Commentary

OPEN ACCESS

molecular systems biology

The evolution of standards and data management practices in systems biology

Natalie J Stanford[1,2,*], Katherine Wolstencroft[3], Martin Golebiewski[4], Renate Kania[4], Nick Juty[5], Christopher Tomlinson[6], Stuart Owen[2], Sarah Butcher[6], Henning Hermjakob[5], Nicolas Le Novère[7], Wolfgang Mueller[4], Jacky Snoep[8,9] & Carole Goble[2]

Mol Syst Biol. (2015) 11: 851

See also: **T Lemberger** (December 2015)

Introduction

Systems biology involves the integration of multiple heterogeneous data sets, in order to model and predict biological processes. The domain's interdisciplinary nature requires data, models and other research assets to be formatted and described in standard ways to enable exchange and reuse.

Infrastructure for Systems Biology Europe (ISBE) is a project to establish essential, centralized services for systems biology researchers throughout the systems biology lifecycle. A key component of ISBE is to support the management, integration and exchange of data, models, results and protocols. To inform further ISBE development, we surveyed the community to evaluate the uptake of available standards, and current practices of researchers in data and model management.

The survey addressed four key areas as follows:

1 Standards usage;
2 Data and model storage before publication;
3 Sharing in public repositories after publication;
4 Reusability of data, models and results.

The survey was sent to major mailing lists targeting the systems biology and computational biology communities and advertised at relevant consortia meetings. It elicited 153 responses, from 17 countries across 6 continents, with a cross section of the systems biology community represented (Appendix Fig S1). Lessons from the survey are being implemented as part of an ISBE supporting project, FAIRDOM (www.fair-dom.org).

To understand how uptake of standards has developed, we compared our findings to a previous study by Klipp *et al* in 2007. Fig 1 shows a summary of the survey results (detailed results in Dataset EV1). A number of acronyms are used within the text, details of which can be found in Table 1.

Standards usage

Formatting and describing data and models using community standards enables them to be understood, compared, exchanged and reused by both collaborators and the wider community. As such, uptake of standards is vital for high-quality, reproducible research. This is especially true for systems biology which naturally requires frequent exchange of data and models. In systems biology, standards are primarily developed by community standardization initiatives such as COMBINE (Hucka *et al*, 2015), and ISO.

In this study, we consider three major types of standards as follows:

1 Standard formats for representing data and models;
2 Standard metadata checklists for describing particular types of data and models;
3 Controlled vocabularies and ontologies to provide a common notation and annotation vocabulary.

In 2007, Klipp *et al* identified formats, in particular those for encoding models, as the most widely used standards. This is still the case now, with SBML (60%) and SBGN (22%) (Hucka *et al*, 2015) dominating. These standard formats allow easy exchange between software tools and databases, improving (re)usability. The availability and uptake of formats has grown rapidly since 2007. Standards for formatting and visualizing models and for some common experimental data are now available.

Metadata standards—standards for data describing the data—were highlighted as requiring significant development in 2007. There are now over 40 *minimum information* checklists that consistently structure the least amount of information required to interpret a data set. These include common data and model types in systems biology

1 Manchester Institute of Biotechnology, The University of Manchester, Manchester, UK
2 School of Computer Science, University of Manchester, Manchester, UK
3 Leiden Institute of Advanced Computer Science, Leiden Institute of Advanced Computer Science, Leiden University, Leiden, The Netherlands
4 Heidelberg Institute for Theoretical Studies, Heidelberg, Germany
5 European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK
6 Department of Surgery and Cancer, Imperial College London, London, UK
7 Babraham Institute, Cambridge, UK
8 Department of Biochemistry, University of Stellenbosch, Matieland, South Africa
9 School of Chemical Engineering & Analytical Science, The University of Manchester, Manchester, UK
*Corresponding author. Tel: +44 161 275 0145; E-mail: natalie.stanford@manchester.ac.uk
DOI 10.15252/msb.20156053

