nature genetics

Data models to GO-FAIR

This journal and *Scientific Data* are calling for submissions containing linked open data models that embody and extend the FAIR principles: that data should be findable, accessible, interoperable and reusable by both humans and machines. These principles are achievable with existing resources, languages and vocabularies to enable computers to combine and reanalyze data sets automatically and lead humans to new discoveries.

This May, individuals who are experts in a range of scholarly disciplines—but relatively unfamiliar with computational data modeling—learned how to make the Internet of Data and Services unearth the full implications of their research. First, in a workshop at the Lorentz Centre in Leiden, the Netherlands (https://www.lorentzcenter.nl/), data from library collections, museum accessions, African mobile healthcare records, tomato breeders and virologists were all subjected to data stewardship according to the FAIR principles (*Sci. Data* **15**, 160018; 2016; https://www.dtls.nl/fair-data/fair-principles-explained/). Then, in a similar but accelerated event at the fifteenth-anniversary conference of Bio-IT World in Boston (http://www.bio-itworldexpo.com/), data sets dealing with pediatric oncology, personal genomic SNP variants and curated human mutations (https://www.ncbi.nlm.nih.gov/clinvar/) were given the FAIRification treatment.

What we learned from these exercises is that it currently takes about two days working with a dedicated repository and a data engineer to add and link the minimum metadata a computer needs to make a data set work with others. We also learned that, despite the motivational potential of the FAIR acronym, there is no un-FAIR. If a computer can host a data set and expose the metadata to describe it in response to queries, it is FAIR enough. Similarly, 'more FAIR' is a concept of ever-improving machine interoperability to be judged by SPARQL queries and SHACL constraints. The proper way to implement community standards is not to exhort with ever more complex checklists but to write queries and constraints that will indicate which FAIR data sets share architecture, concepts and data that can be usefully and legally interoperated.

Each cell in a table (as found in a Supplementary Table of a journal article) typically contains a name or the value of a measurement (string or literal). The row and column headers are each an instance of a class that must be uniquely identified on the web (for example, as a type in an ontology or the uniform resource identifier (URI) of a data item in a repository). Once these elements have been identified, the key to FAIR data modeling is to make a network sketch in which the row and column headings become nodes representing instances of concept classes, with arrows representing predicates (properties or verbs; for example, "is a" or "is part of") that relate them to one another. At this stage, it is astonishing to have to add a number of missing, implicit concepts connecting two adjacent columns of data that only appeared to be connected in a table, but for which there had been no declared relationship. The sketch serves two purposes, first as a wiring diagram to construct a graph in a linked data repository in the subject–predicate–object triplets of RDF or one of the subsequent technologies for representing semantic data. Secondly, the sketch is a map to direct peer scrutiny as a means to improve the precision of the model, to design future models and to collect future data sets.

Just as modeling is enabled by a repository that provides RDF specifications and ontologies, and which imposes the hierarchical DCAT concepts of the collection, data set and distribution (https://www.w3.org/TR/vocab-dcat/), we suggest that hierarchical modular models will be the ones that connect with the largest number of projects. At the generic level, many data sets contain common, interrelated elements, such as "sample", "assay" and "dependent variable." Many tabular data sets resulting from statistical analyses can be represented using the generic Data Cube concepts (https://www.w3.org/TR/vocab-data-cube/). To capture the richness of specific knowledge, there are a number of domain-specific ontologies (https://bioportal.bioontology.org/ and https://www.ebi.ac.uk/ols/index), which in this example would explain how the sample was taken, what technology performed the assay, the units of the dependent variable and that the assay was to measure levels of the metabolite D-glucose.

It will get easier to replace Supplementary Tables with online models. Many of the data sets with the highest reuse value fall into a small number of experimental designs: variant calls from genome sequencing or genotyping, somatic single-nucleotide variant calls for paired tumor–normal samples in a cancer study, RNA sequencing, quantitative trait association or linkage mapping, genome-wide association with a SNP panel and population sequence variants displayed on a phylogenetic tree. Data sets for investigations into the 3D genome, which interrogate epigenomic modifications and structural conformations, are more diverse and complex, but are built of reproducible and increasingly standardized assay modules.

FAIR works best if it is global and open (https://www.dtls.nl/go-fair/), but neither is required for the Internet of Data and Services. If a data set is intended to be FAIR, sufficient metadata must be provided to automatically identify its structure, provenance, licensing and potential uses without having to use specialized parsing tools. Any access protocols should be declared where they do or do not exist (for example, to maintain the privacy and consent conditions inherent in medical metadata).

We have the infrastructure we need thanks to the vision of those who quietly built the remarkable linked data languages and conventions that in many cases anticipated this research need by decades. As for why the time has come for linked data models in this field, genomics and statistical genetics publications have provided a level of reproducibility that justifies keeping and reusing the data. We offer to publish models and their implications as more incentives to go FAIR. ■