

Data Standards: What are they and how are they developed?

Helen Jenkins

Computational Biology Research Group

Department of Computer Science

University of Wales, Penglais,
Aberystwyth

United Kingdom, SY23 3DB

haf@aber.ac.uk

This talk will:

- Discuss the need for data standards
- Describe the nature of data standards
- Describe how data standards are developed
- Identify the benefits of adopting data standards
- Introduce some of the current initiatives in data standardisation

What can data standards do for me?

- Biological experiments can generate large quantities of data.
- Typically experimentalists will want to be able to do one or many of the following:
 - Transmit data sets to collaborators or other external sites.
 - Store data with adequate curation.
 - Make data sets available for statistical analysis and data mining.
- Data standards support these activities by defining:
 - The structure of data sets.
 - The types of information that they should contain.

What are data standards?

- Data standards are:
 - Formal descriptions of the structure and content of data sets.
- They aim to:
 - Identify the data that is *necessary* and *sufficient* to fully describe an experiment.

Necessary data

- Data:
 - The results of an experiment.
- Metadata (data about the data):
 - **Experimental context.** Data items that define the experimental context are required to enable correct interpretation and use of a data set.
 - **Sources of experimental variability.** It is important to anticipate and capture data items that describe sources of experimental variability that may obscure the biological variability we are trying to measure to enable explanation of anomalies in the data.

Data and metadata examples

- Data

- Peak lists
- Spectra
- ...

- Metadata

- Genetic lines
- Greenhouse environment parameters
- Sample preparation and handling protocols
- Instrument settings
- Noise reduction algorithm parameters
- ...

Data standard development: Stage one – definition of content

- Analysis of a biological domain or activity to identify the data items that are necessary and sufficient to describe it fully.
- Method:
 - Requirements analysis performed by discussions with experimentalists, study of existing data and observation of practice.
- Output:
 - Produces a written definition of the data items required that, when verified, will provide the basis for common understanding of the data sets that conform to it.

Data standard definition: Stage two – data modelling

- Construction of formal data models that capture and structure the necessary data items.
- Method:
 - Data modelling is a software engineering task that involves structuring the data items and identifying the relationships between them. It involves iterations with experimentalists to check understanding and in this way often serves to verify the written description.
- Output:
 - Produces a formal data model that may be used as the basis for the design of automated data handling and storage tools.

A word on ontologies

- Key to the usefulness of data standards is common vocabulary
 - It is still possible for a third party to mis-interpret a data set that conforms to a data standard simply because of different use of terminology by the data producer.
- An ontology is a hierarchical formal specification of the concepts within a domain. It specifies the vocabulary that may be used to refer to those concepts and how the concepts may be related.
- The development and use of ontologies and controlled vocabularies is an integral part of nearly all of the data standardisation initiatives that I will introduce later.
- In addition, the FuGO project (<http://fugo.sourceforge.net>) is aiming to bring the work of these groups together to produce a single ontology that will provide a source of terms for consistent annotation of functional genomics experiments.

Benefits of data standards

- Data standards that describe experiment results in their experimental context, enable:
 - Proper interpretation of results.
 - Laboratory interoperability.
 - Meaningful comparison of datasets.
 - Replication of experiments.
 - More options for retrospective analysis of data.

Other benefits of data standards

- Data standards, and the formal data descriptions that underlie, them:
 - Encourage consideration and development of best practice and Standard Operating Procedures.
 - Standardize the reporting of experiments and archiving of data associated with publications.
 - Enable the development of databases and verifiable transmission mechanisms to facilitate the storage, collection and dissemination of logically correct datasets.

Data standards for functional genomics:

- Transcriptomics:

- MIAME (Nat. Genetics., **29**:365-371) defines the data that should be recorded for a microarray experiment.
- MAGE (Spellman et. al., Genome. Biol., **3**, 2002) is an associated formal data description.

- Proteomics:

- PEDRo (Nat. Biotechnol., **21**:247-254) is a formal data description for proteomics.
- PSI-OM (Proteomics, **4**:490-491) is the result of further development of PEDRo by HUPO PSI
- MIAPE is an associated definition of the data that should be recorded for proteomics experiments.

Data standards for metabolomics

- In metabolomics we have the following for defining metabolite complements in their experimental context:
 - SMRS (Nat. Biotechnol., **23**:833-838) is a draft policy for standardisation of reporting for metabolic analysis mainly for pre-clinical drug trials
 - MIAMET (Trends in Plant Science, **9**(9):418-425) is a definition of the data that should be recorded for a metabolomics experiment
 - ArMet (Nat. Biotechnol., **22**:1601-1606) is a formal data description for plant metabolomics.
- All these initiatives are now coming together under the umbrella of the Metabolomics Society's Data Standardisation Initiative (<http://www.metabolomicssociety.org/mstandards.html>)

Integrated data standards

- In the future it will become more and more common for transcriptomic, proteomic and metabolomic analyses to be combined in integrative experiments.
- Therefore, we will need a common data standard for experiment and sample descriptions that must contain the metadata required to fully evaluate the different 'omic' analyses performed on samples from a single experiment
- To this end the following have been produced:
 - SysBio-OM (Bioinformatics, **20**(12), 2004-2015).
 - FGE-OM (Bioinformatics, **20**(10), 1583-1590).
- Other initiatives:
 - The FuGE project (<http://fuge.sourceforge.net/index.php>)
 - SBML (<http://sbml.org/index.psp>)

Finally...

- Some of these initiatives are already widely accepted
 - SBML boasts support by over 90 software systems and MIAME-compliant experiment description is required by many prestigious journals.
- For others only time will tell.
 - Computer Science history contains some good examples of standards created by committee, that cost a lot of money and took years to develop, and that were never fully adopted as community-led *de-facto* standards took over in the meantime.
 - Fundamental to the success of data standards is, therefore, involvement with the experimentalists at all stages of their development, early roll-out of draft versions for feedback and input from as many people as possible.