Beilstein-Institut

Experimental Standard Conditions of Enzyme Characterizations,
September 13th – 16th, 2009, Rüdesheim/Rhein, Germany

Developing Coherent Minimum Reporting Guidelines for Biological Scientists: The MIBBI Project

Chris F. Taylor

European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, U.K.

E-Mail: chris.taylor@ebo.ac.uk

Received: 17th March 2010 / Published: 14th September 2010

Abstract

Modern biological science addresses a variety of subjects using an array of analytical techniques. Few relations between subject and technique are exclusive, making for a very large number of potential workflows, combinatorially-speaking. While this diversity is to be celebrated, it presents informatics challenges that require resolution if the data-sharing ambitions of many funders are to be realised, and the consequent benefits to science obtained. There is increasingly-organised movement towards consensus on data and reporting standards for the biosciences, but significant hurdles remain: scientists must be convinced of the value of the exercise, and user-friendly, time-efficient and robust tools are required. The 'Minimum Information for Biological and Biomedical Investigations' (MIBBI) Project, which promotes and develops guidance on the content that experimental reports should contain, is dependent on progress in both these areas.

Introduction

In recent years the suffix 'omics' has come rather too much into use. While conveying some information about the intent of an investigation (for example, 'proteomics' indicating the study of the protein complement of a [biological] sample), the various omics-terminated terms came more commonly to connote sets of associated technologies (for proteomics, mass spectrometry). For reasons both sociological and technological, nascent societies asso-

ciated with these emerging omics fields spawned standards bodies tasked to smooth out the wrinkles in their data sharing landscape, but those standards bodies' scopes overlapped (for example, mass spectrometry has many uses beyond proteomics), resulting in redundancy of effort. Coordination became essential; not just amongst omics fields, but with the wider bioscience community, within which similar efforts were progressing for a variety of areas, focused on particular fields/workflows.

Given the innumerable combinations of areas of enquiry and techniques of investigation within the biosciences, an appropriate approach is to develop modular solutions (*i.e.*, separable parts; whether of a vocabulary, format or set of guidelines). The Ontology of Biomedical Investigations (OBI; http://purl.obolibrary.org/obo/obi), and at a higher level the Open Biomedical Ontologies project ([1]; http://obofoundry.org/) of which it is part, reflect this approach; as do the structures of data exchange formats such as FuGE [2] and ISA-Tab [3]. Such modules require a general framework to define and contextualise them; if that framework is shared between resources their combined use is simplified. The framework used by an increasing number of projects, including ISA-Tab and the MIBBI Foundry (described later) consists most importantly of three concepts: *Investigation*, *Study* and *Assay* (Figure 1). In this simple framework, an *Investigation* is defined as a body of work consisting of one or more *Studies* (which normally centre on a biological question or a particular source material), each of which may contain one or more *Assays* (usually some kind of data-generating analysis).

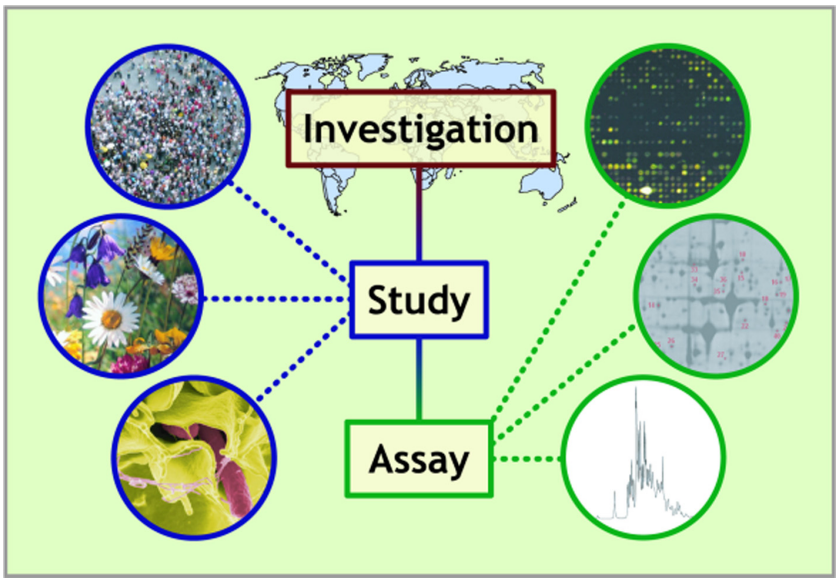

Figure 1. A graphical representation of the Investigation/Study/Assay (ISA) hierarchy, used as a basic framework for the data structures underlying an increasing number of standardisation projects such as MIBBI and ISA-Tab.

