OPEN ACCESS

A sea of standards for omics data: sink or swim?

Jessica D Tenenbaum,[1] Susanna-Assunta Sansone,[2] Melissa Haendel[3]

[1]Duke Translational Medicine Institute, Duke University, Durham, North Carolina, USA
[2]Oxford e-Research Centre, University of Oxford, Oxford, UK
[3]Library and Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland, Oregon, USA

Correspondence to
Dr Jessica D Tenenbaum, Duke Translational Medicine Institute, Duke University, PO Box 17969, Durham, NC 27715, USA;
jessie.tenenbaum@duke.edu

Received 31 May 2013
Revised 8 July 2013
Accepted 10 September 2013
Published Online First 27 September 2013

To cite: Tenenbaum JD, Sansone S-A, Haendel M. *J Am Med Inform Assoc* 2014;**21**:200–203.

ABSTRACT

In the era of *Big Data*, omic-scale technologies, and increasing calls for data sharing, it is generally agreed that the use of community-developed, open data standards is critical. Far less agreed upon is exactly which data standards should be used, the criteria by which one should choose a standard, or even what constitutes a data standard. It is impossible simply to choose a domain and have it naturally follow which data standards should be used in all cases. The 'right' standards to use is often dependent on the use case scenarios for a given project. Potential downstream applications for the data, however, may not always be apparent at the time the data are generated. Similarly, technology evolves, adding further complexity. Would-be standards adopters must strike a balance between planning for the future and minimizing the burden of compliance. Better tools and resources are required to help guide this balancing act.

BACKGROUND

Members of the scientific community are increasingly expected to share data, and to do so in a standards-compliant manner. This is evidenced by the recent mandates, announcements, and requests for information by the funding agencies[1–5] and journals,[6] and numerous essays and announcements by the scientific community,[7–10] including pre-competitive initiatives by the life science industry.[11] However, the scientific community is not necessarily well poised to comply.[12] All stakeholders—funders, journal editors, researchers and those supporting them, struggle to navigate the existing standards and make informed decisions.[13] As an example, in 2009 one of our groups aimed to create a standards-compliant, integrated data repository for clinical and 'omics' data, among other types. This begged the question: with which standards should we comply? Through subsequent efforts to answer this question, three key points have become clear:

1. Different groups and individuals have different definitions for what constitutes a 'data standard'.
2. Even within one domain, no one standard is the 'right' standard across all cases; rather, one must select a standard (or even specific pieces of a standard) based on one's particular needs.
3. Integrated resources and registries are needed to help researchers navigate the fluid standards landscape and to choose and implement the right standard for their respective project.

The focus for that project was on omics data standards, but these points apply across the spectrum of biomedical data types. High-dimensional '*Big Data*' equate to large numbers of parameters, which in turn require yet more data for sufficient statistical power. Importantly, this massive amount of data lends itself to many different analytic approaches, putting comprehensive analysis beyond the capabilities of any one researcher. The size and complexity of these data, combined with growing scarcity of research funding and the quest for personalized medicine, make it increasingly important to maximize the utility of research dollars through data sharing and re-use. Efforts to this end are demonstrated by a spate of new data sharing and aggregation initiatives by academics, private–public partnerships, and publishers, for example Sage Bionetworks,[14] the Pistoia Alliance (http://www.pistoiaalliance.org) and DRYAD,[15] among others.[16–18] At the national level in the USA, the data sharing trend is reflected in programs such as the National Institutes of Health's (NIH) recently announced '*Big Data to Knowledge*' (BD2K) initiative,[19] and the White House office of science and technology policy's recent directive that the results of government-funded research be made publicly available.[20] The Innovative Medicines Initiative (http://www.imi.europa.eu/) is Europe's largest public–private initiative that supports collaborative research projects and builds networks of industrial and academic experts in order to boost pharmaceutical innovation in Europe. Internationally, the Research Data Alliance (https://rd-alliance.org/) has been established by an international steering group from funding agencies in the USA, EU and Australia; and recently the global alliance for genomic and clinical data sharing has brought together over 70 leading healthcare, research, and disease advocacy organizations, involving researchers from more than 40 countries, to enable secure sharing of genomic and clinical data.[21]