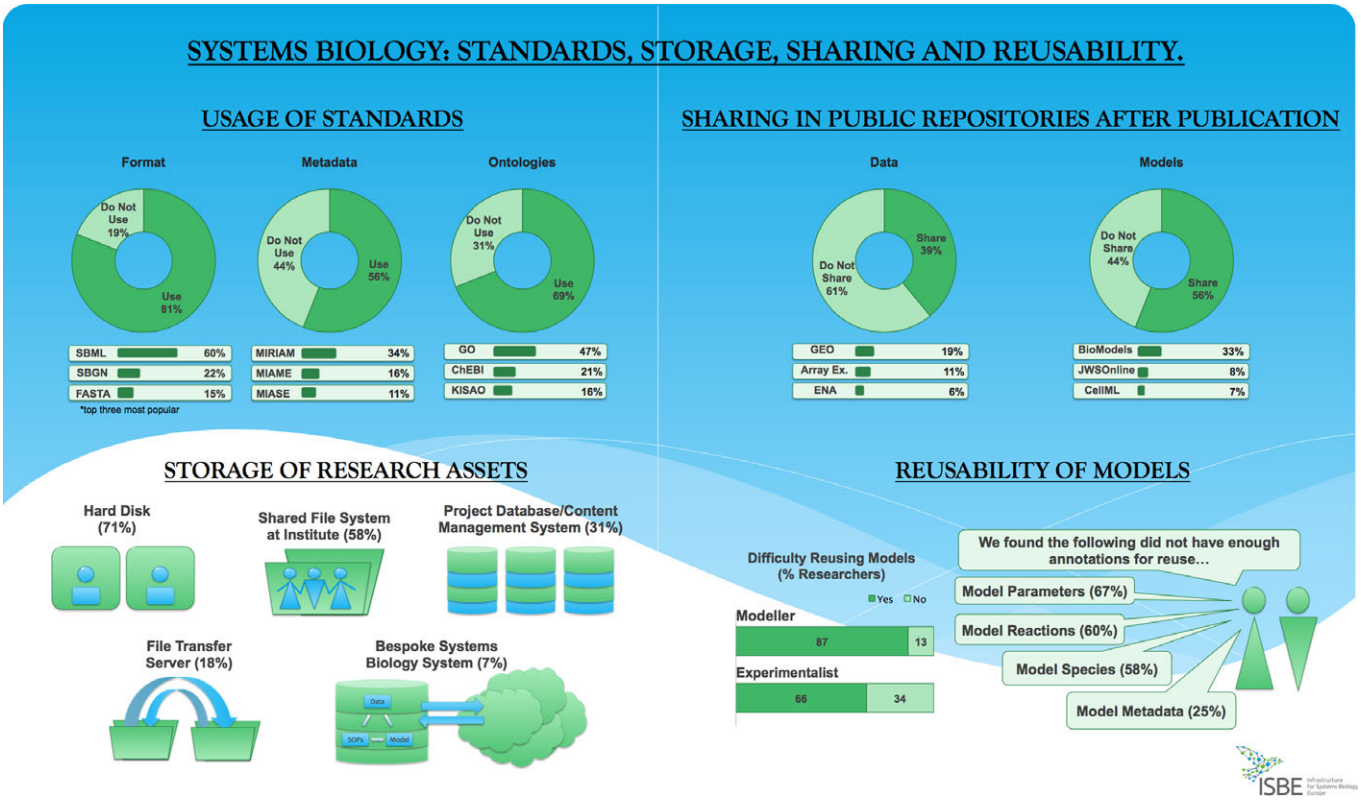

Figure 1. Survey summary.

(see Appendix). MIRIAM (Le Novère *et al*, 2005), MIAME (Brazma *et al*, 2001) and MIASE (Waltemath *et al*, 2011) are the most used by respondents. Ontologies are often used as annotation vocabularies within metadata descriptions. Ontologies for annotating gene functions (GO—47% Ashburner *et al*, 2000), small molecules (ChEBI—21% Hastings *et al*, 2013) and model simulations (KISAO—16% Courtot *et al*, 2011) are the most popular in the community, with growing acceptance since 2007.

Whilst the availability of standards and their growing uptake is encouraging, there is still a dearth of standards for many data types. A priority must be to increase standard availability for common data types not covered. One of the major bottlenecks for uptake is most likely the lack of tools that implement support for standards. If standards compliant results were supported by information management software, it would become part of the research process and thereby reduce the time, knowledge and skills required to achieve compliance, facilitating quicker and more widespread adoption.

