nature genetics

On the table

Data tables are a central element of most scientific papers. Simplified tables with separation of data storage from presentation format are ways to increase the impact and use of research data.

Datasets are growing much faster than we can write about them. The world of information is no longer necessarily subject to the world of words. On the one hand, this shift has been recognized with revolutionary publishing initiatives, for example, the launch of our sister company, Digital Science (http://www.digital-science.com), and with database journals such as (Giga)^n Science from the Beijing Genome Institute (http://www.genomics.cn/en/index.php/). These new tools and modes of publication are most welcome additions to our collective strategies for mining the multidimensional information landscapes of science.

And the explosions from these launching revolutions illuminate the activity of established journals such as this one. We have been caught—not frozen but rather scurrying about—solving problems by evolutionary means. But the evolution of the research paper is both propelled by the proliferation of digital data storage and processing and shaped by the selective forces of human reading and thinking. Compromises between storage and presentation have been developed in print journals to aid reading (which is of course by far the largest economic activity in the world of publishing). In the digital world, these formats are not machine-readable and some of these chimeric creations must be cleaved and re-spliced.

"Table 1" of a typical journal article is currently a hybrid compromise that does not scale well as data increase. Even small tables need to be processed or reformatted when used for comparison with other data or as substrates for further research, limiting the paper's research utility and impact. It is no secret that the majority of many bioinformaticians' time is spent on format conversion tasks, scripting and scraping data from one tabular format to another. This is not a good use for highly trained creative people who could otherwise be doing research. Journal publication should "first, do no harm" to dataset accessibility.

If we were to ask for restrictions on what can go into a table, how simple could the storage format be? We think "as simple as possible, but not more so." It seems fundamental to insist that each cell of a table at the intersection of a row and a column contain a single entity. For example, it would be wrong to list a mean value together with a bracketed range (altogether three numbers plus spaces, brackets and a hyphen). Numbers of a series should be expressed to a compatible number of significant figures or decimal places and in scientific notation. A delimited format is preferable to a table divided by lines, perhaps delimited by tab characters, as commas are often used as separators in US display formats for thousands, and colons can be used to denote hours, minutes and seconds.

Once data are stored in a versatile format, they can be formatted for display. The most familiar formats imposed by journals are to align numbers at the decimal point so that people can assimilate their similarities and differences (and even perform calculations while reading). Hierarchical or extensive row and column headings also belong in the realm of display format, not in the storage table. But, while we are on this topic, should a storage table contain row and column headings at all? Could the rows and columns not simply be numbered and the headings considered metadata destined for a securely attached header file? Both the headings and the contents of simple tables are readily written and checked by eye, but larger tables always require software to create, format and read. So, tables can still be easy to check and to correct, despite having contents and headings stored separately.

We welcome recommendations for standards for human/machine-readable tabular formats from groups who have already made recommendations for the minimal essential storage requirements. We are also interested to hear of ways in which data stored in the minimal tables can be flexibly formatted according to author and journal templates. Because tabulated data are universally used, there are many stakeholders, so we offer to publish community consensus standards for tabulated genetics and genomics data along the lines already suggested by the journal for developing consensus community standards in other fields. Given more simply formatted storage tables of data, existing journals can certainly do a much better job of presenting elegant, easily understood data tables in human-readable papers while also hosting machine-readable data of immediate utility. People should read, machines should work. ■