

Semantic Research Platform for Model Organism Data

Maxime Déraspe^{1,5}, Kalpana Karra⁴, Gail Binkley⁴, Julie Sullivan^{2,3}, Gos Micklem^{2,3}, Jacques Corbeil¹, J. Michael Cherry⁴, and Michel Dumontier⁵

¹ Department of Molecular Medicine, Université Laval, Québec, Canada

² Cambridge Systems Biology Centre, University of Cambridge, Cambridge, United Kingdom

³ Department of Genetics, University of Cambridge, Cambridge, United Kingdom

⁴ Department of Genetics, Stanford University, Stanford, United States

⁵ Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, United States

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 uses InterMine to enable powerful, data-driven bioinformatic analyses and most of the other MODs also expose their data through InterMine so providing a standard platform for MOD data exploration and mining. However bioinformatic analyses also require access to a significantly broader set of biomedical data, which today can be found in structured form in the emerging network of Linked Open Data (LOD). The MODs have expended substantial effort over many years on human curation of the literature and if these gold-standard data alongside other MOD data could be provisioned as FAIR (Findable, Accessible, Interoperable, and Reusable), then scientists could leverage a greater amount of interoperable data in knowledge discovery.

Keywords: linked data, semantic web platform, model organisms, biomedical research

1 Introduction

Model organisms are a set of reference species that the research community use to study basic biology, biodiversity, and help us understand human biology. From fundamental to applied sciences, these guinea-pigs have proved their usefulness in building systems biology, understanding complex phenotypes, uncovering novel biological mechanisms, discovering new drug targets, testing new drugs and studying human diseases. Knowledge about model organisms is captured in Model Organism Databases (MODs) and includes ontologies such as the Gene Ontology (GO) [1], Sequence Ontology (SO) [2], Human Phenotype

Ontology (HPO) [3] and Disease Ontology (DO) [4]. Through the InterMOD project [5] the various MODs are working towards standardizing access to their data through adoption of the InterMine platform [6], a popular system, with over 25 available endpoints. It covers the most widely studied model organisms, such as budding yeast, fruit fly, zebrafish, rat, nematode, mouse and *Arabidopsis* as well as human. Given that MODs rely considerably on open databases and that the biological data provider community (EBI, REACTOME, ENSEMBL, NCBI, DDBJ) has increased its adoption of the Resource Description Framework (RDF), we initiated an effort to provide model organism data as 5-star linked data⁶ so as to integrate these into the wider network of Linked Open Data (LOD). We describe our efforts to develop a novel resource, the Model Organism Linked Database (MOLD⁷), which uses Semantic Web technologies to make the knowledge of six model organisms (budding yeast, fruit fly, zebrafish, rat, mouse, human) available from their respective InterMine endpoints in a FAIR (Findable, Accessible, Interoperable, and Reusable) [7] manner.

2 From MOD to MOLD

In this section, we present the methodologies used to convert the model organism data from InterMine data warehouses into RDF and their integration with other biological LOD.

2.1 RDFization of MOD

InterMine [8] is a model-driven data warehouse system based on PostgreSQL that provides a client API in five programming languages for access to InterMine data. The client API is a graph-based query format that inherits some of its semantics and terminology from SQL. The combination of the API and the object model⁸ allows the user to fetch the content of an InterMine endpoint. We built a script, the InterMine-RDFizer⁹, to make use of these two components to download and process the database content of six MODs: YeastMine [9], ZebrafishMine [10], FlyMine [11], RatMine [12], MouseMine [13] and HumanMine¹⁰. The data flow of the script is illustrated in Figure 1. As it uses the object model specific to each InterMine database, InterMine-RDFizer is flexible enough to be used with any InterMine installation. It can be launched via a command line interface and was used to convert all of the above six MODs. The first step is to query, download, and save all the table content into tab-delimited (TSV) files. There are two different types of tables saved by the script: one that contains information about the resources and another that represents the relations between the tables. The schema of the PostgreSQL database in InterMine is object oriented and loosely

⁶ <https://www.w3.org/DesignIssues/LinkedData.html>

⁷ <http://mo-ld.org>

⁸ The object model can be retrieved in JSON or XML from each InterMine endpoint.

