Managing and sharing experimental data: standards, tools and pitfalls

Norman W. Paton¹
School of Computer Science, University of Manchester, Oxford Road, Manchester M13 9PL, U.K.

Abstract
Experimental processes in the life sciences are becoming increasingly complex. As a result, recording, archiving and sharing descriptions of these processes and of the results of experiments is becoming ever more challenging. However, validation of results, sharing of best practice and integrated analysis all require systematic description of experiments at carefully determined levels of detail. The present paper discusses issues associated with the management of experimental data in the life sciences, including: the different tasks that experimental data and metadata can support, the role of standards in informing data sharing and archiving, and the development of effective databases and tools, building on these standards.

Introduction
Experimental data in the life sciences are increasingly complex and voluminous. This stems both from technology-push (the development of novel, high-throughput methods) and application-pull (the desire to integrate data from multiple sources, e.g. to support the development of mathematical models of behaviour in systems biology). As a result, experimental laboratories increasingly face challenges managing and interpreting locally produced data.

Effective archiving of experimental data can potentially support a range of tasks, such as:

(i) The discovery of relevant results, so that, for example, data sets that use a particular technique or combination of techniques can be identified and studied by experimentalists during experiment design or data analysis.

(ii) The sharing of best practice, whereby, for example in proteomics, approaches that have been successful at identifying membrane proteins or low abundance proteins can be captured alongside the results produced.

(iii) The validation of results, whereby, for example in proteomics, the number of proteins identified or the specific proteins found to be in a sample (or not) can be assessed in the light of the experimental process undertaken.

(iv) The sharing of data sets, so that, for example, public repositories can import or export data or multisite projects can share results to support integrated analysis.

However, providing an infrastructure and capturing the relevant experimental data and associated metadata are not without cost, and the potential benefits are not always experienced by the people who have to pay the costs. However, increasingly, funding bodies are establishing policies that encourage data sharing, journals expect comprehensive experimental results to be made available in standard forms, and analyses require data of different types to be combined. As such, identifying cost-effective techniques and agreeing good practices for data handling and sharing will be a significant challenge for the near future. Important design decisions include identifying what to store, how to store it and why.

For many types of experiments, there is a data pyramid, as illustrated in Figure 1. Raw data, as produced directly from an instrument, are subject to one or several forms of post-processing, to yield derived data that identify new information or properties. Some portion of this derived data is then selected for consumption as the ‘result’ of the experiment in a subsequent analysis activity. For example, in proteomics, the peak lists produced by a mass spectrometer might be considered the raw data. These data are then scanned against a database of protein sequences to yield a collection of peptide identifications, which in turn give rise to presence/absence, absolute or relative measures of proteins in the original sample. Descriptions of peptide hits and the associated evidence on the proteins in a sample constitute derived data. However, a typical consumer of the results of the experiment will apply some threshold to the results and see the experiment as producing a small number of (protein, measure, confidence) data items. What data from the pyramid should be stored and how it should be stored?

(i) Raw data: the motivations for storing raw data are that it may be difficult or impossible to produce again and that it may be desirable to rerun derivations over the data at a later date. For example, in proteomics, there are multiple algorithms for performing identifications, running identification algorithms involves the setting of complicated control parameters that may need to be reviewed, new identification algorithms are produced on a regular basis, and changes to sequence databases lead to changes in identifications.

(ii) Derived data: the motivation for storing derived data is that it provides the explanation as to why specific results were produced. However, if the raw data are stored, it may be
sufficient to describe how the derivation took place, as, at least in principle, the derivation processes can be repeated. It may be challenging to rerun derivations because of changes in software versions or auxiliary data resources used by derivation algorithms. Furthermore, rerunning derivations may be slow, so it may be difficult to recompute derived data on demand.

(iii) Results: the motivation for storing the key results that can be drawn from the derived data is that they are the most frequently consumed part of the pyramid. In addition, they are often compact compared with the raw and derived data.

