Implementing FAIR Identifiers in InterMine

Daniela Butano[1], Sergio Contrino[1], Justin Clark-Casey[1], Josh Heimbach[1], Rachel Lyne[1], Julie Sullivan[1], Yo Yehudi[1], and Gos Micklem[1]

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

Abstract. InterMine is an established platform to integrate and access life sciences data; it provides a web interface and RESTful web services. We examined two possible solutions to make its identifiers FAIRer, one including data source prefixes, and another using InterMine classes. We have decided for the latter schema which, while incorporating some semantic elements, requires much less configuration and is easier to adopt by InterMine maintainers.

1 Introduction

InterMine [1] is a platform to integrate and access life sciences data. It provides a web interface and RESTful web services. Other organizations download and deploy InterMine on their servers. Whilst InterMine comes with a data model for core biological entities and loaders for common data sources, deployments can extend these components to publish any type of data. We plan to improve the format of InterMine identifiers in accordance with FAIR guidelines [2,4].

2 InterMine identifiers

InterMines existing identifiers are URIs that incorporate an internal database ID which is not preserved across releases. We have identified two possible solutions that adopt the recommended practice [4] of using external IDs from one of the data sources integrated.

URI Schema A - data sources prefixes Identifiers.org [3] provides compact identifiers (CURIEs) to uniquely reference records maintained by data resources. A CURIE consists of a prefix and a locally unique identifier (LUI) in the form `<prefix>:<LUI>`. The prefix comes from a registry maintained by identifiers.org and denotes a particular data collection. The LUI is the unique identifier assigned by that collection to its data record. For example uniprot:P12883 denotes a record in the Uniprot KnowledgeBase with the accession P12883. InterMine could appends these CURIEs to the deployments URI. The CURIEs for a particular model class will use the data collection that is most comprehensive across the address space of that class. For instance, the protein class will use the uniprot data collection to create URIs such as `http://mine.org/uniprot:P12883` to denote a protein data record that incorporates information from many sources. This solution does not incorporate semantic information into the

URIs, and enables an easy integration with other systems; however this approach relies on additional manual configuration by InterMine operators in order to match the prefixes with the types of data.
URI Schema B - InterMine class name prefixes This schema will use InterMine class names as prefixes: the same example used before will be `http://mine.org/protein:P12883`. The solution therefore includes semantic elements to form IDs and adds a potential maintenance point since resolution must be preserved when the model is restructured or class names change [4]. This approach will need a configuration file to specify which field is the accession number/identifier for a given class (e.g. Protein.primaryAccess, Pathway.identifier). The configuration will be provided by the InterMine system for the core model, while the InterMine operators would need to create extra configuration only for the classes they add.

Neither solution includes versioning of the identifiers: while a release-level versioning would be technically easy to implement, this is not a feasible course of action for many database maintainers, many of whom lack the resources to provide multiple versioned instances of the resource indefinitely.

3 Conclusions

We decided to use Schema B because of the considerably lower configuration effort. This option requires less human intervention and therefore should increase participation and lower error rates. Independently of the URI schema, we will recommend any InterMine instance to adopt Identifers.org as a Permanent URL (PURL) provider, registering itself as a data collection. For example the mymine instance, registered in Identifiers.org, with prefix mymine and URI `http://identifiers.org/mymine` will provide URIs as `http://identifiers.org/mymine:protein:P12883` which will be redirected to `http://mymine.org/protein:P12883`.

References

1. Smith RN, Aleksic J, Butano D, 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. 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
3. 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
4. McMurry JA, Nick Juty, Niklas Blomberg et al. How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. PLoS Biol. 2017, 15(6):e2001414. https://doi.org/ 10.1371/journal.pbio.2001414