Towards Interoperable Reporting Standards for Omics Data:
Hopes and Hurdles
Susanna-Assunta Sansone\textsuperscript{1,2,\$}, Philippe Rocca-Serra\textsuperscript{1,2,3,\$}, Dawn Field\textsuperscript{2,4}, Chris F Taylor\textsuperscript{1,2}, Weida Tong\textsuperscript{5,\$}, Marco Brandizi\textsuperscript{1}, Eamonn Maguire\textsuperscript{1}, Nataliya Sklyar\textsuperscript{1}

\textsuperscript{1} European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK.
\textsuperscript{2} Natural Environment Research Council Environmental Bioinformatics Centre, Mansfield Road, Oxford OX1 3SR, UK.
\textsuperscript{3} NuGO, The European Nutrigenomics Organisation.
\textsuperscript{4} Molecular Evolution and Bioinformatics Group, Oxford Centre for Ecology and Hydrology, Mansfield Road, Oxford OX1 3SR, UK.
\textsuperscript{5} Center for Toxicoinformatics, National Center for Toxicological Research, US Food and Drug Administration, 3900 NCTR Rd., Jefferson, AR 72079, USA.
\$ Corresponding authors: sansone@ebi.ac.uk, rocca@ebi.ac.uk
\$ Disclaimer: \textit{The views presented in this article do not necessarily reflect those of the US Food and Drug Administration}

Abstract

Background

As the size and complexity of scientific datasets and the corresponding information stores grow, standards for collecting, describing, formatting, submitting and exchanging information are playing an increasingly active role. Several initiatives occupy strategic positions in the international scenario, both within and across domains. However, the job of harmonising reporting standards is still very much a work in progress; both software interoperability and the data integration remain challenging as things stand.

Results

The status quo with respect to standardization initiatives is summarized here, with particular emphasis on the motivation for, and the challenges of, ongoing synergistic activities amongst the academic community focused on the creation of truly interoperable standards.

Conclusions

Groups generating standards should engage with ongoing cross-domain activities to simplify the integration of heterogeneous data sets to the greatest possible extent.

Background

The growing complexity of datasets

In the area of life science, the cycle of data generation and processing is being vastly accelerated by the development of high-throughput experimental methods associated with genomic and post-genomic technologies (e.g. genomics, transcriptomics, proteomics, and metabolomics, hereafter referred as ‘omics’). Biological and biomedical studies commonly range from simple one assay-based to complex multi-assay studies. For the latter type, for example, consider the reporting of a study looking at the effect on a number of subjects treated with different drugs by characterizing the metabolic profile of their urine (by mass spectroscopy), measuring protein and gene expression in the liver (by mass spectrometry and DNA microarrays, respectively), and conducting conventional histological analysis. Omics studies are information intensive and to record their complex structure it is necessary to define and capture the experimental metadata, including experimental design, sample source(s) and treatment(s), the preparation of the sample for the analytical assay, the processes and instruments used throughout, and the final data. It is widely recognized that capturing experimental metadata on
this level of granularity is required to correctly interpret the results that they contextualize and enable efficient data sharing.

**Standardization Initiatives Focused on Particular Domains of Application**

Many groups have risen to this challenge; standards for collecting, describing, formatting, submitting and exchanging both the data and metadata from such complex studies either are under development or have been released [1]. Currently, several standards initiatives occupy strategic positions in the international scenario, largely falling into two groups identifiable by the needs of their respective user communities.

One group of initiatives is driven by regulatory frameworks, and often supported by accredited (*de jure*) Standards Developing Organizations (SDOs). Most significantly, these efforts focus on the Voluntary eXploratory Data Submissions (VXDS) and electronic data submission programs of the US Food and Drug Administration (FDA) [2-4] or around initiatives by other governmental agencies, such as the US Environmental Protection Agency (EPA) [5]. These initiatives also include long-standing efforts in the clinical and non-clinical domains [6] alongside more recent activities in the pharmacogenomics area that add complex omics technologies to biomedical studies [7].