⁹ <https://github.com/mo-ld/intermine-rdfizer>

¹⁰ <http://humanmine.org>

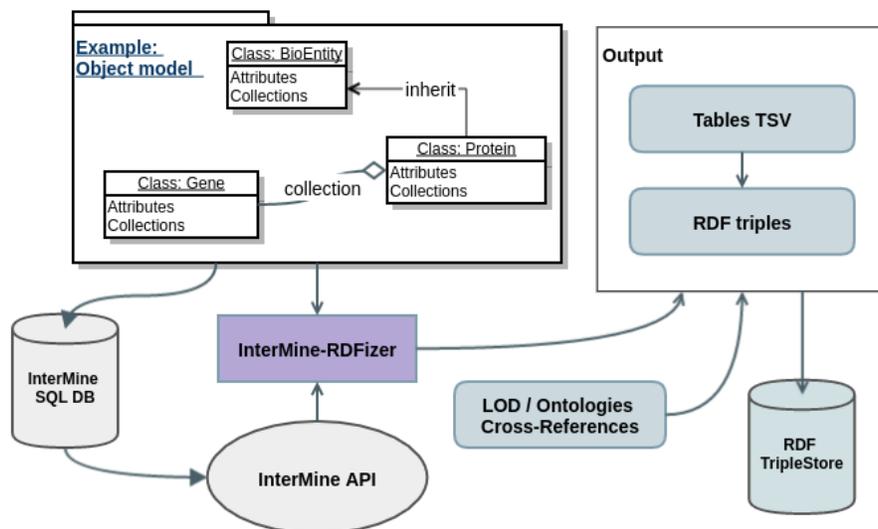


Fig. 1. InterMine-RDFizer data flow

coupled. The reported number of tables range from 89 (MouseMine) to 122 (RatMine) and the number of table relationships from 146 to 223. The script offers the possibility to maintain the data in its original loosely coupled manner, but the default option merges the information for the same resource, i.e. an entity with the same primary key in the SQL database. To maintain flexibility and avoid the need to manually specify hundreds of predicates, InterMine-RDFizer makes no assumptions about each database's vocabulary, and uses the generic prefix `<http://mo-ld.org/mine_vocabulary:>`. However, all the object literals are typed according to their SQL table column name and their database name. Therefore, the user has the ability to extensively query the endpoint with external ontologies and aggregate inter-database object types. For example, each InterMine endpoint has an object type for authors (`:yeastmine_Author`, `:flymine_Author`, etc.), but one can construct a SPARQL query (as shown in Query 1.1.) for the aggregation of the six MODs authors while using the Dublin Core¹¹, FOAF¹² and Schema.org¹³ vocabularies. The statistics of the six MOLD graphs are shown in Figure 2. Consistent with the fly being one of the most commonly used multicellular invertebrate model organisms, 366m triples were derived from FlyMine. Following the fly, in triples, are the human (HumanMine 304m) and the mouse (MouseMine 254m), the two most-studied vertebrate organisms, and in descending order RatMine (92m), YeastMine (83m) and ZebrafishMine (63m). The data

¹¹ <http://dublincore.org/>

¹² <http://xmlns.com/foaf/spec/>

¹³ <http://schema.org/>

from these six MODs together comprise 1.16B triples, 192m distinct subjects, 192m distinct entities, 188m distinct objects, 56m literals, 1081 types, and 977 properties.

