nature genetics

FAIR principles for data stewardship

The FAIR data principles are simple guidelines for ensuring that machines can find and use data, supporting data reuse by individuals. More—and better—research can be generated by designing data and algorithms to be findable, accessible, interoperable and reusable, together with the tools and workflows that led to these data.

We are not lacking standards. Indeed, over 600 content standards for biological data types are listed by the BioSharing registry alone (https://biosharing.org). However, one recent attempt to set standards for all kinds of data generated by scholarly activity gets right to the point: much of our scholarship gets in the way of data reuse because it obscures the machine readability of the data. The consequence of this limitation is that the scale of data reuse by human researchers is restricted (*Sci. Data* **3**, 160018, 2016).

The authors, from diverse backgrounds (including representatives from *Scientific Data* and this journal), conclude that, rather than set yet more standards, we should deposit data and design tools for their formatting, distribution and storage according to the four basic principles of finding, accessing, integrating and reusing all scholarly data. This emphasis is designed to put a stop to the arms race between diversifying data types and metadata annotations on the one side and bespoke mining tools designed to parse those data and metadata on the other. The question to ask in order to be a data steward, to handle data or to simplify a set of standards is the same: "is it FAIR"?

Most types of reusable data that are expensive to produce now have purpose-built databases. FAIR principles dictate the publication of rich metadata to describe these data and to enable discovery of what is contained therein, even in the case of sensitive data that identify persons. The data fields and metadata schema should be accessible, together with the details of any access restrictions, whether or not the underlying data can actually be accessed. In contrast, many of the products of low-throughput bench science do not fit into these standard databases. The repositories so far created for such data are becoming increasingly diverse in purpose and form. The key to taming these is to realize that they will need to be searched by general-purpose open technologies because they contain unpredictable data types unsuited to specialized parsers.

Equally important to good scholarship is the publication of non-data research objects. Explicit analytical workflows, for example, are essential to most forms of knowledge generation. Publication of these according to FAIR principles is essential to ensure transparency of the work as well as maximal use to the community. The key to working with data is to realize that the human touch, the urge to annotate tables with footnotes and cram multiple elements and data types into every cell of a table, gets in the way of computation, automation and scaling up. And this impedes the usefulness of your work for other people. All research objects should be findable, accessible, interoperable and reusable (FAIR) both for machines and for people. ■