Title FAIRDOM: Reproducible Systems Biology through FAIR Asset Management

Natalie J Stanford[1], Finn Bacall[1], Martin Golebiewski[2], Olga Krebs[2], Rostyk Kuzyakiv[3], Quyen Nguyen[2], Stuart Owen[1], Stian Soiland-Reyes[1], Jakub Straszewski[5], Dawie van Niekerk[6], Alan Williams[1], Katy Wolstencroft[4], Lars Malmstroem[3], Bernd Rinn[5], Jacky L. Snoep[1,6], Wolfgang Müller[2] and Carole Goble[1]

[1]School of Computer Science, The University of Manchester, United Kingdom
[2]Heidelberg Institute for Theoretical Studies, Germany
[3]University of Zürich, Switzerland,
[4]Leiden Institute of Advanced Computer Science, Netherlands
[5]ETH Zürich, Switzerland
[6]Stellenbosch University, South Africa
Corresponding author: natalie.stanford@manchester.ac.uk

Introduction

The FAIR data principles state [1] that data should be Findable (and citable), Accessible (with appropriate caveats for sensitive data), Interoperable (can be combined, typically through adherence to standards) and Reusable (can be reused later or reproduced later from a publication). These principles lie at the heart of a number of Research Infrastructures within Europe such as ELIXIR (http://www.elixir-europe.org) and ISBE – Infrastructure for Systems Biology in Europe (http://project.isbe.eu/). FAIRDOM (http://fair-dom.org) [2] is a European Research Infrastructure initiative sponsored by ISBE and the ERANet ERASysAPP. FAIRDOM's primary mission is to support researchers, students, trainers, funders and publishers by enabling Systems and Synthetic Biology *projects* - their team members and collaborators - to make their Data, Operating procedures and Models FAIR. Integrative biology inherently has multiple parts, typically includes multiple parties and practically scatters its outcomes across multiple resources. These characteristics have consequences for reproducibility, sharing and ultimately wider publication.

Multiple and interrelated components - Mathematical modelling methods and laboratory experiments are *combined* in order to understand and predict dynamic processes in living systems. Heterogeneous data, including multiple 'omics datasets, are integrated and interlinked with mathematical models to allow results to be interpreted, compared and/or reused. The Standard Operating Procedures associated with the data are essential in order to interpret the data and to inform the modelling. Thus any investigation will include many data files, SOPs and models; this contextualized and structured collection of outputs backs up the findings of the associated article.

Multiple and dispersed disciplines - Systems Biology projects typically involve the exchange of data between partners who are often distributed in different institutions. Data from several labs may need to be integrated into a single model, possibly with little input from the original experimentalists. Annotation with rich metadata and adherence to metadata standards is essential. Numerous standards influence the work of biologists, ranging from discipline-agnostic recommendations to specialized community formats, checklists and ontologies [3]. However, standardized metadata has to be supported by resources, skills and enabling tools such as the CombineArchive [4].

Multiple and dispersed repositories - Recorded outcomes are typically held in local lab stores or deposited in type-specific specialized public repositories and catalogues, for example BioModels, GEO, and PRIDE (see [3] for a survey). In order to facilitate domain specific research it is important to support these specialist public repositories. However, integrative investigations require a further platform to draw together this scattered content and give it context.

Methods

FAIRDOM addresses each of these challenges, working together with the different sources and publication possibilities for data, models, and SOPs and supporting Systems, and more recently Synthetic, Biology project teams. The initiative has three chief components:

- *Community activities* for knowledge exchange, engagement with standards bodies and training. We aim to: raise the profile of reproducibility and stewardship amongst funders, publishers and researchers; improve the curation, data citation and reproducibility skills of researchers; and create a self-supporting community of stewardship-aware investigators and policy makers. See http://www.fair-dom.org.
- *Project support* for major projects through a blend of asset lifecycle and process planning; technical support; metadata and curation support (data template making, technical model reproducibility and validation) and training. Project ambassadors (called "PALs") bridge between project teams and the FAIRDOM consortium.
- *Software infrastructure* for FAIR asset stewardship that addresses the issues introduced earlier and enables *considered* sharing and supports *reproducible* publishing. This is the subject of the rest of the abstract.

