ED_ng.1099.0212.SO_PR

nature genetics

It's not about the data

Researchers, funders and journals are in broad agreement that data must be accessible to support the conclusions of scientific publications and for the research to have impact. What is lacking is agreement on timing, formatting and attribution.

In December 2011, the US National Science Board (NSB) presented its report Digital Research Data Sharing and Management, which makes recommendations for the US National Science Foundation (NSF) to implement with its associated scientific and engineering communities. The report acknowledges that there are a broad range of challenges inherent in sharing research data and a need to provide instructions, support and trained professionals to enable data management. The report warns that "one-size-fits-all solutions cannot adequately address most digital research data policy issues because each research community is best suited to address the nuances of its own data." We agree that some communities have more sophisticated approaches to data access than others and that both the style of data presentation and the deal-breaking issues preventing access may differ somewhat by field. However, presenting many different solutions will do nothing to promote interdisciplinary data access. So, our recommendation is that we learn from each field's best examples and then all concentrate on the three crucial issues of timing, formatting and attribution. Each party can then bring what it does best to bear on solving these problems, whether that is funding research, teaching, programming, generating data or publishing.

While keeping up pressure for access to data resources ("No second thoughts about data access"; *Nat. Genet.* **43**, 389, 2011; http://www.nature.com/ng/journal/v43/n5/full/ng.827.html), we have been advocating the use of citable data management plans in line with the proposals of major funding agencies. Like the US National Institutes of Health, the NSF wants a formal declaration of the data resources in each large resource project and their use conditions, whereby "using the Data Management Plan to determine the timeline for initiating the data sharing process recognizes the rights and responsibilities of investigators." The report also recommends that "data should be shared using persistent electronic identifiers, which enable automatic attribution of authors and award funding." As an example of excellent practice in integrative data management, we laud the International Cancer Genome Consortium (ICGC; http://dcc.icgc.org/), which laid out its data policy for its 34 constituent studies in a marker paper (*Nature* **464**, 993–998, 2010). We particularly like the way in which a data management plan written at the grant stage evolves from an explanation of the project and the resources it will generate. As the project progresses, the plan is versioned to detail the databases and data fields that will be generated, with a detailed timeline for data use. The plan finally matures into a 'data descriptor', which we define as a user guide to the resources, accession codes and use conditions accompanying a completed project or publication. One ICGC study currently has a data descriptor in the database of Genotypes and Phenotypes (dbGAP), with accession code phs000370.v1.p1 linking the associated publication (*Science* **333**, 1157–1160, 2011; doi:10.1126/science.1208130) and 883 sequence data depositions in the Short Read Archive (SRA) database. We note that all versions of data plans and descriptors can be citable by digital object identifier (DOI) and can reside online in databases, project websites or journals.

Reformatting data is a full-time job for many researchers, even before the minimum reporting guidelines, terminologies and formats of each field are taken into consideration. In this issue, we present a Commentary and a Perspective suggesting solutions to these problems that have been developed by a process of community consultation and open review to which the journal was a party. In the Commentary, Susanna-Assunta Sansone and colleagues identify one central problem, namely that "most repositories are designed for specific assay types, necessitating the fragmentation of complex datasets," and they offer a unified view of the metadata formatting that will be needed to ensure that biomedical research datasets become interoperable. This solution is the overarching ISA framework, where the acronym stands for 'Investigation' (the project context), 'Study' (a unit of research) and 'Assay' (analytical measurement) (p 121). This proposal shifts the sets of reporting standards agreed upon by each community into the infrastructure and formatting of the data files themselves. Sansone and colleagues also list a set of participant communities that can pioneer the approach and teach by example. In the Perspective, Jonathan Derry, Stephen Friend and colleagues lay out the infrastructure requirements for a data commons in which all of the data depositors, curators and users become participants who engage with each other and the data by sharing tools and datasets. Their common uniting purpose would be improving preclinical drug design via multidimensional molecular modeling of human disease (p 127).

Within a data commons, attribution for scholarly contributions can be tracked and acknowledged. So, too, in the market of peer citation, and in this issue, the web of coauthorship during the recent years of genome-wide association studies (GWAS) is discussed by Brendan Bulik-Sullivan and Patrick Sullivan (p 113). Recognition of coauthor groups as well as formally declared consortia is the first step to establishing responsibilities for stewardship over complex datasets spanning multiple institutions, journals, databases and funders. Recognizing this need, a complementary approach is being taken by Neil Caporaso and Siiri Bennett (http://hdl.handle.net/10101/npre.2011.6680.1), who sent a survey to the participants in at least 110 of the named consortia in the GWAS field. Consortium information can be sent to these authors or updated by participants via the survey on the WikiGenes site (http://www.wikigenes.org/GWAS/consortia.html), to be published in a future issue. We anticipate that the more complete and granular information about the people who generated knowledge in this field will contribute to sustainable access to the datasets in perpetuity. ■