On Coordination

There are various efforts to coordinate the development of standards (OBI/OBO, MIBBI, FuGE, ISA-Tab, *etc*.; [4]). The situation is still in flux, but promises to continue to settle down as patterns and precedents become more established. In recent years, one significant driver has been policy development by funding agencies. Their various statements and regulations relating to data sharing [5] all parallel statements made by the Organisation for Economic Cooperation and Development (OECD; http://oecd.org/), which holds that publicly-funded research data are a public good, produced in the public interest, and should be openly available to the maximum extent possible.

Broadly, there are three kinds of standard: data formats such as ISA-Tab; vocabularies/ontologies such as those in OBO; and Minimum Information (MI) checklists, which are the focus of this paper. MI checklists are guidance documents specifying the information that should be provided when reporting research work. Historically, these have been developed by groups working within particular biological or technological domains. The resulting fragmentation made it difficult to obtain an overview of ongoing projects, or to track progress. Furthermore, having been developed independently, checklists were frequently partially redundant against each other, and arbitrary decisions on wording and substructuring in the overlaps ensured that integration would be laborious. This made it impractical for potential users to combine parts of checklists to make guidance documents appropriate for their own workflow.

The MIBBI Project

The Minimum Information for Biological and Biomedical Investigations (MIBBI) Project is an international collaboration between communities developing MI checklists. In addition to its general goals of promoting the use of MI checklists to funders, publishers and the community, and of promoting the development of checklist-supportive tools and databases, the project has two specific goals, with two corresponding sections on the website:

The MIBBI Portal (http://mibbi.org/index.php/MIBBI_portal) exists as an organising point for the many MI checklist projects underway in the community. As a 'shop window' for such projects, the MIBBI Portal increases the likelihood that users and potential contributors will discover a project appropriate to their needs. It also acts as a rallying point for new projects, bringing them into the 'checklist community', encouraging collaboration.

The MIBBI Foundry (http://www.mibbi.org/index.php/MIBBI_foundry) takes the overlapping, hard-to-integrate contents of the various Portal checklists and reworks them into a suite of checklist modules that can be reassembled to fit any number of workflows. A browse interface (http://mibbi.org/index.php/MICheckout) allows users to specify their areas of interest and download compiled sets of these modules, as HTML (for browsing), as XML

Schema (for model-driven software, *inter alia*), as a tab-delimited file (simple spreadsheet), or as a configuration for ISAcreator (http://isatab.sourceforge.net/), for data collection/management/submission.

Ultimately, we envision a situation where different communities 'own' parts of the MIBBI Foundry's content, with shared ownership over common features such as 'organism', 'person', or 'environment'. This body of guidance then becomes a unified expression of the information required when reporting bioscience research, which should then be referenced by software and database developers, viewing it as their driving 'use case'.

The Objections to Fuller Reporting

The potential benefits of the widespread sharing of well-annotated data sets are significant; the extraction of further value from existing data; the re-use of data in meta studies; the maintenance of a solid evidence base for claims made in the literature; and so on. However, this is not the whole picture. For those who generate the data, whose worth is assessed through the value they extract from it while they have sole access, and who must invest the time to encode, describe and share it, things can look rather different. Although most are governed by funding conditions that mandate data sharing, such strictures rarely obtain greater than the minimum of effort; 'data generators' must be positively encouraged to share their data.

The arguments against data sharing are of four kinds: (1) that the loss of intellectual property reduces the chances for researchers to maximally exploit their data; (2) that neither the money nor the tools are available to support data sharing; (3) that other's shared data will potentially be of low quality, a problem exacerbated by the difficulty of *post hoc* quality assessment; (4) that others (especially bioinformaticians) will benefit from the investment in the shared data without credit accruing to the originator. However, all of these issues can or have been addressed.