In terms of how data should be stored, this depends principally on the form of access that is required. Raw data are often first produced in proprietary file formats and read in their entirety by analysis programs; as a result, the data need to be available in a format that can be read by these programs. Thus such data are typically stored in the original format, or in some standard representation that can be used by multiple analysis programs. Derived data that support specific decisions can often be produced using many alternative techniques and are more likely to be queried than raw data. As a result, if access to derived data is to be straightforward for validating downstream results, it is problematic that such data are represented in inconsistent ways; therefore some form of consistent model and storage infrastructure is valuable. The top-level results of an experiment are often effectively the results of queries over the derived data and thus may be straightforward in structure.

In addition to the data described above, there is also likely to be a significant amount of associated experimental metadata, to describe the sample, how it was obtained, the processing steps to which it has been subject, the equipment used etc. In practice, although such data are often much smaller than the raw or derived data, they may be as complex in structure and more expensive to capture, as some level of manual data entry is often required from experimentalists. However, subsequent search and interpretation activities often lean heavily on experimental metadata, so their quality and consistency are important to practical archiving and reuse.

**Standards**

The trends and challenges discussed in the ‘Introduction’ section are encouraging systematic work towards the development of community standards and practices, which support the interpretation and sharing of experimental data. In life science informatics, there have been a fair number of de facto standard file formats, but functional genomics and systems biology are increasing the profile of de jure standards (see, e.g., [1–3]).

Standards can support different but related tasks. For example, Figure 2 illustrates the different roles that may be played by minimum information guidelines, good practice guidelines and formats. Minimum information guidelines (e.g. MIAME (Minimum Information About a Microarray Experiment) for microarray data [4] and MIAPE (Minimum Information About a Proteomics Experiment) for proteome data [5]) describe the level of detail that must be provided about an experiment to support certain tasks, such as the validation of the experimental results. Good practice guidelines (see, e.g., [6]) describe the way an experiment should be carried out to enable a specific form of conclusion to be drawn (e.g. the number of replicates required). There are more minimum information guidelines than good practice guidelines, as the latter are much more challenging to produce; as an experiment needs to provide some level of evidence in support of a conclusion, changes to the nature of the conclusion affect the appropriateness of a protocol or practice. Conformance to the minimum information guidelines should enable an informed reader to determine whether or not they are...
comfortable with the experiment design. Both minimum information and good practice guidelines tend to result in textual descriptions. Formats, by contrast, are designed to support computational searching and manipulation of the data, and typically include some form of structured file format, often represented using XML, and some form of terminology to be used when populating data elements in the XML file. For example, MAGE-ML (Microarray Gene Expression Markup Language) [2] and the MGED (Microarray Gene Expression Data) Ontology [7] are a format and an ontology for use with microarray data.

Several standards bodies are working on different forms of experimental data [e.g. the MGED Society, the PSI (Proteomics Standards Initiative) and the MSI (Metabolomics Standards Initiative)]. Both the MGED and the PSI have produced minimum information proposals, XML data formats and ontologies. In addition, all of these standards bodies are starting to make use of a common foundation model, the FuGE (Functional Genomics Experiment) model [3], thereby providing some consistency in practice between the communities and easing the development of consistent representations for multiple omics experiments. In the systems biology community, there are minimum information guidelines for specifying what should be said about a biochemical model [8], the widely used Systems Biology Markup Language [9] provides a format by which models can be shared, and a Systems Biology Ontology is under development. These de jure standards are often at an early stage in their ongoing development, but there is increasing support for their use from software tools, public repositories and journals, so it seems likely that they will become well established. If they do, this should significantly ease the development of databases and tools for managing such data in the laboratory.