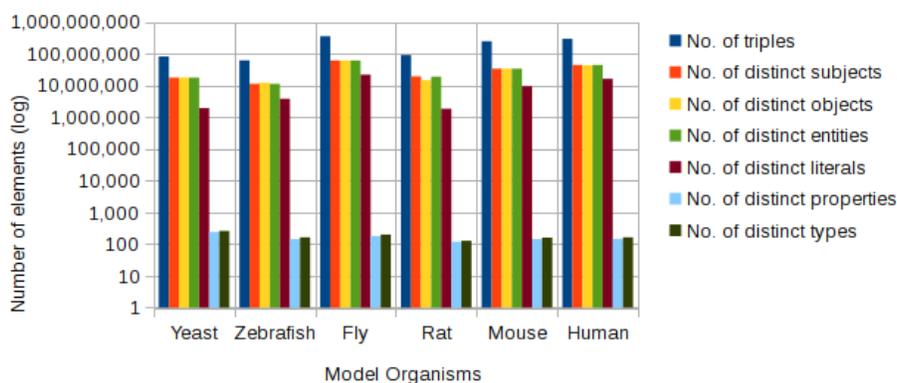


Fig. 2. Basic metrics of the 6 MODs graphs.

```

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX mold: <http://mo-ld.org/resource/>
PREFIX mold_voc: <http://mo-ld.org/mine_vocabulary:>
PREFIX schema: <http://schema.org/>
CONSTRUCT {
  ?s dc:contributor ?author .
  ?author foaf:name ?name .
  ?author rdf:type ?author_type, schema:Person
}
WHERE {
  ?subject mold_voc:hasAuthor ?author .
  ?author rdfs:label ?name .
  ?author rdf:type ?author_type
}
LIMIT 100

```

Query 1.1. SPARQL query with external ontologies

2.2 Linking of MODs

The InterMine-RDFizer also allows the user to create cross-references with other databases and ontologies. It is the only feature of the script that requires an educated user input. The user needs to provide a CSV file, which is then used by the script to map the external cross-references of the database with other linked data endpoints and ontologies. The script also validates each link with the targeted URI by making a simple ASK query¹⁴ over the linked entity. It creates

¹⁴ ask {<target URI> ?predicate ?object}

two different outputs, one RDF/N-TRIPLES file with all the putative links and a second one with only the validated links. Bio2RDF ([14], [15], [16]) is one of the broader LOD network for life sciences that can assign node identifiers for over 2000 datasets. With such a large network of supported databases, it was a natural choice to start connecting our graph to a larger network of biological data. We used prefixcommons¹⁵ to identify the right prefix for each cross-reference's datasource, which was then used to generate a proper Bio2RDF HTTP identifier. All database cross-references and ontologies contained in an InterMine instance were collected for further linking. Of the 92 external datasources used in the MODs, 38 (41%) were also present in Bio2RDF. Thus, linking the MODs with Bio2RDF added value to the original MODs. It is important to notice that even the links that are not currently validated were also incorporated into MOLD datasets. Because of the standardization of Bio2RDF URIs creation, we are guaranteed that once the targeted database is converted on their side, the links will bind. Figure 3 shows the connections between the 6 MOLD datasets and the 38 datasets. Only the Gene Ontology and the PantherDB are found in all the six MODs. Databases focused on genomics (genes, proteins, etc.) show the highest number of links. PantherDB (2,048,167), RefSeq (936,568), UniGene (399,637) and NCBI gene (352,585) databases have most links overall. Nonetheless, other important biological aspects such as phenotypes (HGNC and OMIM databases) and diseases (human disease ontology (DO)) are present. Cumulatively between the six MODs, 57% (3,382,672 of 5,923,399 links) of all their cross-references were existing in Bio2RDF. However, the connectivity of the MODs could be greatly increased if a few additional databases were made available as part of the network of linked open data. For instance, the entire conversion of PantherDB would increase the total connectivity of the six MODs by 22%, obtaining a global coverage of 79%.

3 Semantic platform for MOLD

This section presents the platform of the Model Organism Linked Database. We also outline our efforts to improve the deployment and reuse of the linked data platform using Docker.

3.1 MOLD Architecture

The MOLD Web application was built with simplicity in mind aiming to reuse state-of-the-art Semantic Web software. It comes with all the functionalities a user would expect to find in a Semantic Web platform: support for querying, browsing and exploring the data. Figure 4 shows the technologies used in the MOLD architecture. First, the SPARQL query editor and results viewer, respectively YASQE and YASR, are two components of YASGUI [17], a very user-friendly and commonly used editor. The editor is also customizable and comes with interesting features out of the box, such as auto-completion of predicates¹⁶