The software infrastructure has three threads:

- *The FAIRDOM openSEEK platform* comprising three main components. The SEEK (http://seek4science.org) is our web-based front end cataloguing and metadata platform [5]. openBIS (https://sis.id.ethz.ch/software/openbis.html) is our back-end LIMS and analytics platform for scalable local data collection and processing [6]. Finally, we employ a pool of plug-in tools and resources, which includes our own specialist software (e.g. JWS Online, RightField), third party specialist software (e.g. LabArchives, BiVes, Cytoscape), specialist public archives (e.g. PubMed, BioModels, GEO, JBEI-ICE) and general purpose public archives (e.g. Zenodo, Figshare). openSEEK is open source and freely available.
- *Specific installations of the platform* that enable data and model management to be customised, locally managed and combined with local data management platforms. Installations range from national programmes and projects to institution-based laboratories.
- The *FAIRDOMHub* (http://fairdomhub.org) is our public data and model Community Commons hosted at HITS. An instance of the SEEK component, it includes a yellow pages (programmes, projects, institutions and people), their experiments, and their research assets (data, models, SOPs, samples, organisms and publications). The Hub is a one-stop catalogue and showcase for projects; a safe store for uploaded files, and a directory of researchers.

The FAIRDOMHub demonstrates our approach to the reproducibility, sharing and publication challenges outlined in the introduction. It provides a synthesis - a common, related view - of the data, SOPs and models of a project or a group of projects. This related view can be used before, during, and after publication. Our Just Enough Results Model (JERM) describes the interrelations between assets and the metadata fields required to describe them. We use the ISA format [7] which allows the aggregation of individual assays or analyses into related studies and investigations (Figure 1). We reuse existing Minimum Information Models for life science data (e.g. MIAME, MIAPE, MIASE) with the aim to capture the least amount of information needed for someone to understand and interpret an experiment. Elements in common (sample information, type of data, what was measured) are augmented with data-specific elements. JERM templates are defined and shared in spreadsheet format for different types of experimental data, using our RightField annotation framework (http://www.rightfield.org.uk) [8]. Our support for Systems Biology metadata standards, notably SBML and SED-ML, enables models to be simulated, validated, and compared using plug-ins, and ensure exchangeability and reproducibility of simulation experiments. Metadata about resources in SEEK is available as both XML and RDF through content negotiation. The RDF is structured around our Just Enough Results Model (JERM) ontology [9]. It is possible to run SEEK together with a triple store and SPARQL endpoint [10]. We plan to expose a SPARQL endpoint for the FAIRDOMHub shortly.

As assets may well be stored in different repositories hosted by different organizations FAIRDOMHub organizes files, models, data, and protocols in one place *regardless of where they physically reside*. Thus it is an over-arching, cross-repository catalogue. Its metadata describes,

references and accesses content spread over distributed stores, thereby retaining the connections to the investigation context and aggregating a project's assets. By linking through FAIRDOMHub we reuse international public repositories (e.g. BioModels), aid submission to them and enable users to store their data and models in the appropriate public archives. Moreover, by including third party resources in our catalogue we can securely link to local stores where data is too large or too sensitive to relocate. The projects control what is public or private, making it easy and safe to collaborate with colleagues or just team members. Recent work has experimented with how we frame sharing options to "nudge" contributors to more open sharing practices [11]. FAIRDOMHub tracks licenses, versions, derivations and attributions of all the content it manages, and also tracks asset ownership. Individual research assets and their ISA aggregates are identified with unique and persistent HTTP URLs. Our support for ORCiD aims to track and build credit for contributors

When it comes to publishing, each asset or the whole ISA view can be snapshot and registered with a DOI, and made public for peers to be able to review, reuse and cite. As FAIRDOM's ISA-structured view of an experiment is effectively a mechanism for packaging metadata we can readily adopt the Research Object (http://www.researchobject.org) framework [12]. The framework takes a systematic and standards based approach for physically and logically (through identifiers referencing entries in repositories) bundling the digital components into off-the-shelf container platforms (such as Zip, BagIt and Docker). We create and publish Research Object snapshots at the Investigation and at the Studies and Assays/Analysis level, along with assigning a DOI. These FAIRDOM Research Objects contain bundles of assets that are portable and can be deposited in community stores, like Zenodo. They also afford a means of exchange between different SEEK installations. Research Object bundles are powerful mechanisms for reproducibility. With this in mind their specification and that of the COMBINEArchive OMEX file format [13] have been aligned.