These types of initiatives, together with the evolving portfolio of grass-roots standards, have enhanced the need to maximize awareness and discoverability of standards. Such efforts are becoming more common,[22–26] but they lack integration or unification. There is a clear need for some level of coordination, without taking the form of a top-down authority. How can we avoid requiring would-be standard adopters to spend considerable time and effort becoming well versed with a multitude of standards solely in order to rule most of them out?

WHAT IS A DATA STANDARD?

The International Organization for Standardization defines a standard as '…a document that provides requirements, specifications, guidelines or characteristics that can be used consistently to ensure that materials, products, processes and services are fit for their purpose'.[27] Standards range from de jure, that is, ordained by some official organization such as the International Organization for Standardization or the American National Standards Institute, to de facto, that is, developed by grass-root initiatives and commonly adopted, but not prescribed by an official or specific authority. The BioSharing registry (http://biosharing.org/) houses a fairly comprehensive,

curated list of data standards (primarily de facto) in the life science, environmental, and biomedical space. These standards are divided into three categories. First, content standards take the form of reporting guidelines, for example, minimum information checklists. These vary from general guidance to itemized prescriptions of the information that should be provided (ie, curation guidelines), including both data and metadata. The second category consists of syntax standards in the form of representations and formats that facilitate the exchange of information. These fall broadly into two types: delimited text, or a 'markup language' such as XML. Third are the semantic standards in the form of terminology artifacts, such as controlled vocabularies or ontologies. These add an interpretive layer to the data by defining the concepts or terms in a domain, and in some cases the relationships between them.

Other discussions of standards include the notion of a data model, which extends beyond terms and their definitions to describe the relationships between concepts in a domain.[28] Other groups also use additional terms such as conceptual model, conceptual schema, ontology, or domain analysis model,[29–32] but generally differ on what each of these terms means. This is in fact part of the confusion—even data standard experts do not agree on what constitutes a data standard. Nevertheless, focusing just within the context of transcriptomics, preliminary investigation yielded a list of 15 potentially relevant standards (table 1). Note that this list could grow depending on the type of sample and organism used, as many terminologies are species specific. Now imagine if a researcher has an associated dataset from a proteomics investigation, for example. How is a mere mortal to sort through these?

FIT FOR PURPOSE

In biomarker discovery, the phrase 'fit-for-purpose' refers to the notion that the degree of rigor for assay validation should be tailored to the intended purpose of a given biomarker study.[33] The same is true for data standards adoption. While each individual project will inevitably have its own specific requirements, it can be useful to group projects across a spectrum of rigor. At the lowest level, there is the use case of data sharing within a laboratory or between collaborators. While minimum information guidelines should be followed, for the most part any documentation need only be human readable, and issues requiring clarification are merely a walk down the hall or an e-mail away (at least until the student graduates or the postdoc moves on). Data that are to be shared publicly, for example, accompanying a publication, require more rigor. Ideally, a prospective consumer of the data can both understand and reproduce those data without needing to contact the original author. Furthermore, much of the content of publications is now aggregated and curated by various online resources. These value-added services can be much more efficient and effective at making content available via secondary sources when quality data standards are used. Minimally structured data can be very helpful for such purposes; for example, the use of a unique identifier to describe a molecule or a standardized vocabulary term to denote the disease area under study. The highest level of rigor is needed for contribution of data to a structured data repository. In this case, additional effort is warranted in the form of structured fields and a standardized, machine-readable format. Such rigor enables querying across multiple datasets and integrative meta-analysis combining more than one set.