¹⁵ <http://prefixcommons.org/>

¹⁶ Auto-completion is based on <http://prefix.cc>

find a three way relationship between mouse, a specific gene annotation and human. Relfinder would then find the genes annotated for both of the organisms. The last piece of the MOLD Web application is the REST API. To adhere to best practice in API descriptions, we used the OpenAPIs¹⁸ specification and the Swagger-UI¹⁹. Our implementation currently supports five different commands: *search*, *describe*, *inlinks*, *outlinks* and *sparql*, that can be called via HTTP GET. The *describe* command is used to describe a resource identified by a URI. Specified via an option, *describe* can return either a long (with all the links) or short (attributes only) description. The two *link* (*in* and *out*) commands can be used to find other resources that the targeted URI connects with. The *sparql* instruction, as its name implies, provides a SPARQL query call that can also be sent via an HTTP POST, if needs be. Other components of the MOLD interface include a quick search and an interactive network of database connectivity that can be found in the *about* section.



Fig. 4. MOLD Web technologies

3.2 MOLD in the Cloud

To ease deployment of the MOLD infrastructure in the cloud, we built Docker images for the Virtuoso triple store, the MOLD web application and the MOLD API. The images are publicly hosted in the docker hub registry²⁰, along with documentation to allow users to launch their own MOLD containers. The code is available on GitHub²¹, licensed under the MIT license²², and includes InterMine-RDFizer, the Web application, the API, and the different docker configuration files. A Google group²³ has been created to allow for discussions about MODS and Linked Data. These resources will provide a *de facto* place to share common use cases and best practice.

¹⁸ <https://openapis.org/specification>

¹⁹ <http://swagger.io/swagger-ui/>

²⁰ <https://hub.docker.com/u/mold/>

²¹ <https://github.com/mo-ld>

²² <https://opensource.org/licenses/MIT>

²³ <https://groups.google.com/forum/#!forum/mo-ld>

4 Pan-Organism Analysis with MOLD

One of the key advantages of linked data is the use of a standardized language and access protocols to break down data silos and improve interoperability. To demonstrate the value of our model organism linked data platform, we have pursued two relevant use cases involving queries across the model organisms.

The first use case focuses on examining the set of orthologous genes between two or more species. Orthologous genes are genes from different species that share a common function and whose genetic lineage matches the species tree. To do this, we constructed a query (Query 1.2) to count the number of orthologous genes between human and yeast using PantherDB, a database of evolutionary relationships. While PantherDB²⁴ is not part of Bio2RDF or the network of Linked Open data, the graph-like nature of the representation of RDF triples allow us to find common PantherDB identifiers linked to by the mouse and human genes. The initial two-species query can be expanded to other species using a UNION clause. The results for all the organism-organism associations are reported in Table 1. Surprisingly, the largest number of uniquely identified orthologous genes were found in the zebrafish. The two organisms that shared the most entities were zebrafish and human, which we wouldn't expect to be the closest one, in evolutionary terms. Notice that the InterMine instances could be biased due to their representation in PantherDB. Yet this example proved the ease with which we can build cross-datasets statistics.

```
PREFIX mine_vocab: <http://mo-ld.org/mine_vocabulary:>
SELECT (COUNT (DISTINCT ?pantherOrtholog) as ?Count)
WHERE {
  GRAPH <http://human.mo-ld.org> {
    ?shuman skos:exactMatch ?pantherOrtholog .
    ?shuman mine_vocab:hasDataSource ?datasource .
    ?datasource rdfs:label ?dslabel .
    FILTER (lcase(str(?dslabel)) = "panther") }
  GRAPH <http://yeast.mo-ld.org> {
    ?syeast skos:exactMatch ?pantherOrtholog . }
}
```

Query 1.2. SPARQL query for PantherDB orthologous genes between the yeast and human

Model Organism	Yeast	Zebrafish	Fly	Rat	Mouse	Human
Yeast	2151	1842	785	788	261	1543
Zebrafish	1842	4862	1604	1869	526	3419
Fly	785	1604	2426	721	199	2411
Rat	788	1869	721	2483	26	1509
Mouse	261	526	199	26	796	409
Human	1543	3419	2411	1509	409	5024