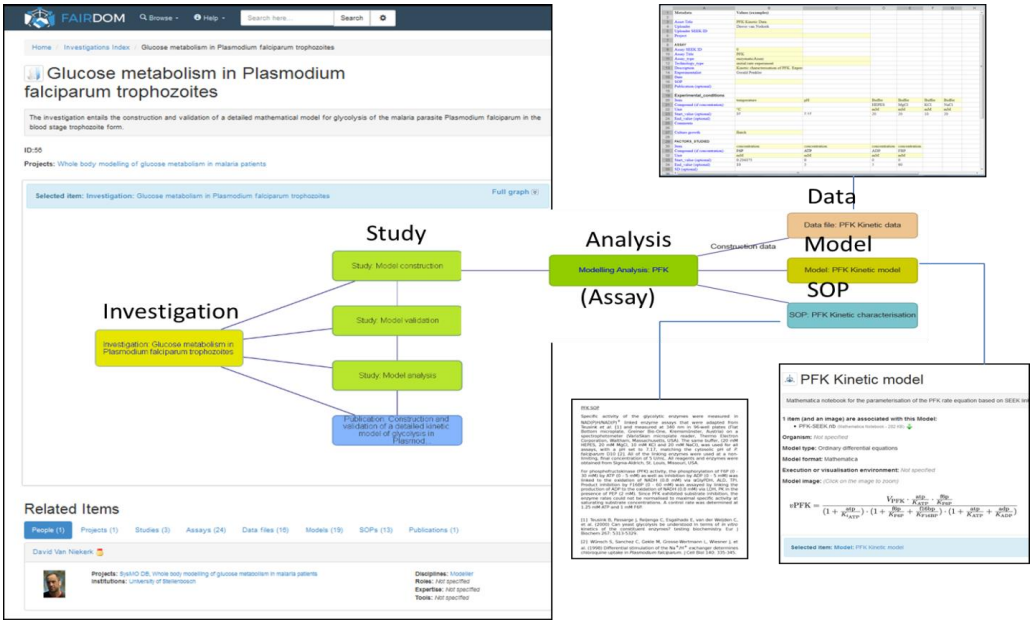

Figure 1: a montage navigating through a section of the ISA to a model with its data and SOP used in model construction for doi [15] associated with [14].

Results

Consider [14], an article on the kinetic model of glycolysis in Plasmodium falciparum supported by 24 analyses, 16 data files, 19 models, and 13 Standard Operating Procedures. The DOI [15] resolves to the full investigation entry on FAIRDOMHub (Figure 1). All the components are available for reproducing this Systems Biology modeling experiment, and it is richly annotated by metadata. The assets indicated may be uploaded to the Hub or may reside in other repositories to be accessed through their identifiers and fetched through APIs (subject to credentials if necessary). The FAIRDOMHub currently contains over 1200 data files, over 130 Systems Biology models, and over 200 SOPs contributed from more than 40 different Systems Biology projects.

These range from pan-national consortia in large EU initiatives, such as ERASysAPP, de.NBI and SysMO, to smaller projects from independent labs. Since December 2015 users have been able to independently register and self-manage their own projects and content. The Hub is governed by regulations and policies for deposition, metadata standardization, and FAIR use, reuse and sharing. HITS guarantee the Hub's sustainability until 2029. We are working with publishers to encourage the use of the Hub as a companion resource for supplementary materials and with funders to encourage that the Hub be used as a resource for retaining and promoting the outcomes of their managed programmes.

The FAIRDOM platform has been independently installed by over 22 organisations, ranging from large national facilities to centres of excellence and national projects. Recently the platform has been adopted and adapted by UK Synthetic Biology centres, adding extensions to support externally defined theoretical pathways and to track their synthesized version through to collation with the results needed for documenting the reproducibility of those experiments.

Conclusions

The FAIR principles for Systems Biology are supported by the capabilities of the Hub and in particular our SEEK software that underpins it. By drawing together the multiple components of investigations, regardless of their physical location, we contextualize experiments and richly annotate and interlink the components. By using the FAIRDOM software as they run their projects, either through the Hub or independent installations, researchers can prepare for reproducible publication and more effective exchange with collaborators. Although developed for Systems Biology, the FAIRDOM software infrastructure is generic and can be applied to a much wider range of Life Science investigations. Although infrastructure is important, the FAIRDOM initiative dedicates equal resources to community activities. To support this, continued technical developments and to scale out project service support, we are establishing the FAIRDOM Association and are actively seeking members.

Acknowledgements

This work was supported by the Biotechnology and Biological Sciences Research Council (BBG0102181, BB/I004637/1, BB/M013189/1); Bundesministerium für Bildung und Forschung (0315749, 031A525, 0315781, 031A371); Klaus Tschira Foundation (KTS); SystemsX; DST/NRF - South Africa, (SARCHI 82813 and TTK14051967526); and the Netherlands Organisation for Scientific Research (832.14.004).

