nature genetics

Discussing standards

Whereas plans for data generation and public release can be agreed upon between data producers and their funders, community standards for the reporting, analysis and publication of high throughput data require wider discussion and broad consensus.

Keeping track of data over the lifespan of large projects has recently become easier thanks to the practice of producing data management plans (marker papers).

On page 919 of this issue, the coordinators of the Human Microbiome Project (HMP) explain the funders' motivation to make these marker papers a part of key resource projects in order to achieve their stated aims with respect to the tripartite responsibilities of data producers, data users and funders. Early examples from our pilot project describing details for prepublication data release in accordance with funder policies can be seen on the *Nature Precedings* site (http://precedings.nature.com/collections/human-microbiome-project).

Other large projects are taking a similar approach; for example, the International Cancer Genome Consortium has its own data access site. Among the features on the site that promote best practice in data citation are template letters to facilitate communication between data users and data producers and journals, respectively, (http://www.icgc.org/icgc/cgp-template-letters) and at least one excellent example of a comprehensive and pragmatic data plan (ftp://data.dcc.icgc.org/version_1/Breast_Carcinoma-WTSI-UK-1/README.txt). We are confident that these careful efforts to explain the projects and their resources will pay off for the researchers by augmenting their reputation and increasing the citation of both the data and the resulting publications.

In contrast, although there is a parallel swell of enthusiasm to define and popularize community standards for research practice in reporting, analysis and publication of data, field by field, it is clear that this is a more difficult task than the data description efforts highlighted above. Previous editorials have highlighted the phenomenon of the profusion of standards documents authored by small groups, as well as problems with compliance even with those standards that are widely accepted in the research community (for example, *Nat. Genet.* **41**, 135, 2009). Notable successes in standard setting have been achieved in several fields in high-throughput genomics, for example, sequence quality scores, microarray reporting standards and genome-wide genotyping and statistical evidence for association. As the fields mature, community standards emerge and are upheld by peer referees and journals. One group has succeeded in evolving a set of standards for the Systems Biology Markup Language by posting drafts for public comment and iterative revision, gaining authors and users in the process (http://sbml.org/Documents/Specifications).

In order to help such standard-setting efforts succeed, the journal will prioritize those fields where the user group is large, where stakes are high,and wherever transparency and efficiency are at a premium. In this vein, we are pleased to announce the criteria for judging the Archon Genomics X PRIZE (page 917), where a substantial prize is offered as an incentive to develop high standards for clinical-quality human genome sequencing. The coordinators of this initiative have already consulted widely in order to come up with a consensus validation protocol. Here, their correspondence announces that the draft validation protocol will soon to be open for comments from the scientific community. The journal welcomes suggestions regarding ways to achieve broad consultation and fair incorporation of varying views on technical improvements and critical considerations. We believe that input from all stakeholders and from qualified users will make the standards more useful and that consensus may even serve to pare the guidelines down to their essentials.

The principle that everything on the web has a unique address is widely accepted but is only partially implemented. Because research data and information about the way in which the data were generated (metadata) are currently stored in a variety of databases and formats, there is always a tradeoff between the efforts required to keep track of resources versus the effort it takes to carry out research on the data. We recognize that current citation practice and database infrastructure may not support universal standardization efforts, and we may need to wait for semantic web infrastructure for full implementation. However, we have already seen the benefits to the journal and its authors of precise data citation and consensus field-specific standards, and we think that many of the tools and experts are available if we could help to link them all together. ■