nature biotechnology

Credit where credit is overdue

A universal tagging system that links data sets with the author(s) that generated them is essential to promote data sharing within the proteomics and other research communities.

Science progresses most rapidly when researchers provide access to their data. This is not only good scientific practice. It facilitates the confirmation of original results. It provides others with a starting point to explore new or related hypotheses. It speeds the identification of errors and discourages fraud. And it minimizes inefficient use of funding in duplicating experiments. And yet, full data disclosure in proteomics, and many other fields, remains a work in progress. If practicing scientists are to be truly incentivized to spend time and effort on sharing data, funders and publishers need to develop a universally recognized tagging system that would link investigators to their deposited data. In this way, publicly disclosed data sets would become part of a researcher's publication record, allowing such efforts to be recognized by employers and funders alike.

Next month marks the two-year anniversary of the publication of guidelines specifying the minimum reporting requirements for papers describing proteomics and molecular interaction experiments (*Nat. Biotechnol.* **25**, 887–893, 894–898, 2007). Both sets of standards encourage deposition of data in public repositories, a practice that at the time was not universally adopted in proteomics.

We have carried out an informal survey of all manuscripts published in the year following publication of the two guidelines by the 68 authors of those two papers. The analysis reveals that a majority of the guideline authors published at least one manuscript last year for which no accompanying data were archived. If the proponents of data-reporting guidelines—most of whom are better resourced than other researchers in the field—are not depositing all of their data in a public repository, it is unlikely that the wider community is doing so either.

One issue that inhibits openness is the perception that full data disclosure may result in the loss of an edge over competing research groups. Occasionally, data are withheld while intellectual property is secured. More often, though, a failure to share simply reflects the considerable time and effort associated with formatting, documenting, annotating and releasing data. In this regard, the availability of new tools, such as an application (p. 598) to facilitate deposition of data in PRIDE (a public archive for mass spectrometry and protein identification data) should prove helpful.

For proteomics, the rapidly evolving technology and the complexity of the data itself pose particular challenges. Concerns about the quality of proteomics data generated by mass spectrometry have long plagued the field, raising the issue of whether peers have sufficient faith in other groups' work to not only value the data lodged in public repositories but also make the effort to deposit their own. Here too, though, progress is being made. A study reported in this issue (p. 633) demonstrates the high reproducibility of a targeted proteomic approach for biomarker discovery from plasma among several laboratories. Such a result would have been difficult to achieve using the technology and approaches of a few years ago.

But data quality is only part of the problem in overcoming the community's reticence about disclosure. For many researchers, the software provided by the public repositories for searching and analyzing proteomics data is not as efficient and user friendly as it could be. An analysis published last month by the Human Proteomics Organization cited the misassignment of peptides to ambiguously annotated proteins by database search engines as one of the major hindrances to researchers in the field (*Nat. Methods* **6**, 423–430, 2009). What's more, despite the recent launch of yet another archive for mass spectrometry and protein identification data—the US National Center for Biotechnology Information's Peptidome repository (p. 600)—the various proteomics databases have yet to introduce a standardized data format that would allow the seamless exchange of data. Contrast this with the genome databanks, where the pooling of nucleotide sequence data in a common format has been integral to consistency, accessibility and, above all, utility of sequence data for reanalysis.

With all of these impediments, it's not surprising that proteomics researchers have been slow to embrace data disclosure. It is equally clear that disclosure edicts and recommendations from funding agencies and scientific journals have been insufficient to ensure widespread proteomics data release, despite evidence that the papers of researchers who share their data have an increased citation rate (*PLoS ONE* **2**, e308, 2007). Clearly, other incentives are needed.

One option would be to provide researchers who release data to public repositories with a means of accreditation. This would take the form of a universally standardized tag for data that could be searched and recognized by both funding agencies and employers. An ability to search the literature for all online papers that used a particular data set would enable appropriate attribution for those who share. In essence, the tag would be a digital object identifier (DOI), currently best known for its use in unambiguously identifying papers online.

Similar to citation information about publications, citation information about a researcher's data DOIs could be gathered by funders assessing future support and used by institutions in performance evaluation. Researchers who disclose data sets that subsequently prove particularly useful to the community would end up with highly cited data DOIs, and could thereby be rewarded accordingly.

Such a system would not solve all the problems slowing data disclosure in proteomics and elsewhere. But it would provide greater incentive than the present system of evaluation, which is skewed almost exclusively to publications in high-profile journals and citation metrics. Data DOIs would not only enhance a researcher's reputation but also establish priority of data generation. Most important of all, they would provide a way to acknowledge the time and effort individuals must invest in sharing data, which ultimately benefits the scientific community as a whole.