A personal take on science and society

World view

By Barend Mons

Invest 5% of research funds in ensuring data are reusable

It is irresponsible to support research but not data stewardship, says Barend Mons.

Many of the world's hardest problems can be tackled only with data-intensive, computer-assisted research. And I'd speculate that the vast majority of research data are never published. Huge sums of taxpayer funds go to waste because such data cannot be reused. Policies for data reuse are falling into place, but fixing the situation will require more resources than the scientific community is willing to face.

In 2013, I was part of a group of Dutch experts from many disciplines that called on our national science funder to support data stewardship. Seven years later, policies that I helped to draft are starting to be put into practice. These require data created by machines and humans to meet the FAIR principles (that is, they are findable, accessible, interoperable and reusable). I now direct an international Global Open FAIR office tasked with helping communities to implement the guidelines, and I am convinced that doing so will require a large cadre of professionals, about one for every 20 researchers.

Even when data are shared, the metadata, expertise, technologies and infrastructure necessary for reuse are lacking. Most published data sets are scattered into 'supplemental files' that are often impossible for machines or even humans to find. These and other sloppy data practices keep researchers from building on each other's work. In cases of disease outbreaks, for instance, this might even cost lives.

I tell research institutions that, on average, 5% of overall research costs should go towards data stewardship. With €300 billion (US$325 billion) of public money spent on research in the European Union, we should expect to spend €15 billion on data stewardship. Scientists, especially more experienced ones, are often upset when I say this. They see it as 5% less funding for research.

Bunk. First, taking care of data is an ethical duty, and should be part of good research practice. Second, if data are treated properly, researchers will have significantly more time to do research. Consider the losses incurred under the current system. Students in PhD programmes spend up to 80% of their time on 'data munging', fixing formatting and minor mistakes to make data suitable for analysis – wasting time and talent. With 400 such students, that would amount to a monetary waste equivalent to the salaries of 200 full-time employees, at minimum. So, hiring 20 professional data stewards to cut time lost to data wrangling would boost effective research capacity. Many top universities are starting to see that the costs of not sharing data are significant and greater than the associated risks.

> "Funders hold the stick: they should disburse no further funding without a data-stewardship plan."

Barend Mons is a professor at Leiden University Medical Center in the Netherlands, and a board member of the Leiden Centre of Data Science. He also co-leads GO FAIR and the International Science Council's committee on data.
e-mail: barend.mons@go-fair.org

Data stewardship offers excellent returns on investment.

A 2018 European Commission report estimates that problems with the reuse of data cost the EU at least €10 billion each year in the academic sector alone, and €16 billion in lost innovation opportunities. I translate that as roughly €100 billion lost annually at the global level. That's not even counting related reproducibility problems.

The FAIR guiding principles are now cited three times per day, but citations do not equate to practice. My colleagues and I, along with European Open Science Cloud, an initiative aimed at promoting open-science practices, scoped requirements for the continent-wide data-sharing infrastructure. We estimated that Europe will have at least 10 million serious data producers among its 70 million science and technology professionals and 1.7 million researchers. So we will need to educate about 500,000 data stewards of various kinds to support researchers through experimental design and data capture, curation, storage, analytics, publication and reuse.

These tasks are too complex and time-consuming to leave to researchers (I wrote a whole book on why it would be a grave mistake to train every future scientist to be a fully knowledgeable data steward). Few active researchers see data stewardship as their core business, especially because current incentive systems discourage data sharing and entrench an archaic and almost exclusively narrative-based culture of scholarly communication.

Fortunately, research institutions around the world are setting up data-competence centres to help, including the one at Wageningen University in the Netherlands and at the University of Turin, Italy. But this movement is in its infancy.

Funders hold the stick: they should disburse no further funding without a properly reviewed and budgeted data-stewardship plan. The carrot is that FAIR data allow much more effective artificial intelligence (FAIR can also mean 'fully AI ready'), which will open up unprecedented research opportunities and increase reproducibility.

Some data-intense organizations, such as pharmaceutical companies and medical centres, are committing the necessary funds. (Note that FAIR is not the same as 'open'; confidentiality rules still apply.) In Leiden, we now have a programme for computational oncology, in which a machine-actionable information file will be created for every tumour we profile in our patients, without practitioners' manual intervention. About 10% of this programme's budget is dedicated to the design and stewardship of these files, which we call FAIR digital twins.

More research institutions and funders must make similar commitments. The key is to build capacity, enable groups to collaborate nationally and internationally and share good practices so that good data stewardship becomes the rule, not the exception.