A second group of initiatives that address particular (omics or other) technologies or defined domains of application (*e.g.* system biology, pathways) have emerged from the academic community, in many cases with the support of commercial organizations such as instrument vendors. Such initiatives (*e.g.*, [8-14]) are focused on supporting tool interoperability and data exchange among public and proprietary systems through the development of three kinds of (*de facto*) reporting standards: ‘minimum information’ checklists (*specifications of data set content, however encoded*), ontologies (*semantics*) and file formats (*syntax*). Minimal information checklists are easy-to-read, structured documents that reflect the consensus view of the essential pieces of information that should be reported; ontologies provide the semantics needed to describe the minimal information requirements and file formats the syntax to transmit and exchange these. Combining these three kinds of reporting standards a submission tool, for example, should guide researchers through the process of meeting the reporting requirements made by a given minimal information specification, enable straightforward practical use of ontology terms and export the collected information in a standard format to a given database.

**Fragmentation of Standards**

Domain-specific initiatives are regarded as important because they address ‘real world’ data reporting requirements. Unfortunately, focusing on particular communities’ interests or technologies leads to duplication of effort. More seriously, the development of (largely arbitrarily) different standards severely hinders data integration. Nowadays researchers are able to perform multi-assay studies where the same sample is run through the full range of ‘omics and conventional technologies, in combination. In this specific case, it is critical that the reporting standards are designed to be interoperable and fit neatly into a jigsaw, with users being able to take the pieces that are relevant to report their study. The fragmentation severely hinders the interoperability of databases and tools, implementing such reporting standards: this scenario is illustrated by the ArrayExpress [15] and PRIDE [16] – two EBI public repositories for transcriptomics and proteomics data respectively. These systems implement (non-interoperable) standards applicable only for their ‘omics’ domains. Consequently, users have to deal with different submission formats, diverse representations of the metadata and terminologies when depositing their datasets in these systems, and similarly when downloading other datasets. Such fragmentation has a strong impact on the user community, particularly by hampering deposition and integration of multi-assay studies.

**Results and Discussion**

**Integrative Cross-Domain Standardization Initiatives**

Fortunately, amongst the academic community a number of initiatives aim to foster the harmonization and consolidation of the three kinds of reporting standard previously described

**Content:** Twenty-seven groups now participate in the Minimum Information for Biomedical or Biological Investigations (MIBBI) project, which offers a one-stop shop for those exploring the range of extant ‘minimum information’ checklists (*such as MIAME [17]*) and fosters their collaborative, integrative development [18].

**Semantics:** More than 70 groups participate in the Open Biological Ontology (OBO) Foundry. The objective of the project is to encourage the development of orthogonal, interoperable ontologies [19].

**Syntax:** Several groups participate in the Functional Genomics (FuGE) project to develop a single generic data model that will underpin a variety of XML-based file formats by providing a single common framework [20]. Recently, a complementary initiative has been begun by a
Consistent reporting has a positive and long-lasting impact on the value of collective scientific outputs. This has also been recognized by funding agencies that are increasingly playing an active role in the strategic stewardship of omics data, often through the development of data policies encouraging the use of (existing) standards and public standards-compliant repositories for data collection and management [29].

Conclusions

This paper has illustrated the growing number of standards and the complexity facing those attempting to use them, for example, to report or integrate datasets from multiple domains. We have also indicated the existence of a number of synergistic projects seeking to simplify the process of integrating reporting standards, where possible. Of course, this is not an exhaustive list; several coordinative infrastructure initiatives work to address the problem of sharing and archiving large amounts of data, according to common standards (e.g., [30-32]).

There are many benefits accruing to the development and acceptance of reporting standards. For example, by limiting the range and variability of standards, the development and maintenance costs for commercial and academic software developers of standards-compliant products comes down. This results in more appropriate resources for the biomedical and scientific community, making the job of capturing, annotating, integrating, sharing and exploiting (meta)data simpler, increasing the prima facie value of the data to others (secondary users), and by extension, increasing the return on the investment of (public) funds that supported their generation.

Above all actions a ‘top-down’ coordination is needed to help bringing these standardization efforts closer to address the fragmentation issue. Although, regulatory- or biomedical-driven initiatives have far stricter guidelines than academia, much could be learned from exchange of ideas and practices of these sectors.