One key point in differentiating between these levels of rigor is that there are different 'flavors' of annotation. At every level, there is a difference between what needs to be documented, and what needs to be documented in a structured and queryable fashion. While the option exists to select a standard that allows for maximum structure and adopt it only loosely, complexity can turn off would-be standards adopters, as well as waste time in development if such rigor will ultimately never be needed.

Categories of criteria to be used in evaluating data standards for adoption include:

Table 1 A sampling of (some of the) standards related to microarray-based transcriptomics, generated by non-experts for evaluation of relevance to a project involving microarray-based transcriptomics data

Standard	Type	Description
MIAME	Reporting guideline	Minimum Information About a Microarray Experiment Specifies six components that must be included to describe a microarray experiment, for example, raw and processed data, experimental design, sample annotation, protocols. MIAME does not specify how these components must be represented, for example, in any given format, or using any given terminology
ISA-TAB	Exchange format	Generic format for experimental representations; conversion tools to MAGE-Tab, MIMiML and other formats exist
MAGE-TAB	Exchange format	MicroArray and Gene Expression-Tabular Simple tab-delimited, spreadsheet-based format. Used by ArrayExpress
MAGE-ML	Exchange format	MicroArray and Gene Expression-Markup Language. No longer supported
SOFT	Exchange format	Simple Omnibus Format in Text. Line-based, plain text format designed for rapid batch submission of data. Used by GEO
MIMiML	Exchange format	MIAME Notation in Markup Language. Optimized for microarray and other high-throughput molecular abundance data Used by GEO
GO	Terminology artifact	Gene Ontology. Controlled vocabulary for annotation of gene function and cellular location. Part of the OBO Foundry
EFO	Terminology artifact	Experimental Factor Ontology. Provides a systematic description of many experimental variables. Used by ArrayExpress
OBI	Terminology artifact	Broader scope for experimental representations. Part of the OBO Foundry
MGED Ontology	Terminology artifact	Integrated in OBI
MAGE-OM	Object model	MicroArray and Gene Expression—Object Model. The object model from which MAGE-ML was derived
FuGE	Object model	Generic object model for functional genomics
SEND	Exchange format	Standard for Exchange of Nonclinical Data—an implementation of the CDISC (Clinical Data Interchange Standards Consortium) SDTM (Standard Data Tabulation Model)
GEML	Exchange format	These three standards have since been deprecated and/or replaced by other standards, but that progression may not always be clear to novice users
FUGO	Terminology artifact	
MAML	Exchange format	

- The standard itself
  - specification documentation
  - ease of implementation (eg, level of documentation, requirement for programmer support)
  - human and machine readability
  - formal structure
  - expressivity—the breadth of information that can be represented
  - ease of use, for example, minimal required fields, text-based interface familiarity to biologists.
- Adoption and user community
  - broad adoption and implementation, outside the initial group
  - support supplied by the user community
  - use by community databases
  - software development that supports the standard (eg, for curating, submitting to databases)
  - responsiveness to community requests
  - availability of examples of use
  - requirements of relevant authoritative bodies, for example, funders (NIH, National Science Foundation, Centers for Medicare & Medicaid Services), publishers, etc.
- Additional factors
  - integration/compatibility with other standards
  - extensibility and flexibility to cover new domains
  - conversion and mapping, when applicable
  - cost (eg, open vs licensing fee).

Of course, specific projects may have additional criteria to add, and different projects will place different weight on the different items. Unfortunately, standards adoption, when it happens, is often determined less by an objective criteria-based evaluation and more based on historical precedent ('my advisor used standard X'), marketing ('I saw a press-release about standard X') or sociopolitical circumstance ('I know someone on the standard X team'). What makes it even more difficult to select standards empirically, based on objective criteria, is that standards are often complex. Even well-documented standards can be dense and impenetrable to prospective users who were not involved in their development. This is one reason why standards are often duplicated or reinvented. Other factors include the desire for some level of control, or recognition for doing the work.