Table 1. Shared orthologous genes MOLD from PantherDB

²⁴ <http://www.pantherdb.org/>

The second use case involves a pan-organism analysis to find genes with a specific function. Query 1.3 aims to find extrinsic components of a cell membrane (GO:0019898), a term that is specified in the Gene Ontology. In a nutshell, the query identifies reactions from the KEGG database for mice, zebrafish, and yeast genes annotated with the specified GO term. To do so, the federated query asks the Bio2RDF SPARQL endpoint to find the Enzyme Classification (EC) number contained in MOLD and fetches the reaction activity from KEGG. An interesting extension of the query would be to ask for the biological pathways associated with the resulting enzymes from KEGG, but this addition was omitted for brevity. In the context of drug development research, this kind of query could be useful to explore potential drug targets from gene annotations, or to evaluate drug safety with pathway analysis.

```

PREFIX mine_vocab: <http://mo-ld.org/mine_vocabulary:>
PREFIX b2f_go: <http://bio2rdf.org/go:>
PREFIX b2f_keyvoc: <http://bio2rdf.org/kegg_vocabulary:>
PREFIX void: <http://rdfs.org/ns/void#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?gene_entity ?kegg ?kegg_rx_label
WHERE {
  ?oTerm skos:exactMatch b2f_go:0019898 .
  ?oTerm mine_vocab:hasOntologyAnnotation ?gene_annotation .
  ?gene_annotation mine_vocab:hasBioEntity ?gene_entity .
  ?gene_entity mine_vocab:hasCrossReference ?xref .
  ?xref mine_vocab:hasDataSource ?ds .
  ?xref skos:exactMatch ?bio2rdf_ec .
  FILTER (?ds = <http://mo-ld.org/mousemine:9331717> ||
    ?ds = <http://mo-ld.org/zebrafishmine:14401557> ||
    ?ds = <http://mo-ld.org/yeastmine:1034153>)
  SERVICE <http://bio2rdf.org/sparql> {
    ?kegg b2f_keyvoc:x-ec ?bio2rdf_ec .
    ?kegg b2f_keyvoc:reaction ?kegg_rx .
    ?kegg_rx rdfs:label ?kegg_rx_label .
  }
}

```

Query 1.3. Federated SPARQL query to gather enzyme reaction for GO annotated genes

5 Conclusion

Our work creates a new and sustainable avenue by which model organism databases that use InterMine can be exposed as Linked Data. While our efforts focused on only 6 of the MODs, many more could also be exposed in a similar fashion. Our analysis of the network of linked data revealed the resources that are unique and/or shared by the MODs, and we demonstrate the utility of our transformation through pan-MOD queries. We use a common InterMine vocabulary to increase the interoperability of the data produced, and demonstrate how we can use SPARQL construct queries to expose these data with other vocabularies such as schema.org. Structuring model organism data for bioinformatics research is not new ([16], [21]). However, our approach of simultaneously engaging the MOD community and using W3C standards to expose data in a manner that

allows others to reproduce and extend our work yields a concrete milestone in generating Linked Data similar to other institutional efforts ([19], [22]). The software and data in this project are open source and available to the community, thus offering additional support towards the reproducibility of scientific research. Our current work is not without limitations. First, data available from a MOD website may differ from that of the InterMine instances, because MODs do not necessarily rely on InterMine as their primary store. In fact, some MODs, such as SGD, selectively move data into InterMine from a relational database, thereby yielding different results. Second, our approach does not attempt to structure data in a manner that has been promoted by the community. For instance, the FALDO[23] vocabulary has been put forward as a standard for describing the location of genomic features. As we continue to develop our approach, we will strive to include better integration of formal ontologies, including SIO[24], or the work on genotype-phenotype integration that is ongoing at the Monarch Initiative²⁵. We will also enhance our work by conducting a user experience evaluation of the MOLD platform and by collecting more use cases from the MOD community.

6 Acknowledgments

This work was supported by NIH/NHGRI U41HG001315 (M. Cherry, K Karra, G Binkley, J Sullivan) and supplement 3U41HG001315-21S1 (M. Dumontier, M Déraspe), NIH/NHGRI U41HG002659 (supplement subcontract to G.Micklem), the Wellcome Trust grant 099133 (G.Micklem), and J. Corbeil acknowledges the Canada Research Chair in Medical Genomics. The content is solely the responsibility of the authors and does not necessarily represent the official views of any of the funding bodies.