References

[1] Field D, Sansone S-A:. A special issue on Omic data standards. OMICS 2006, 10(2): 84-93.
[2] US HHS/FDA Guidance for Industry: Pharmacogenomic data submissions. (2005). [http://www.fda.gov/ohrms/dockets/98fr/2003d-0497-gdl0002.pdf]
[3] Frueh FW: Impact of microarray data quality on genomic data submissions to the FDA. Nat Biotechnol 2006, 24(9):1105-1107.
[4] Tong W, Harris SC, Fang H, et al. An Integrated Bioinformatics Infrastructure
Essential for Advancing Pharmacogenomics And Personalized Medicine in the Context of the FDA's Critical Path Initiative. Drug Discovery Today 2007, 4 (1):3-8.

US EPA Potential Implications of Genomics for Regulatory and Risk Assessment Applications at EPA. (2004). [http://www.epa.gov/osa/genomics.htm]

CDISC Standards pages. http://www.cdisc.org/standards/

Shabo A: Clinical genomics data standards for pharmacogenetics and pharmacogenomics. Pharmacogenomics 2006, 7(2):247-53.

Lindon JC, Nicholson JK, Holmes E, et al. Summary recommendations for standardization and reporting of metabolic analyses. Nature Biotechnol 2005, 23(7):833-8.

Le Novère N, Finney A, Hucka M, et al. Minimum information requested in the annotation of biochemical models (MIRIAM). Nature Biotechnol 2005, 23(12):1509-15.

Ball CA, Brazma A: MGED standards: work in progress. OMICS 2006, 10(2):138-44.

Sansone SA, Fan T, Goodacre R, et al.: The metabolomics standards initiative. Nature Biotechnol 2007, 25(8):846-8.

Taylor CF, Paton NW, Lilley KS, et al. The minimum information about a proteomics experiment (MIAPE). Nature Biotechnol 2007, 25(8):887-93.

Deutsch EW, Ball CA, Berman JJ, et al. Minimum Information Specification For In Situ Hybridization and Immunohistochemistry Experiments (MISFISHIE). Nature Biotechnol 2006, 26(3):305-12.

Field D, Garrity G, Gray T, et al. The minimum information about a genome sequence (MIGS) specification. Nature Biotechnol 2008, 26(5):541-7.

ArrayExpress: http://www.ebi.ac.uk/arrayexpress

PRIDE: http://www.ebi.ac.uk/pride/

Brazma A, Hingamp P, Quackenbush J, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nature Genet 2001, 29(4):365-71.

Taylor C, Field D, Sansone SA, et al. Promoting Coherent Minimum Reporting Requirements for Biological and Biomedical Investigations: The MIBBI Project. Nature Biotechnol 2008, 26(8):889-96. [http://www.mibbi.org/]

Smith B, Ashburner M, Rosse C, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnol 2007, 25(11):1251-5. [http://www.obofoundry.org/]

Jones AR, Miller M, Aebersold R, et al. The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics. Nature Biotechnol 2007, 25(10):1127-33. [http://fuge.sf.net/]

Sansone SA, Rocca-Serra P, Brandizi M, et al. The First RSBI (ISA-TAB) Workshop: “Can a Simple Format Work for Complex Studies?” OMICS 2008, 12(2):143-9. [http://isatab.sf.net/]

Study Data Tabulation Model. [http://www.cdisc.org/models/sdtm]

BioInvestigation Index: [http://www.ebi.ac.uk/bioinvindex]

ISAcreator: http://isatab.sourceforge.net/isacreator.html

Workshops on reporting standards: [http://www.ebi.ac.uk/net-project/projects.html#workshop]

Editorial. Standard operating procedures: Nature Biotechnol 2006, 24(11):1299.

Editorial. Democratizing proteomics data: Nature Biotechnol 2007, 25:262.

Editorial. Standardizing data: Nature Cell Biol 2008, 1123-1124

List of funding agency data policies: http://biosharing.org/2009/01/data-policies.html

CaBIG: http://cabig.cancer.gov

BIRN: http://www.nbirn.net/

ELIXIR: http://www.elixir-europe.org