

# Implementing FAIR Principles in InterMine

Daniela Butano<sup>1</sup>, Justin Clark-Casey<sup>1</sup>, Sergio Contrino<sup>1</sup>, Josh Heimbach<sup>1</sup>, Rachel Lyne<sup>1</sup>, Kevin Herald Reierskog<sup>1</sup>, Julie Sullivan<sup>1</sup>, Yo Yehudi<sup>1</sup>, and Gos Micklem<sup>1</sup>

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

**Abstract.** InterMine is an established platform to integrate and access life sciences data providing a web interface and RESTful web services. In order to make the data integrated in the different InterMine deployments even more Findable, Accessible, Interoperable and Reusable, we have been improving InterMine adherence to FAIR principles, adopting concepts as such as persistent URIs, standards for embedding data descriptions into web pages, describing data with ontologies, and data licences.

## 1 Introduction

InterMine [1] is a platform to integrate and access life sciences data, providing flexible querying through a web interface as well as RESTful web services [2]. Whilst InterMine comes with a core data model for common biological entities, different deployments can extend these components to publish any type of data. InterMine is an established platform first released in 2006, and already includes some FAIR principles such as search and structured query functionalities, web services, and cross-references to other InterMine instances and resources. We will describe here how we are improving InterMine adherence to FAIR principles [3].

## 2 Persistent URIs

InterMine already has unique URLs to identify the report pages for biological entities, but these are based on internal InterMine IDs that change at every database build. To achieve data **findability** and **accessibility**, we have generated new navigable URLs based on the InterMine class names combined with local IDs provided by the data resource providers. For example, in FlyMine, the URL of the report page for the protein, with UniProt accession Q9V4E1, will be `https://www.flymine.org/flymine/protein:Q9V4E1`. Adding the InterMine database instance to third party resolvers, as such Identifiers.org [4], we can generate persistent URIs with pattern: `http://identifiers.org/a_mine_unique_namespace/class_name:local.ID`.

### 3 Describing data with ontologies

The InterMine system is based on a core data model, described in an XML file which defines classes (the entities in the model) and the relationships between them. InterMine already automatically applied terms from the Sequence Ontology [5] to its data model, but, to improve data **interoperability** and **reusability**, we have added more ontologies to its core data model and provided InterMine instance administrators with the ability to apply any other ontology describing their data model extension. The ontologies applied are available in the data model and will be used in the generation of RDF.

### 4 Marking up web pages

In order to improve **findability**, we have applied structured data in JSON-LD format to InterMine web pages, using Bioschemas.org [6] *DataCatalog* profile in the home page and *DataSet* profile in the report page for DataSet. The Bioschemas.org types *Gene* and *Protein* are in development stage.

### 5 Publishing Data Licences

To improve one of many aspects related to data **reusability**, InterMine has updated its model, adding the attribute *licence* to include the licences that govern the data sets that have been integrated. As a data integrator, we must propagate the licences provided for the underlying data by displaying them in the dataset report pages and in query results. At the moment only a minority of data sets have a licence. We will propagate the licence information when generating RDF.

## References

1. Smith RN, et al. InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. *Bioinformatics*. 28(23):3163-5 (2012) <https://doi.org/10.1093/bioinformatics/bts577>
2. Kalderimis A, et al. InterMine: extensive web services for modern biology. *Nucleic Acids Res*. 42(Web Server issue):W468-72 (2014)
3. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 3: 160018 (2016) <https://doi.org/10.1038/sdata.2016.18>
4. Sarala M. Wimalaratne, Nick Juty, John Kunze et al. Uniform resolution of compact identifiers for biomedical data. *Scientific Data*. 5: 180029 (2018). <https://doi.org/10.1038/sdata.2018.29>
5. Eilbeck K., Lewis S.E., Mungall C.J., Yandell M., Stein L., Durbin R., Ashburner M. The Sequence Ontology: A tool for the unification of genome annotations. *Genome Biology* 6:R44 (2005)
6. Gray, A.J.G, Goble, C.A. and Jimenez, R., 2017. Bioschemas: From Potato Salad to Protein Annotation. In International Semantic Web Conference (Posters, Demos & Industry Tracks).