FAIRDOM approach for semantic interoperability of systems biology data and models

Olga Krebs[1*], Katy Wolstencroft[3], Natalie Stanford[2], Norman Morrison[2], Martin Golebiewski[1], Rostyk Kuzyakiv[4], Stuart Owen[2], Quyen Nguyen[1], Jacky Snoep[2], Wolfgang Mueller[1], and Carole Goble[2]

[1] Heidelberg Institute for Theoretical Studies, Germany
[2] School of Computer Science, University of Manchester, UK
[3] Leiden Institute of Advanced Computer Science, Leiden, NL
[4] University of Zurich, Switzerland

ABSTRACT

Motivation: The ability to collect and interlink heterogeneous data and model collections is essential in systems biology. Effective data exchange and comparison requires sufficient data annotation. This is particularly apparent in systems biology, where data heterogeneity means that multiple community metadata standards are required for the annotation of a whole investigation, including data, models and protocols.

Results: FAIRDOM (http://fair-dom.org/) is an initiative to enable the systems biology community to produce and publish FAIR Data, Operating procedures and Models. It allows research assets to be aggregated, interlinked and shared in the context of the systems biology investigations that produced them. Here we present the FAIRDOM strategy in the context of semantic data integration, and how it supports the whole life cycle of data collection, annotation, sharing and reuse of systems biology data and resources.

Availability: https://fairdomhub.org

*** Contact:** olga.krebs@h-its.org

1 INTRODUCTION

Data integration is an essential part of systems biology. Scientists need to combine different sources of information in order to model biological systems, and relate those models to available experimental data for validation. Metadata is an important aspect of data management and data sharing. Annotating experimental results with a consistent set of information allows for easier discovery of relevant data as well as enabling others to potentially reuse it. Metadata ranges from simple descriptions about when an experiment was done to more detailed descriptions of where biological samples originated, how they were prepared, and what the experimental conditions were at the time of the experiment. Currently, only a small fraction of the data and models produced during systems biology investigations are deposited for reuse by the community, and only a smaller fraction of that data is standards compliant, semantically enriched content.

FAIRDOM project is a joint action of ERA-Net ERASysAPP (https://www.erasysapp.eu/) and European Research Infrastructure ISBE (http://project.isbe.eu/) to establish a data and model management service facility for systems biology. Its prime mission is to support researchers, students, trainers, funders and publishers by enabling systems biology projects to make their Data, Operating procedures and Models, Findable, Accessible, Interoperable and Reusable (FAIR). FAIRDOM builds on the outcomes of the successful SysMO-DB and SyBIT data management projects, uniting their tool and database development as well as their experience serving large systems biology projects. FAIRDOMHub is a web-based platform comprising two main components: SEEK (http://seek4science.org) as a web-based front-end cataloguing and metadata platform and openBIS as a back-end LIMS for scalable local data collection and processing (https://sis.id.ethz.ch/software/openbis.html). Here we present the semantic data integration in SEEK, and how it supports the whole life cycle of data collection, annotation, sharing, and reuse of systems biology data and resources.

2 APPROACH

The SEEK [1] is based on the ISA infrastructure (Investigations, Studies and Assays), a standard format for describing how individual experiments (assays) are aggregated into wider studies and investigations [2]. The Just Enough Results Model (JERM) describes the interrelations between assets and the metadata fields required to describe them. For example, for each dataset uploaded to SEEK, the JERM describes what type of experiment it was, what was measured, and what the values in the dataset mean. The JERM captures the core elements of MIBBI metadata, allowing users to comply with these standards as well as capturing the information required for linking in SEEK. The JERM Ontology (available from the BioPortal, http://bioportal.bioontology.org/ontologies/1488) is an application ontology designed to describe the relationships between items in SEEK (for example, data, models, experiment descriptions, samples, protocols, standard operating procedures and publications); and to enable these relationships to be expressed with formal semantics. It is based on the idea of the Minimal Information Models (https://www.biosharing.org), which have been collected under the umbrella of MIBBI (Minimum Information for Biological and Biomedical Investigations).

3 METHODS

The majority of laboratory scientists use spreadsheets for the daily management and manipulation of data, so the RightField semantic spreadsheet application [3] (also part of this work) is used to embed semantic annotation into the data. Individual cells, columns, or rows in spreadsheets can be restricted to particular ranges of allowed classes or instances from chosen ontologies. By embedding the JERM

metadata model in a spreadsheet format, and enabling the use of JERM (and other) vocabulary terms for annotation, the process of standardized semantic data collection can become part of the existing data management activities in the laboratory. Bioinformaticians, with experience in ontologies and data annotation, can prepare RightField-enabled spreadsheets with embedded ontology term selection support for distribution across the consortium.

JERM-compliant spreadsheet templates have been developed for a wide range of experimental data types, their collection is available from http://docs.seek4science.org/help/templates.html.

By embedding semantic technologies into familiar data management tools, the SEEK enables semantic annotation of new data and the generation and querying of Linked Data - compliant datasets, whilst hiding the complexities of ontologies and metadata from its users.

Underlying semantic web resources additionally extract and serve SEEK metadata in RDF (Resource Description Format). RDF enables rich semantic queries, both within SEEK and between related resources in the web of Linked Open Data.

ACKNOWLEDGEMENT

This work was funded by the BBSRC (BBG0102181, BB/I004637/1, BB/M013189/1), and by the BMBF grants 0315749, 20315781 and 031A525. We would like to thank the FAIRDOM PALS and users for their valuable feedback, testing and comments.

REFERENCES

1. Wolstencroft et al (2015) SEEK: a systems biology data and model management platform. BMC Systems Biology (9)33 DOI:10.1186/s12918-015-0174-y
2. Rocca-Serra, P., Brandizi, M., Maguire, E., Sklyar, N., Taylor, C., Be-gley, K., Field, D., Harris, S., Hide, W., Hofmann, O. et al. (2010) ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. Bioinformatics, 26, 2354-2356.
3. Wolstencroft, K., Owen, S., Horridge, M., Krebs, O., Mueller, W., Snoep, J.L., du Preez, F. and Goble, C. (2011) RightField: embedding ontology annotation in spreadsheets. Bioinformatics, 27, 2021-2022