RESOURCES WANTED

The recent data and informatics working group report to the advisory committee to the director of the NIH included recommendations to establish a minimal metadata framework for data sharing, and to create catalogs and tools to facilitate data sharing.[2] A truly minimal set of metadata elements is important if we are to have any hope of compliance because the activation energy required for data curation and annotation represents a significant hurdle in facilitating data sharing. The minimum information for biological and biomedical investigations (MIBBI) project, part of the broader BioSharing effort, worked with different research communities to coordinate their 'minimum information' checklists,[3 4] but each community has some unique requirements. Also, data annotation presents an inherent tension: the easier we make it for investigators to annotate their datasets, the harder it will be to ensure discoverability. Conversely, the more discoverable we make the datasets, for example, through annotation using controlled terminologies, the more burden we put on the data generators.

Researchers need better tools and resources to identify, evaluate, and implement standards. BioSharing is a great resource to register and discover standards, and has adopted the initial set of criteria described above, requiring the communities to do a self-appraisal and tag their entries accordingly. The standards development community also has an active role to play if they wish to maximize the use and uptake of their work. Reviewers of publications and associated adherence to data standards should include biocurators. In the absence of widely agreed upon metrics to evaluate community standards, the decision about which is the right standard falls on the researcher. For reasons described above, this situation is problematic. Table 2 lists some potential resources/functionalities to address this problem. For any of these resources, it is important to note that technology is dynamic, and therefore so are any associated standards. Relevant resources must be similarly dynamic and up to date.

DISCUSSION

While one can conjure up motivating scenarios from a regulatory or archiving standpoint, the value proposition behind adherence to standards only really makes sense if data are to be shared beyond the team that originally created them. Thanks in part to policies put in place by some funders and publishers,[8] many high throughput datasets are made publicly available and, at some level, standards compliant. However, these policies have a number of restrictions that make them fall short. Some apply only to data generation through grants that exceed US$500 000.[2] Some require only a very low bar of compliance, and data are still difficult if not impossible to interpret. In many cases, the policies are simply not enforced,[7] although the government and the NIH have recently taken steps to rectify that fact.[3 19 20]

Table 2 Potential resources to assist in the selection and adoption of appropriate standards

Resource	Notes
Lay person's primer to standards	This would be a text document for the lay person to describe the standard, what problem it helps solve, and how it achieves that. Although FAQs address a number of these questions, one must first identify the standard and find the respective FAQ. This would be a centralized collection of documentation that requires no previous knowledge
'Consumer reviews'	This would be a rating system along the lines of Amazon product reviews. Ontology registries such as the NCBO and the OBO Foundry enable or perform reviews, but the reviews are few in number, not substantive, or infrequent. As discussed above, the utility of a standard depends on the purpose for which it is being used, so information beyond numeric scores is needed
Standard-selection wizard	Decision support methods could be used to ask a researcher about the intended goals and make recommendations accordingly. For example, 'what instrument type was used to generate the data?' and, 'will these data be deposited in a public data repository? If so, which one?' etc. Clearly this would require significant resources and ongoing maintenance
Standards-adoption 'helpdesk'	This would be a centralized resource of real humans with expertise across a number of standards. Once a standard has been selected, many have rich user communities and distribution lists for help with questions. However, for an individual investigator who wants to be standards-compliant and does not know where to begin, expert advice can save significant time in researching options
Quality assurance tools	Similar to syntax validators such as for RDF, tools to gauge or validate standards compliance are useful for data submitters as well as reviewers

NCBO, National Center for Biomedical Ontology (http://www.bioontology.org/); RDF, Resource Description Framework.

