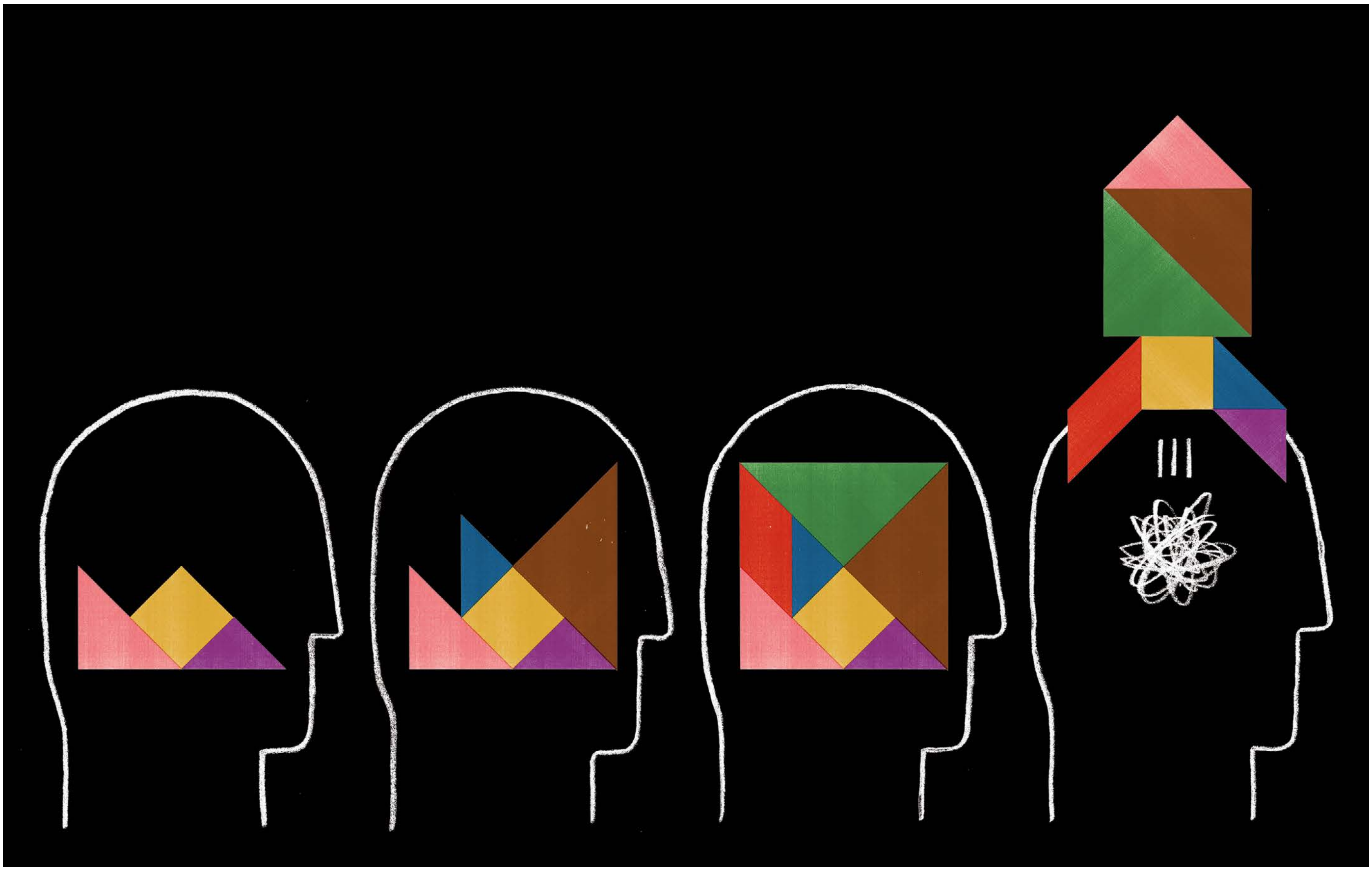

ILLUSTRATION BY THE PROJECT TWINS

SEVEN TIPS FOR SHARING SCIENTIFIC DATA

Specialists offer advice on how to make your data accessible, discoverable and useful. **By Jeffrey M. Perkel**

Miguel Acevedo typically gets two questions about his research on malaria in lizards. "Do lizards really get malaria?" (The answer is yes.) And, "Will I get malaria from a lizard?" (Not likely.)