References

1. Wilkinson M, Dumontier M et al (2016) The FAIR Guiding Principles for scientific data management and stewardship Scientific Data 3, doi:10.1038/sdata.2016.18
2. FAIRDOM: Data, operations and model stewardship for Systems Biology http://fair-dom.org
3. Stanford NJ, Wolstencroft K, Golebiewski M, Kania R, Juty N, Tomlinson C, Owen S, Butcher S, Hermjakob H, Le Novère N, Mueller W, Snoep J and Goble C (2015) The evolution of standards and data management practices in systems biology, Molecular Systems Biology 11: 851 DOI 10.15252/msb.20156053
4. Scharm M, Wendland F, Peters M, Wolfien M, Theile T, Waltemath D (2014) The CombineArchive Toolkit - facilitating the transfer of research results. PeerJ PrePrints 2:e514v1 https://doi.org/10.7287/peerj.preprints.514v1
5. Wolstencroft K, Owen S, Kreb O, Nguyen Q, Stanford NJ, Golebiewski M, Weidemann A, Bittkowski M, An L, Shockley D, Snoep JL, Muelle W, Goble C (2015) SEEK: a systems biology data and model management platform", BMC Systems Biology", 2015", 9(1):1-12 doi="10.1186/s12918-015-0174-y",
6. Bauch A, Adamczyk I, Buczek P, Elmer F-J, Enimanev K, Glyzewski P, Kohler M, Pylak T, Quandt A, Ramakrishnan C, Beisel C, Malmström L, Aebersold R and Rinn B (2011) openBIS: a flexible framework for managing and analyzing complex data in biology research, BMC Bioinformatics 12(468) DOI: 10.1186/1471-2105-12-468

7. Sansone SA, Rocca-Serra P, Brandizi M, Brazma A, Field D, Fostel J, Garrow AG, Gilbert J, Goodsaid F, Hardy N, Jones P, Lister A, Miller M, Morrison N, Rayner T, Sklyar N, Taylor C, Tong W, Warner G, Wiemann S, and Members of the RSBI Working Group. OMICS: A Journal of Integrative Biology (2008), 12(2): 143-149. doi:10.1089/omi.2008.0019.
8. Wolstencroft K, Owen S, Horridge M, Krebs O, Mueller W, Snoep JL, du Preez F, Goble C (2011) RightField: embedding ontology annotation in spreadsheets Bioinformatics 27(14):2021 – 2022, DOI:10.1093/bioinformatics/btr312.
9. Wolstencroft K, Owen S, Krebs O, Mueller W, Nguyen Q, Snoep JL, Goble C (2013) Semantic Data and Models Sharing in Systems Biology: The Just Enough Results Model and the SEEK Platform. The Semantic Web – ISWC 2013 LNCS Volume 8219: 212-227. doi:10.1007/978-3-642-41338-4_14
10. SEEK Sparql endpoint http://docs.seek4science.org/tech/setting-up-virtuoso.html
11. Garza K, Goble C, Brooke J, Jay C (2015) Framing the Community Data System Interface, figshare https://dx.doi.org/10.6084/m9.figshare.1300051.v5
12. Bechhofer S, Buchan I, De Roure D, Missier P, Ainsworth J, Bhagat J, Couch P, Cruickshank D, Delderfield M, Dunlop I, Gamble M, Michaelides D, Owen S, Newman D, Sufi S, Goble C (2013) Why linked data is not enough for scientists Future Generation Computer Systems 29(2): 599 – 611 doi:10.1016/j.future.2011.08.004.
13. Bergmann FT, Adams R, Moodie S, Cooper J, Glont M, Golebiewski M, Hucka M, Laibe C, Miller AK, Nickerson DP, Olivier BG, Rodriguez N, Sauro HM, Scharm M, Soiland-Reyes S, Waltemath D, Yvon F and Le Novère N (2014) COMBINE archive and OMEX format: one file to share all information to reproduce a modeling project, BMC Bioinformatics15:369 DOI: 10.1186/s12859-014-0369-z
14. Penkler G, du Toit F Adams W, Rautenbach M, Palm DC, van Niekerk DD and Snoep JL (2015) Construction and validation of a detailed kinetic model of glycolysis in Plasmodium falciparum, FEBS J 282(8): 1481–1511, https://dx.doi.org/10.1111/febs.13237
15. Penkler G, du Toit F Adams W, Rautenbach M, Palm DC, van Niekerk DD and Snoep JL (2015) Glucose metabolism in Plasmodium falciparum trophozoites; FAIRDOMHub. http://dx.doi.org/10.15490/seek.1.investigation.56