Under the policy of most funders, intellectual property is protected for a period of years after data are generated, allowing for most value to be extracted before the data are shared with the wider community. It is also fairly normal these days to apply for funds for data management as part of a funding application; and as the standards that will support data sharing settle down, so we have seen tools begin to emerge. The issue of data quality is a difficult one. Many funders require that all data are shared, including data that were not the subject of a publication, and therefore have not been peer-reviewed. The solution will most likely come from databases, who will (and in some cases already do [6]) use metadata and statistical analyses to assess the likely quality of a data set, but even data of moderate quality can be useful if suitably annotated (data processing techniques may improve, for example).

The objection that no credit accrues from data reuse is challenging. To acknowledge one's peers is central to the scientific process, yet the re-use of downloaded data is often unacknowledged. Even where it is acknowledged, there is normally no way for standard assessments of academic performance to count the 'impact' of such a thing. The solution is firstly to make data sets 'citeable' through the use of Digital Object Identifiers (DOIs), secondly to ensure that referees are as ardent about missing citations for data sets used as for papers, and lastly to ensure that literature databases, funders and faculties count these as the equivalent of normal citations [7]. A side benefit of making data sets citeable is that the quality of annotation may improve further: if the re-use of a data set is governed by its utility, then it follows that higher quality annotation, by improving utility (both because the data set is better characterised, and because the additional annotation may increase the chance of orthogonal re-use) will cause the data set to be re-used more frequently, with more credit accruing to the generator of those data.

Summary

The MIBBI Project seeks to coalesce fragmented efforts to develop MI checklists, and ultimately to provide a unified set of checklist modules by re-using the work of individual communities. This parallels unification efforts for other kinds of standard (formats, vocabularies) and complements the desire of funders and publishers for increased data sharing. However, the success of the MIBBI Project, and of efforts to increase data sharing generally, depend on the instigation of a system to uniquely identify data sets, on the use of that system by those who assess the worth of researchers, and on the availability of appropriate, robust tools.

References

[1] Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., OBI Consortium, Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* **25**(11):1251 – 5.
doi: http://dx.doi.org/10.1038/nbt1346.

[2] Jones, A.R., Miller, M., Aebersold, R., Apweiler, R., Ball, C.A., Brazma, A., Degreef, J., Hardy, N., Hermjakob, H., Hubbard, S.J., Hussey, P., Igra, M., Jenkins, H., Julian, R.K. Jr, Laursen, K., Oliver, S.G., Paton, N.W., Sansone, S.A., Sarkans, U., Stoeckert, C.J. Jr, Taylor, C.F., Whetzel, P.L., White, J.A., Spellman, P., Pizarro, A. (2007) The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics. *Nat. Biotechnol.* **25**(10):1127 – 33.
doi: http://dx.doi.org/10.1038/nbt1347.

[3] Sansone, S.A., Rocca-Serra, P., Brandizi, M., Brazma, A., Field, D., Fostel, J., Garrow, A.G., 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., Members of the RSBI Working Group. (2008) The first RSBI (ISA-TAB) workshop: ''can a simple format work for complex studies?''. *OMICS* **12**(2):143 – 9. doi: http://dx.doi.org/10.1089/omi.2008.0019.

[4] Taylor C.F. (2007) Standards for reporting bioscience data: a forward look. *Drug Discov Today* **12**(13 – 14):527 – 33. doi: http://dx.doi.org/10.1016/j.drudis.2007.05.006.

[5] Field, D., Sansone, S.A., Collis, A., Booth, T., Dukes, P., Gregurick, S.K., Kennedy, K., Kolar, P., Kolker, E., Maxon, M., Millard, S., Mugabushaka, A.M., Perrin, N., Remacle, J.E., Remington, K., Rocca-Serra, P., Taylor, C.F., Thorley, M., Tiwari, B., Wilbanks, J. (2009) Megascience. 'Omics data sharing. *Science* **326**(5950):234 – 6. doi: http://dx.doi.org/10.1126/science.1180598.

[6] Kapushesky, M., Emam, I., Holloway, E., Kurnosov, P., Zorin, A., Malone, J., Rustici, G., Williams, E., Parkinson, H., Brazma, A. (2010) Gene expression atlas at the European bioinformatics institute. *Nucleic Acids Res.* 38(Database issue): D 690 – 8.

[7] Thorisson, G.A. (2009) Accreditation and attribution in data sharing. *Nat. Biotechnol*. **27**(11):984 – 5. doi: http://dx.doi.org/10.1038/nbt1109-984b.