References

1. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, and Others, “Gene Ontology: tool for the unification of biology,” *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.
2. K. Eilbeck, S. E. Lewis, C. J. Mungall, M. Yandell, L. Stein, R. Durbin, and M. Ashburner, “The Sequence Ontology: a tool for the unification of genome annotations,” *Genome biology*, vol. 6, no. 5, p. R44, 2005.
3. P. N. Robinson and S. Mundlos, “The human phenotype ontology,” *Clinical genetics*, vol. 77, no. 6, pp. 525–534, 2010.
4. L. M. Schriml, C. Arze, S. Nadendla, Y.-W. W. Chang, M. Mazaitis, V. Felix, G. Feng, and W. A. Kibbe, “Disease Ontology: a backbone for disease semantic integration,” *Nucleic acids research*, vol. 40, no. D1, pp. D940—D946, 2012.
5. J. Sullivan, K. Karra, S. A. T. Moxon, A. Vallejos, H. Motenko, J. D. Wong, J. Aleksic, R. Balakrishnan, G. Binkley, T. Harris, B. Hitz, P. Jayaraman, R. Lyne, S. Neuhauser, C. Pich, R. N. Smith, Q. Trinh, J. M. Cherry, J. Richardson, L. Stein,

²⁵ <https://monarchinitiative.org>

- S. Twigger, M. Westerfield, E. Worthey, and G. Micklem, "InterMOD: integrated data and tools for the unification of model organism research.," *Scientific reports*, vol. 3, p. 1802, 2013.
6. A. Kalderimis, R. Lyne, D. Butano, S. Contrino, M. Lyne, J. Heimbach, F. Hu, R. Smith, R. Stépán, J. Sullivan, and G. Micklem, "InterMine: extensive web services for modern biology.," *Nucleic acids research*, vol. 42, pp. W468–72, jul 2014.
 7. M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, "The FAIR Guiding Principles for scientific data management and stewardship.," *Scientific data*, vol. 3, p. 160018, 2016.
 8. R. N. Smith, J. Aleksic, D. Butano, A. Carr, S. Contrino, F. Hu, M. Lyne, R. Lyne, A. Kalderimis, K. Rutherford, R. Stepan, J. Sullivan, M. Wakeling, X. Watkins, and G. Micklem, "InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data.," *Bioinformatics (Oxford, England)*, vol. 28, pp. 3163–5, dec 2012.
 9. R. Balakrishnan, J. Park, K. Karra, B. C. Hitz, G. Binkley, E. L. Hong, J. Sullivan, G. Micklem, and J. M. Cherry, "YeastMine—an integrated data warehouse for *Saccharomyces cerevisiae* data as a multipurpose tool-kit.," *Database : the journal of biological databases and curation*, vol. 2012, p. bar062, 2012.
 10. L. Ruzicka, Y. M. Bradford, K. Frazer, D. G. Howe, H. Paddock, S. Ramachandran, A. Singer, S. Toro, C. E. Van Slyke, A. E. Eagle, D. Fashena, P. Kalita, J. Knight, P. Mani, R. Martin, S. A. T. Moxon, C. Pich, K. Schaper, X. Shao, and M. Westerfield, "ZFIN, The zebrafish model organism database: Updates and new directions.," *Genesis (New York, N.Y. : 2000)*, vol. 53, pp. 498–509, aug 2015.
 11. R. Lyne, R. Smith, K. Rutherford, M. Wakeling, A. Varley, F. Guillier, H. Janssens, W. Ji, P. McLaren, P. North, D. Rana, T. Riley, J. Sullivan, X. Watkins, M. Woodbridge, K. Lilley, S. Russell, M. Ashburner, K. Mizuguchi, and G. Micklem, "FlyMine: an integrated database for *Drosophila* and *Anopheles* genomics.," *Genome biology*, vol. 8, no. 7, p. R129, 2007.
 12. S.-J. Wang, S. J. F. Lauderkind, G. T. Hayman, J. R. Smith, V. Petri, T. F. Lowry, R. Nigam, M. R. Dwinell, E. A. Worthey, D. H. Munzenmaier, M. Shimoyama, and H. J. Jacob, "Analysis of disease-associated objects at the Rat Genome Database.," *Database : the journal of biological databases and curation*, vol. 2013, p. bat046, 2013.
 13. H. Motenko, S. B. Neuhauser, M. O'Keefe, and J. E. Richardson, "MouseMine: a new data warehouse for MGI.," *Mammalian genome : official journal of the International Mammalian Genome Society*, vol. 26, pp. 325–30, aug 2015.
 14. F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, and J. Morissette, "Bio2RDF: towards a mashup to build bioinformatics knowledge systems.," *Journal of biomedical informatics*, vol. 41, pp. 706–16, oct 2008.
 15. M.-A. Nolin, P. Ansell, F. Belleau, K. Idehen, P. Rigault, N. Tourigny, P. Roe, J. M. Hogan, and M. Dumontier, "Bio2RDF network of linked data.," in *Semantic*

