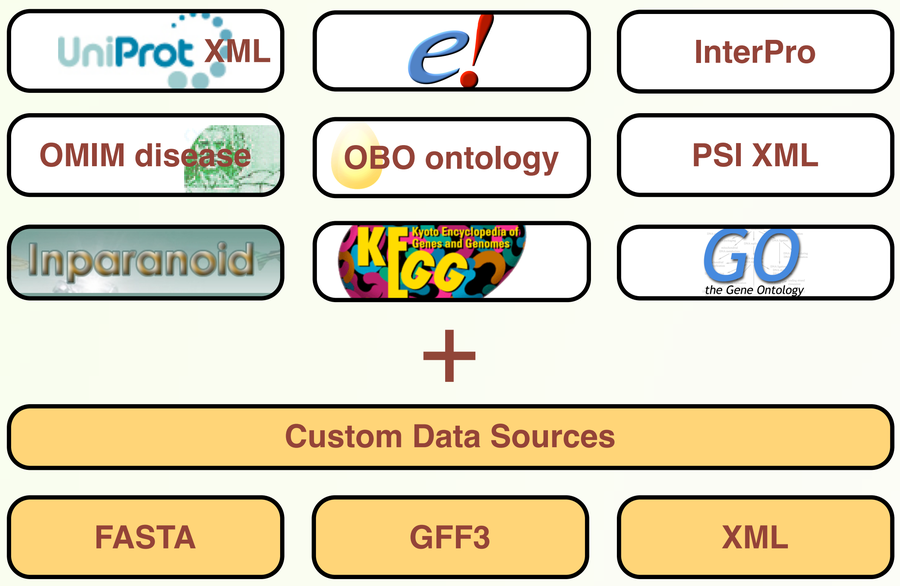
InterMine was developed to enable FlyMine (www.flymine.org) and is now used in several other projects worldwide, including the $57m modENCODE project (www.modencode.org).

InterMine is written in Java, all code is freely available under the LGPL license.

## Software Overview

InterMine is used in several projects worldwide.




Configurable data integration

## InterMine Data Warehouse

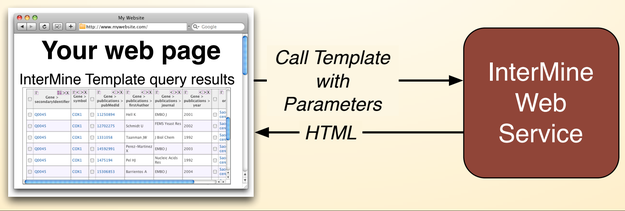
InterMine makes it easy to integrate multiple data sources into a central data warehouse. It has a core data model based on the sequence ontology and supports many biological data formats. The object-based data model is defined as XML from which a database schema and Java classes are automatically generated. It is simple to add custom sources to extend the data model and integrate your own data. Java and Perl APIs are provided for converting data.

Supported formats include GFF3, FASTA, Chado, GO gene association, UniProt XML, PSI XML, PDB XML and Ensembl. A central configuration file defines the sources to load into an InterMine instance and specific organisms and data files to include.

## 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. Queries can exported as XML from the web application for use with the web service or created using client APIs in Java or Perl (Summer 2008).

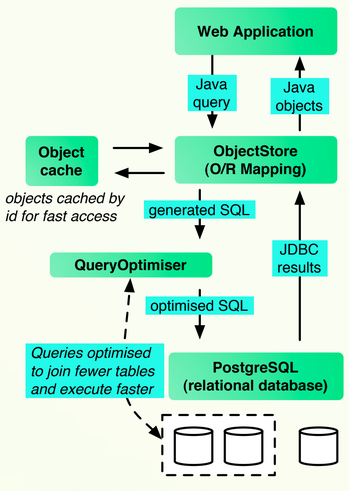
It is also possible to fetch HTML to include query or widget results directly in other web pages. All template queries have an 'embed this query' link which provides a URL to include results in an existing report page, this allows external sites to present data without the need for local data integration.




## Architecture

InterMine is written in Java and built around a custom **object/relational mapping** tool optimised for read-only performance. The data model is object-based and defined by an XML file from which a database schema and Java classes are automatically generated.