Tools
In-house data handling for large-scale experimentation is not yet straightforward. While equipment vendors provide data capture and analysis capabilities for their products, at least in academic settings, laboratory information management, in particular in support of downstream analysis, tends to be a work in progress. In support of data standards, several data management platforms have been developed for microarray data that support the level of detail required by MIAME and that enable export of data in MAGE-ML (see, e.g., [10]), and several commercial offerings export data conforming to PSI formats [11], but the other functional genomics data standards are less well established, and comprehensive tool support necessarily lags behind the completion of the standards. Furthermore, many experimental disciplines lack accepted standards for some or all of their activities, and thus development activities are designing models in parallel with their software infrastructures, for example in structural biology (see, e.g., [12]) and microscopy (see, e.g., [13]). The need for rapid development of software infrastructure to support the evolving experimental landscape has led some researchers to advocate the use of model-driven software engineering (see, e.g., [12,14]), but there are still significant challenges keeping software infrastructures in step with experimental and analytical practices.

A requirement that further complicates the biological data management landscape is the need to integrate multiple sources of data to draw meaningful conclusions. Data integration infrastructures can often be classified as loosely or tightly coupled. In a loosely coupled infrastructure, the data to be integrated stay in their original store, and are extracted and transformed as required to answer requests in conjunction with other data sets. This integration on demand model has low up-front costs and often supports consistent access to local and remote data, for example using workflow technologies (see, e.g., [15]). However, inconsistencies in naming, terminology and representations in the data sets to be integrated surface at integration time, and thus the process of conducting integration can be cumbersome and error prone. In a tightly coupled infrastructure, the data to be integrated are copied and reorganized into a form that eases integrated analysis. This integration in anticipation model, which gives rise to data warehouses, has a high up-front cost, but once this has been paid should enable individual integrated analyses to be carried out in a straightforward manner. Which approach is most suitable depends on the setting. The loosely coupled approach suits unpredictable settings, in which speculative enquiries are to be made and approximate answers are sufficient. The tightly coupled approach suits a more stable context, where focused analyses are required over predictable data sets.

Pitfalls
Space prevents a comprehensive review of the many pitfalls of scientific data management. The following are common examples:

(i) Over ambition: eager informaticians design complex systems for capturing comprehensive descriptions of experiments that are time-consuming to populate and thus never used.

(ii) Under investment: grant applications include minimal or no resource to support the capture and management of complex data sets, leading to ad hoc solutions.

(iii) Reinvention: development of local solutions to problems that many people have solved before and for which solutions already exist.

(iv) Free software fixation: reluctance to spend £1000 on commercial software when there is the opportunity to spend ten times that much on staff development, or use ‘flaky’ but ‘free’ alternatives.

Conclusions
The consequences, for experimental laboratories, of the increasing ubiquity of high-throughput experimental techniques, combined with the associated requirement for
analysis of the complex data sets produced, have not yet been fully worked through. Many experimental results when published are accompanied by fairly minimal analyses of the novel data. There are several reasons for this: (i) complex analyses often require different skills from those required to design and carry out the original experiment; (ii) many projects significantly underestimate their data analysis requirements, and thus resources are skewed towards data production rather than interpretation; (iii) many laboratories are struggling to provide effective infrastructures for managing their own data, so relating a new data set to earlier local or remote data sets may be cumbersome; and (iv) few domains have well established infrastructures for capturing, managing and analysing experimental results, in particular where multiple technologies are used together in an investigation.

Some of this simply reflects the fact that there are few well established data integration infrastructures in the life sciences; laboratories with integrated data management infrastructures have often carried out significant software development or tailoring activities in house. These activities represent significant duplication of effort, so it must be hoped that the growing activity in data standards will provide a foundation whereby the community can agree as to what is important, so that tool suppliers can then compete in the provision of effective platforms.

Standards production is often surprisingly slow; developing standards requires (i) a measure of technical and political consensus, (ii) an organizational framework within which to focus efforts and (iii) individuals who are willing to contribute time and expertise. However, a measure of emerging consensus around a methodology that first develops minimum information guidelines and then moves on to formats, significant agreement on the use of XML as the foundation on which to construct the formats, and cross-community collaboration in both format and ontology development may yet lead to a coherent space within which tool developers can work. This in turn will make it easier for laboratories to manage, integrate and analyse data, which in turn should establish a balance between experimentation and data analysis that reflects changes in experimental practice.

References


Received 10 August 2007

doi:10.1042/BST0360033