The MO-LD project: Enhancing the FAIRness of Yeast and other model organism data

Maxime Déraspe\textsuperscript{1,2}, Gail Binkley\textsuperscript{3}, Kalpana Karra\textsuperscript{3}, Gos Micklem\textsuperscript{4}, Julie Sullivan\textsuperscript{4}, Michael J. Cherry\textsuperscript{4}, and Michel Dumontier\textsuperscript{1}

\textsuperscript{1}Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, United States; \textsuperscript{2}Department of Molecular Medicine, Université Laval, Quebec, Canada; \textsuperscript{3}Department of Genetics, Stanford University, Stanford, United States; \textsuperscript{4}Department of Genetics, University of Cambridge, Cambridge, United Kingdom

Abstract

Model organisms such as budding yeast provide a common platform to interrogate and understand cellular and physiological processes. Knowledge about model organisms, whether generated during the course of scientific investigations, or extracted from published articles, are integrated and made available by model organism databases (MODs) such as the Saccharomyces Genome Database (SGD). SGD and many MODs use InterMine, a system for integrating, analysing, and republishing biological data from multiple sources that also enables data-driven bioinformatic analyses through a web user interface and programmatic web services.

Here, we developed a cloud-ready dockerized platform that uses Semantic Web technologies to transform and make available model organism data in a manner that makes it easier to discover, explore, and query. First, we developed a pipeline to extract, transform, and load a Linked Data representation of the InterMine store. Second, we use Docker to package both software and data for local or remote deployment. Third, we built a lightweight dashboard that packages together existing and SPARQL-aware applications to search, browse, explore, and query the InterMine-based data. Our work extends the InterMine platform, and supports new query functionality across InterMine installations and the network of open Linked Data.

Specific Aims

1. To improve interoperability of MOD data by publishing Linked Data
2. To enable and demonstrate federated queries between MOD data and the network of Linked Data
3. To package our software and data for easier local and cloud-based deployment

Availability

Website: http://mo-ld.org

Tutorial: https://github.com/mo-ld/mold-dock

Docker (container engine) to build and deploy the MOLD infrastructure

https://hub.docker.com/u/mold

Microservices architecture for reusability and extensibility:

Web application, API and Virtuoos images

• Cloud-Ready - tested on Amazon EC2

The Intermine RDFizer queries the InterMine API with XML description of the object model. It then upconverts the tabular results into triples (RDF). It merges resources with the same primary keys and links data with external datasets. Finally, the data is loaded into a RDF-compatible database (triple store).

References

Déraspe, M; Karra, K; Binkley, G; Micklem, G; Corbel, M; Cherry, M; Dumontier, M: Semantic research platform for Model Organisms. Proceedings of the 1st International Workshop on Biomedical Data Integration and Discovery (BMDI 2016). Kobe, Japan.

Acknowledgements

Funding for this work was provided as an administrative supplement to the Genomic Database for the Yeast (http://mo-ld.org).