Lizard malaria is a model for vector-borne disease ecology and evolution[1]. A colleague had been pursuing the same problem, at the same site in Puerto Rico, since the 1990s, and Acevedo, a wildlife ecologist at the University of Florida in Gainesville, wanted to combine those older data with his own to perform a long-term analysis. It was easier said than done. Whereas Acevedo's data were logged using a standardized data-entry template, the colleague's data were recorded in a mix of paper notebooks, Excel spreadsheets and hand-drawn maps. "It was some of the most organized data of that era, but we didn't have the standards then that we have today," he says. Columns weren't necessarily consistent from sheet to sheet, nor did they use the same units, and it wasn't always clear which sampling sites were being measured.

In the end, what could have been a morning's effort took "six or seven months", Acevedo says. "It's a lot of work, and it's not fun work, you know?"

Funder and publisher mandates, coupled with a growing emphasis on open science and reproducibility, mean that researchers are increasingly depositing data alongside their publications. Other scientists can use those data to drive new research. But not every journal requires that authors make their data sets available, and some authors decline to do so, either for fear of getting scooped or for lack of time. (The research data policy for Springer Nature, which publishes *Nature*, "strongly encourage[s] that all datasets supporting the analysis and conclusions of the paper are made publicly available at the time of publication", and mandates "the sharing of community-endorsed data types".)

Nature asked data scientists about their best practices for publishing usable, high-quality data – here's what they said.

Craft metadata

If there's one thing scientists can add to maximize their data's value, it's "metadata, metadata, metadata", says environmental scientist Patricia Soranno at Michigan State University in East Lansing.

Metadata are data that describe data – the timestamp and geolocation details that a smartphone camera stores with every image, for instance. Metadata basically explain what data mean, and are key to making data FAIR – findable, accessible, interoperable and reusable[2]. "Data without metadata", says Acevedo, "is like a Lego set without the instructions."

What those instructions should say varies from experiment to experiment – microscopy data require different metadata than do gene sequences, say. But according to Sarah Supp, an ecologist at Denison University in Granville, Ohio, they can generally be put into a simple 'README' text file that lists when, where and how the data were collected, and by whom; the licence under which they are released; whether data collection is complete; and their status – raw or processed, for instance.

It's worth including a 'codebook' that defines experimental variables, units, abbreviations, expected ranges and how missing data are denoted (using 'NA', for example). If there are many tables or files, then explain how they interrelate. And if software was used for data processing, detail the tools, version numbers and runtime parameters, says Anne Brown, a product-development scientist at Bayer US Crop Science in Chesterfield, Missouri. Template README files, data dictionaries and project summaries have been shared on Twitter by Crystal Lewis, a research-data management consultant in St Louis, Missouri (see go.nature.com/43kvzt2).

For Acevedo, good metadata practice has made his lizard malaria project maintainable. "It's like learning from trauma," he says.

Over-share

What with raw numbers, exploratory dead ends and the final processed data set, "at the end of the project, there's actually thousands of versions of the data", says Ciera Martinez, a research data scientist at the Eric and Wendy Schmidt Center for Data Science and Environment in Berkeley, California. So which one should scientists publish?

"If you're able to share both the raw data and the derived data, do so," says Karthik Ram, a data scientist at the Berkeley Institute for Data Science. Processed data underlie the analysis, but raw data let other researchers test your assumptions and processing strategies.

That said, raw data sets can be unwieldy and expensive to store. In that case, says Martinez, a "good rule of thumb" is to publish the data that were used to generate your figures.

Ultimately, says Brown, publishing data shouldn't simply tick a box, but should serve the scientific community. So, ask yourself what others are likely to want from the data, and how they might use them. "Knowing that can help you understand, OK, if other researchers are going to use this data then I am going to make sure that they're able to understand it."

Embrace standards

Every project is different, as are the expectations of which data should be published and how that should be done. So, look to the broader community for guidance, Martinez says. Many disciplines have dedicated data repositories, such as Genbank and the Protein DataBank for DNA sequences and protein structures, respectively. But data can also be posted to general archives such as Zenodo, Figshare and Dryad. Ask whether your publisher (or funder) has a preferred storage location and file format, Brown advises. Or, consult your institutional resource librarian, suggests Jacqueline Campbell, a plant geneticist at the US Department of Agriculture (USDA) Agricultural Research Station in Ames, Iowa.