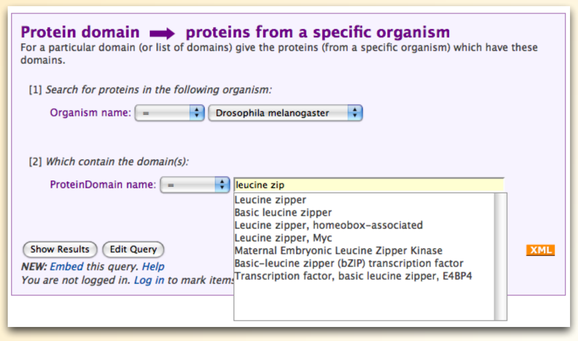
A **cost-based query optimiser** attempts to re-write incoming queries to make use of pre-computed tables to join fewer tables and run faster. New pre-computed tables can be added at any time to adapt to actual usage. In contrast to traditional data warehouses performance optimisation is thus separated from schema design.



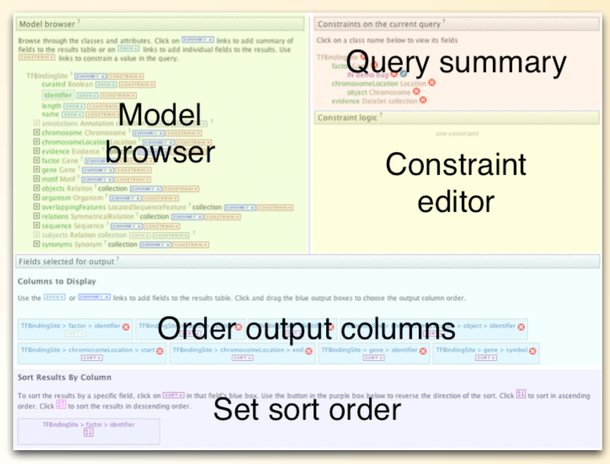

*NOTE: ObjectStore is not related to commercial product of the same name*

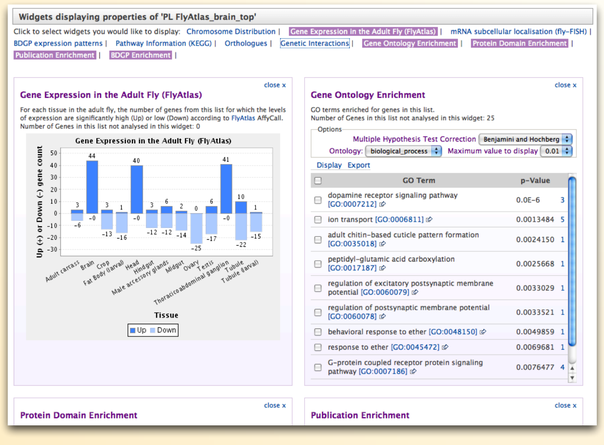
## Web Application

A web application works 'out of the box' for any data model to provide flexible query access to the data warehouse, designed for functionality beyond looking up an identifier and viewing a report. The application is highly customisable through configuration and development of new data displayers or integration of tools.




A **QueryBuilder** (right) allows non-programmers to create custom queries. Any query can be turned into a re-usable **template query** (above) providing a simple form with a description and editable constraints.




**Lists** of any type (e.g. genes, proteins) can be uploaded or created from query results. Lists can be used in any query or template in place of a single identifier. A **list analysis** page (left) summarises properties of a list through interactive **widgets**, for example graphs of gene expression or statistical enrichment of GO terms or protein domains. A framework is provided for creating new widgets.

A **MyMine** account allows users to save lists and queries between sessions.

An important feature is that much of the presentation can be controlled by non-programmers: an **admin user** can create and publish new template queries and lists at any time and can adapt report pages by applying tags.

wellcome trust

FlyMine is funded by the Wellcome Trust (grant no. 067205), awarded to M. Ashburner, G. Micklem, S. Russell, K. Lilley, and K. Mizuguchi.

Systems Biology Centre, University of Cambridge, Tennis Court Road, Cambridge CB2 1QR, UK
Tel: +44 1223 760262 Email: info@flymine.org