Storage of research assets

Systems biology researchers need to exchange experimental data, computer code and models between collaborators within their institute and with distributed, external partners. Despite this exchange being a key activity, the majority of researchers still only store their work on their local hard disc (71%), or shared file systems within their institute (58%). This can make versioning or snapshotting research assets difficult and raises barriers for sharing with collaborators, or, for example, when key personnel leave a team. Content management systems and bespoke systems biology platforms are more amenable to organizing, versioning and sharing, but are only used by 31% and 7% of researchers, respectively. Bespoke platforms require more investment in upload and updating, but provide users with more security for data backup, and offer versioning and easier sharing options.

Sharing in public repositories

Using public repositories is more common to share models (56%) than data (39%). BioModels (Chelliah *et al*, 2015) is the most popular models database (33%)—it is also one of the most popular for finding models after publication (22%). Data are often published in dedicated repositories, grouped by data type (e.g. metabolomics data in a metabolomics database), rather than by function (e.g. all data on human liver). This can make identifying complementary datasets for integration into models difficult, even if the data are well annotated. A major disadvantage for systems biology results is that data sets that were generated from the same samples to address specific biological processes can be separated and submitted to several independent repositories, which results in a loss of experimental context. Some researchers use content aggregator commons, such as SEEK (7%) (Wolstencroft *et al*, 2015), which support functional linking for data and model integration, helping retain experimental context.

Sharing data and models solely through supplementary material in journal articles is still common practice. This represents a publication-centric view of the data, which

means finding related data might be more difficult than it would be when data are submitted to public repositories.

Reusability of models

Being able to reuse data and models in different studies allows a maximized return on research investments. The majority of respondents found it difficult to reuse models and associated data. Model parameters and the traceability of their origins were particularly notable as areas that needed improvement (67% finding issues). These could be improved with better annotation of the original data and better semantic linking of the models to the experimental data that was used to construct them.

Conclusions and outlook

It is clear from the research that we need:

1 Software tools that support standards, thereby facilitating their adoption;
2 Shared/cloud-based platforms to disseminate assets across the community;
3 Annotate and curate assets to enable their meaningful integration;
4 Intimately and persistently, link structured and annotated data and models.

To address the issues above, we suggest that centralized coordinated infrastructures like ISBE, in collaboration with standardization initiatives such as COMBINE, take lead in improving availability, adoption and long-term sustainability of standards. This can be achieved through the training of researchers as well as tool development to support their work flows. The community should also look towards encouraging data and model sharing through incentives such as credit mechanisms and appropriate mandates on practices from journals.

Table 1. Glossary of acronyms.

Acronym	Description	Link
Array Ex.	Array Express—archive of functional genomics data	https://www.ebi.ac.uk/arrayexpress/
BioModels	Database for storing curated and non-curated systems biology computational models	https://www.ebi.ac.uk/biomodels/
CellML	Standard for formatting models, as well as a model repository	https://www.cellml.org/
ChEBI	Chemical Entities of Biological Interest—a dictionary of molecular entities	https://www.ebi.ac.uk/chebi/init.do
COMBINE	Computational Modelling in Biology Network	http://co.mbine.org
ENA	European Nucleotide Archive—a comprehensive record of nucleotide sequences	http://www.ebi.ac.uk/ena
FAIRDOM	Findable Accessible Interoperable Reusable Data standard Operating Procedures and Models	http://fair-dom.org
FASTA	Text-based format for representing nucleotide sequences	https://en.wikipedia.org/wiki/FASTA_format
GEO	Gene Expression Omnibus—repository for functional genomics data	http://www.ncbi.nlm.nih.gov/geo/
GO	Gene Ontology—a controlled vocabulary of gene and gene product attributes	http://geneontology.org/
ISBE	Infrastructure for Systems Biology Europe	http://project.isbe.eu
ISO	International Standards Organization	http://www.iso.org
JWS Online	Tool for online simulation of systems biology models	http://jjj.mib.ac.uk/
KISAO	Kinetic Simulation Algorithm Ontology, for identifying algorithms and associated set-up of simulations	http://co.mbine.org/standards/kisao
MIAME	Minimum Information about a Microarray Experiment	http://fged.org/projects/miame/
MIASE	Minimum Information about a Simulation Experiment	http://co.mbine.org/standards/miase
MIRIAM	Minimum Information Required in the Annotation of Models	http://co.mbine.org/standards/miriam
SBGN	Systems Biology Graphical Notation	http://www.sbgn.org/
SBML	Systems Biology Mark-up Language	http://sbml.org/
SEEK	Bespoke systems biology data management platform, which works as an aggregated content commons, and a database	http://fair-dom.org/SEEK