Ideally, it should be noted, researchers themselves would be shielded from the complexity of data standards. Developers, informaticists, and curators are perhaps better equipped to delve into data standards than would be a clinician or bench scientist, but even they are typically not experts in specialized standards. In an ideal world, data generators would have access to user-friendly tools that enable the seamless use of relevant standards and can be customized to fit the different data and domain needs.[9] The actual standards would be hidden from the data generators, and their use made automatic through intuitive, user-friendly tools.

Although we have described some tools for the discovery and evaluation of standards if one is so inclined, the real challenge is incentivizing researchers to go to the trouble. This will probably need a combination of proverbial carrots and sticks. On the penalty side, funders and publishers must continue to develop and publicize progressive data-sharing policies, and to enforce those policies through the delay of publication or future funding, if necessary. On the incentives side, a formal system for data citation must be developed, and those citations acknowledged and valued by funders, professional organizations, and university promotion and tenure committees. Recent activity in the realm of data publishing has been an important first step.[35 36] Only when obstacles are minimized and incentives are properly aligned will investigators be able to justify the effort required to do the right thing.

Correction notice This article has been corrected since it was published Online First. In the Discussion section 'US$500 million' has been changed to 'US$500 000.'

Acknowledgements The authors would like to thank contributors to the BioSharing catalog and members of the CTSA Omics data standards working group, particularly Simon Lin who provided valuable feedback on early drafts of this manuscript, and Bill Barry, David Beck, Colette Blach, Jim Cimino, Todd Ferris, Carol Haynes, R Curtis Hendrickson, Carol Hill, Ken Kawamoto, Tahsin Kurc, John Osborne, Jeff Pennington, and Sarah Wheelan, who performed preliminary investigation into existing omics standards.

Contributors JDT conceived the idea for the paper. JDT, SAS, and MH co-wrote the manuscript.

Funding JDT was funded by NIH UL1RR024128 and a gift from David H. Murdock. MH was funded by NIH R24OD011883 and CTSA 10-001:100928SB23. SAS was funded by Oxford e-Research Centre and the UK Biotechnology and Biological Sciences Research Council (BBSRC) BB/I000771/1 and BB/I025840/1.

