World view

By Mark A. Musen

Demand standards to sort FAIR data from foul

Without appropriate metadata, data-sharing mandates are pointless.

Last month, the US government announced that research articles and most underlying data generated with federal funds should be made publicly available without cost, a policy to be implemented by the end of 2025. That's atop other important moves. The European Union's programme for science funding, Horizon Europe, already mandates that almost all data be FAIR (that is, findable, accessible, interoperable and reusable). The motivation behind such data-sharing policies is to make data more accessible so others can use them to both verify results and conduct further analyses.

But just getting those data sets online will not bring anticipated benefits: few data sets will really be FAIR, because most will be unfindable. What's needed are policies and infrastructure to organize metadata.

Imagine having to search for publications on some topic – say, methods for carbon reclamation – but you could use only the article titles (no keywords, abstracts or search terms). That's essentially the situation for finding data sets. If I wanted to identify all the deposited data related to carbon reclamation, the task would be futile. Current metadata often contain only administrative and organizational information, such as the name of the investigator and the date when the data were acquired.

What's more, for scientific data to be useful to other researchers, metadata must sensibly and consistently communicate essentials of the experiments – what was measured, and under what conditions. As an investigator who builds technology to assist with data annotation, it's frustrating that, in the majority of fields, the metadata standards needed to make data FAIR don't even exist.

Metadata about data sets typically lack experiment-specific descriptors. If present, they're sparse and idiosyncratic. An investigator searching the Gene Expression Omnibus (GEO), for example, might seek genomic data sets containing information on how a disease or condition manifests itself in young animals or humans. Performing such a search requires knowledge of how the age of individuals is represented – which in the GEO repository, could be age, AGE, age (after birth), age (years), Age (yr-old) or dozens of other possibilities. (Often, such information is missing from data sets altogether.) Because the metadata are so ad hoc, automated searches fail, and investigators waste enormous amounts of time manually sifting through records to locate relevant data sets, with no guarantee that most (or any) can be found.

Some optimists assume this problem can be solved by referencing data sets in published manuscripts, which at least include experimental details. But often, no published manuscript ever appears; if it does, its descriptions are rarely adequate to understand the data in the form in which they are deposited. Thus, metadata for data sets must stand on their own, and they need to follow community-accepted guidelines, enumerating the key attributes of experiments.

When metadata standards exist, technology can be helpful. The CEDAR Workbench, developed by my group at Stanford University in California, offers a general-purpose approach to creating standardized metadata. CEDAR relies on machine-readable libraries of metadata-reporting guidelines and controlled terminologies adopted by specific disciplines, and it automatically generates forms that prompt those depositing data online to fill in metadata fields and to annotate data sets with all the experimental descriptors endorsed by a given scientific community. (The tool has been used in such disparate fields as biomarker investigations and wind-energy experiments.) The result is metadata that can be searched reliably and that use specific, consistent terms that indicate what an experiment was actually about. (For example, there is only one way to indicate 'age'.) But the workbench is useless in disciplines that lack basic metadata standards – including most fields of science.

If we're serious about sharing data, we need to develop standards that can make data FAIR. Funding agencies must go beyond simple mandates for FAIR data. The Netherlands Organization for Health Research and Development (ZonMw) in The Hague, for example, hosts workshops to develop simple metadata standards that its grant recipients can use. Already, this process has produced guidelines for reporting results related to COVID-19 and antimicrobial resistance, and many more workshops are planned. As a condition of funding, ZonMw requires new grant recipients to use these standards. Other funders should adopt this more participatory approach, providing tailored assistance alongside issuing requirements. This would lead not only to better data sets for targeted research programmes, but also to the creation of metadata standards that communities will want to apply.

If we really want FAIR data, the new international mandates for data sharing will have a significant price tag. The leaders of the ZonMw workshops estimate that development of a single standard costs €40,000 (US$40,000) when considering the labour contributed by the workshop attendees.

Scientists and their funders need to recognize that FAIR data will require more than just a mandate – such data will require an enormous investment. The research community must commit to creating discipline-specific standards for metadata and to applying them throughout the scientific enterprise.

> **"If we're serious about sharing data, we need to develop the standards that can make data FAIR."**

Mark A. Musen is professor of medicine (biomedical informatics) and biomedical data science at Stanford University in California.
e-mail: musen@stanford.edu

MARK A. MUSEN

© 2022 Springer Nature Limited. All rights reserved.