Expanded View for this article is available online.

Acknowledgements

The paper was supported primarily by the European Union under the Preparatory Phase Projects in the framework of FP7 (project reference 312455). NJS is additionally grateful for funding under grant code BB/M013189/1 (DMMCore), and BBSRC BB/I004637/1 (SysMODB2). MG, RK and WM received additional funding from the German Federal Ministry of Education and Research (BMBF) via grants 031A540A (de.NBI) and FKZ 0315749 (VirtualLiver Network) and the Klaus Tschira Foundation. MG also received funding from the German Federal Ministry for Economic Affairs and Energy (BMWi) via the NormSys project (grant FKZ 01FS14019). JS received funding from NRF-SARCHI-82813. NLN also receives strategic funding from the BBSRC (BBS/E/B/000C0419).

References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25: 25 – 29

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S *et al* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* 29: 365 – 371

Chelliah V, Juty N, Ajmera I, Raza A, Dumousseau M, Glont M, Hucka M, Jalowicki G, Keating S, Knight-Schrijver V, Lloret-Villas A, Natarajan K,

Pettit J-B, Rodriguez N, Schubert M, Wimalaratne S, Zhou Y, Hermjakob H, Le Novère N, Laibe C (2015) BioModels: ten year anniversary. *Nucleic Acids Res* 43: D542 – D548

Courtot M, Juty N, Knüpfer C, Waltemath D, Zhukova A, Dräger A, Dumontier M, Finney A, Golebiewski M, Hastings J, Hoops S, Keating S, Kell DB, Kerrien S, Lawson J, Lister A, Lu J, Machne R, Mendes P, Pocock M *et al* (2011) Controlled vocabularies and semantics in systems biology. *Mol Syst Biol* 7: 543

Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, Muthukrishnan V, Owen G, Turner S, Williams M, Steinbeck C (2013) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res* 41: D456 – D463

Hucka M, Nickerson D, Bader G, Bergmann F, Cooper J, Demir E, Garny A, Golebiewski M, Myers C, Schreiber F, Waltemath D, Le Novère N (2015) Promoting coordinated development of community-based information standards for modeling in biology: the COMBINE initiative. *Front Bioeng Biotechnol* 3: 19

Klipp E, Liebermeister W, Helbig A, Kowald A, Schaber J (2007) Systems biology standards – the community speaks. *Nat Biotechnol* 25: 390 – 391

Le Novère N, Finney A, Hucka M, Bhalla US, Campagne F, Collado-Vides J, Crampin EJ, Halstead M, Klipp E, Mendes P, Nielsen P, Sauro H, Shapiro B, Snoep JL, Spence HD, Wanner BL (2005) Minimum Information Requested In the Annotation of biochemical Models (MIRIAM). *Nat Biotechnol* 23: 1509 – 1515

Waltemath D, Adams R, Beard DA, Bergmann FT, Bhalla US, Britten R, Chelliah V, Cooling MT, Cooper J, Crampin E, Garny A, Hoops S, Hucka M, Hunter P, Klipp E, Laibe C, Miller A, Moraru I, Nickerson D, Nielsen P *et al* (2011) Minimum Information About a Simulation Experiment (MIASE). *PLoS Comput Biol* 7: 4

Wolstencroft K, Owen S, Krebs O, Nguyen Q, Stanford NJ, Golebiewski M, Weidemann A, Bittkowski M, An L, Shockley D, Snoep JL, Mueller W, Goble C (2015) SEEK: a systems biology data and model management platform. *BMC Syst Biol* 9: 33