Competing interests SAS is the principal investigator for the BioSharing and ISA Tools projects.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Request for Information (RFI): input into the Deliberations of the Advisory Committee to the NIH Director Working Group on Data and Informatics. 2012. http://grants.nih.gov/grants/guide/notice-files/NOT-OD-12-032.html (accessed 30 May 2013).
2 National Institutes of Health. NIH Data Sharing Policies. 2013. http://www.nlm.nih.gov/NIHbmic/nih_data_sharing_policies.html (accessed 1 Jul 2013).
3 Upcoming Changes to Public Access Policy Reporting Requirements and Related NIH Efforts to Enhance Compliance. 2012. http://grants.nih.gov/grants/guide/notice-files/NOT-OD-12-160.html (accessed 16 May 2013).
4 National Science Foundation. NSF Supports U.S. Participation in the Launch of a New International Effort Aimed at Making Data Easier to Share Among Researchers. 2013. http://www.nsf.gov/news/news_summ.jsp?cntn_id=127299 (accessed 30 May 2013).
5 Federal Register. Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research. 2011. https://www.federalregister.gov/articles/2011/11/04/2011-28621/request-for-information-public-access-to-digital-data-resulting-from-federally-funded-scientific (accessed 30 May 2013).
6 Nature Publishing Group. Raising standards. Nat biotech [Editorial]. 2013;31:366. doi:10.1038/nbt.2588.
7 Alsheikh-Ali A, Qureshi W, Al-Mallah M, *et al*. Public availability of published research data in high-impact journals. *PLoS One* 2011;6:e24357
8 Field D, Sansone S-A, Collis A, *et al*. Megascience. 'Omics data sharing'. *Science* 2009;326:234–6.
9 Sansone SA, Rocca-Serra P, Field D, *et al*. Toward interoperable bioscience data. *Nat Genet* 2012;44:121–6.
10 National Research Council. Toward Precision Medicine: building a knowledge network for biomedical research and a new taxonomy of disease. The National Academies Press. 2011.
11 Barnes MR, Harland L, Foord SM, *et al*. Lowering industry firewalls: pre-competitive informatics initiatives in drug discovery. *Nat Rev Drug Discov* 2009;8:701–8.
12 Waldrop M. Data's shameful neglect. *Nature* 2009;461:145.
13 Sansone S-A, Rocca-Serra P. On the evolving portfolio of community-standards and data sharing policies: turning challenges into new opportunities. *GigaScience* 2012;1:1–3.
14 Friend SH, Norman TC. Metcalfe's law and the biology information commons. *Nat Biotechnol* 2013;31:297–303.
15 DRYAD. Dryad Digital Repository. http://datadryad.org/ (accessed 31 May 2013).
16 One Mind. One Mind for Research. http://1mind4research.org/
17 Eastman P. ASCO's continuous learning prototype passes proof-of-principle test. *Oncol Times [serial on the Internet]* 2013;35: http://journals.lww.com/oncology-times/Citation/2013/05100/ASCO_s_Continuous_Learning_Prototype_Passes.4.aspx.
18 European Translational Information and Knowledge Management Services. http://www.etriks.org (accessed 30 May 2013).
19 National Institutes of Health. Big data to knowledge. 2013. https://commonfund.nih.gov/bd2k/index.aspx (accessed 1 Jul 2013).
20 Holdren J. Memorandum: increasing access to the results of federally funded scientific research. In: Office of Science and Technology Policy, Executive Office of the President. Washington, DC, 2013.
21 Hayden EC. Geneticists push for global data-sharing. *Nature* 2013;498:16–17.
22 Smith B, Ashburner M, Rosse C, *et al*. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 2007;25:1251–5.
23 Noy NF, Shah NH, Whetzel PL, *et al*. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res* 2009;37 (Web Server issue):W170–3.
24 Common Data Element Resource Portal. 2013. http://www.nlm.nih.gov/cde/ (accessed 23 May 2013).
25 Komatsoulis GA, Warzel DB, Hartel FW, *et al*. caCORE version 3: implementation of a model driven, service-oriented architecture for semantic interoperability. *J Biomed Inform* 2008;41106–23.
26 Field D, Sansone S, Delong EF, *et al*. Meeting report: BioSharing at ISMB 2010. *Stand Genomic Sci* 2010;3:254–8.
27 International Organization for Standardisation. Standards. http://www.iso.org/iso/home/standards.htm (accessed 1 Jul 2013).
28 Scheuermann RH, Kong M, Dahlke C, *et al*. Ontology-based knowledge representation of experiment metadata in biological data mining. In: Chen J, Lenardi S, eds. *Biological data mining*. Chapman & Hall, 2010.
29 Coronel C, Morris SA, Rob P. Database systems: design, implementation, and management: Cengage Learning. 2012.
30 Teorey TJ, Lightstone SS, Nadeau T, *et al*. Database modeling and design: logical design: Morgan Kaufmann. 2011.
31 Freimuth RR, Freund ET, Schick L, *et al*. Life sciences domain analysis model. *J AmMed Inform Assoc* 2012;19:1095–102.
32 Simon J, Dos Santos M, Fielding J, *et al*. Formal ontology for natural language processing and the integration of biomedical databases. *Int J Med Inform* 2006;75:224–31.
33 Lee JW, Devanarayan V, Barrett YC, *et al*. Fit-for-purpose method development and validation for successful biomarker measurement. *Pharm Res* 2006;23:312–28.
34 Taylor CF, Field D, Sansone SA, *et al*. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat Biotechnol* 2008;26:889–96.
35 Edmunds SC. Peering into peer-review at GigaScience. *GigaScience* 2013;2:1–3.
36 NPG. Scientific data. 2013. http://www.nature.com/scientificdata/ (accessed 8 Jul 2013).