Small data sets can be deposited on the code-sharing site GitHub, but that doesn't guarantee persistence, warns Ethan White, an environmental data scientist at the University of Florida. Data can be deleted or modified at any time, so archive the data formally, as well.

Never post data to personal websites, says Tracy Chen, a scientific analyst at the NASA Exoplanet Science Institute in Pasadena, California, who co-authored a best-practices document for astrophysics data[3]. If you change jobs or retire, links to personal websites can become obsolete.

Consider the format

Data should be in an open, non-proprietary file format, says Ellen Bledsoe, who teaches ecological data science at the University of Arizona in Tucson; otherwise, they could become unreadable. Bledsoe encountered that problem when she had to extract data from Lotus 1-2-3 – a now-obsolete commercial spreadsheet program. "Trying to finagle that data added a whole other step," she says.

Text-based file formats, such as CSV (comma-separated values), can be read by many tools and programming languages, achieving the 'I' in FAIR data. And unlike with binary files, it's easy to track how text files change over time. Above all, avoid using PDF files for tables, says Campbell, who is an assistant curator for the USDA's soybean genetics database SoyBase. Spreadsheets are easy to import, she says. But PDF tables must be manually keyed in – a slow, painful and error-prone process.

Include code

If you used code for data analysis, post it alongside the data. Code reveals the many steps and decisions you made, "providing, in effect, a more detailed version of the methods section", says White. Before publishing, test that the code runs in a clean computational environment – that is, one with no objects in memory. Remove computer-specific elements, such as hard-coded file paths. Add comments to show what you're doing, and detail how to run the code, suggests John Guerra Gómez, a computer scientist at Northeastern University's campus in San Francisco, California. "Think as a time traveller," he says. "What would I want John in the future to know about this?"

Finally, suggests Kari Jordan, executive director of The Carpentries, find a coding partner. The Carpentries, based in Oakland, California, runs workshops on scientific computing and data analysis, and one point that it makes during instructor training is to "never teach alone", Jordan says. "Never teach alone, don't learn alone, don't do anything alone."

For instance, says White, you could ask a more advanced programmer to provide high-level feedback: "What are a couple of big things that you can do to make this easier to understand?" White's typical response to this question is to suggest breaking up long blocks of code into discrete functions, eliminating repetitious code and ensuring that function and variable names are informative. If a third party can understand and execute your code, Supp says, "you've probably done a pretty decent job at making your code readable".

Think accessibility

Big-data projects often expect a certain level of technical infrastructure on the part of prospective users. And they make assumptions about how people will consume, query and manipulate the data.

These assumptions often don't hold, says Sabina Leonelli, who teaches the philosophy and history of science at the University of Exeter, UK. "The idea that you're creating platforms that are for universal use, that can be infinitely repurposed, fails in practice because it doesn't take account of the fact that there may be groups around the world which are working under different conditions."

Leonelli's advice: consult organizations, such as the Research Data Alliance or the Committee on Data of the International Science Council, for feedback on your data standards and assumptions. And, where possible, consider "low-tech solutions", she says. Can you develop a low-bandwidth version of a database, for instance, or release both low- and high-resolution images?

Fail to consider a range of requirements, says Leonelli, and the result will be a resource that only you and others like you can use. "You run the risk of producing a resource that doesn't take any of those needs into account."

Take the plunge

Open science, says Bledsoe, "is not an all-or-nothing game"; anything you can do adds value. "Even if you don't know how to go all the way to zero-to-60 open science, zero-to-20 is also really good," she says.

So, release your data – that gives data consumers more to analyse, and data providers more opportunities for collaboration.

It's also scary, Supp admits: sharing means opening oneself to scrutiny. "There's a certain level of vulnerability with that," she says. "But that's also how we get better."

Jeffrey M. Perkel is technology editor at *Nature*.

1. Otero, L., Schall, J. J., Cruz, V., Aaltonen, K. & Acevedo, M. A. *Parasitology* **146**, 453–461 (2019).
2. Wilkinson, M. D. *et al. Sci. Data* **3**, 160018 (2016).
3. Chen, T. X. *et al. Astrophys. J. Supp. Ser.* **260**, 5 (2022).