- Web Challenge; International Semantic Web Conference (ISWC 2008)*, Citeseer, 2008.
16. A. Callahan, J. Cruz-Toledo, P. Ansell, and M. Dumontier, "Bio2RDF Release 2: Improved Coverage, Interoperability and Provenance of Life Science Linked Data," in *The Semantic Web: Semantics and Big Data* (P. Cimiano, O. Corcho, V. Presutti, L. Hollink, and S. Rudolph, eds.), vol. 7882 of *Lecture Notes in Computer Science*, pp. 200–212, Springer Berlin Heidelberg, 2013.
 17. L. Rietveld and R. Hoekstra, "Yasgui: Not just another sparql client," in *The Semantic Web: ESWC 2013 Satellite Events*, pp. 78–86, Springer, 2013.
 18. C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, "DBpedia - A crystallization point for the Web of Data," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 7, pp. 154–165, sep 2009.
 19. S. Jupp, J. Malone, J. Bolleman, M. Brandizi, M. Davies, L. Garcia, A. Gaulton, S. Gehant, C. Laibe, N. Redaschi, S. M. Wimalaratne, M. Martin, N. Le Novère, H. Parkinson, E. Birney, and A. M. Jenkinson, "The EBI RDF platform: linked open data for the life sciences.," *Bioinformatics (Oxford, England)*, vol. 30, pp. 1338–9, may 2014.
 20. P. Heim, S. Hellmann, J. Lehmann, S. Lohmann, and T. Stegemann, "RelFinder: Revealing relationships in RDF knowledge bases," in *Semantic Multimedia*, pp. 182–187, Springer, 2009.
 21. E. Antezana, W. Blondé, M. Egaña, A. Rutherford, R. Stevens, B. De Baets, V. Mironov, and M. Kuiper, "BioGateway: a semantic systems biology tool for the life sciences.," *BMC bioinformatics*, vol. 10 Suppl 1, p. S11, 2009.
 22. G. Fu, C. Batchelor, M. Dumontier, J. Hastings, E. Willighagen, and E. Bolton, "PubChemRDF: towards the semantic annotation of PubChem compound and substance databases.," *Journal of cheminformatics*, vol. 7, p. 34, 2015.
 23. J. Bolleman, C. J. Mungall, F. Strozzi, J. Baran, M. Dumontier, R. J. P. Bonnal, R. Buels, R. Hoehndorf, T. Fujisawa, T. Katayama, and P. J. A. Cock, "FALDO: A semantic standard for describing the location of nucleotide and protein feature annotation.," *bioRxiv*, 2014.
 24. M. Dumontier, C. J. Baker, J. Baran, A. Callahan, L. Chepelev, J. Cruz-Toledo, N. R. Del Rio, G. Duck, L. I. Furlong, N. Keath, D. Klassen, J. P. McCusker, N. Queralt-Rosinach, M. Samwald, N. Villanueva-Rosales, M. D. Wilkinson, and R. Hoehndorf, "The SemanticScience Integrated Ontology (SIO) for biomedical research and knowledge discovery.," *Journal of biomedical semantics*, vol. 5, no. 1, p. 14, 2014.