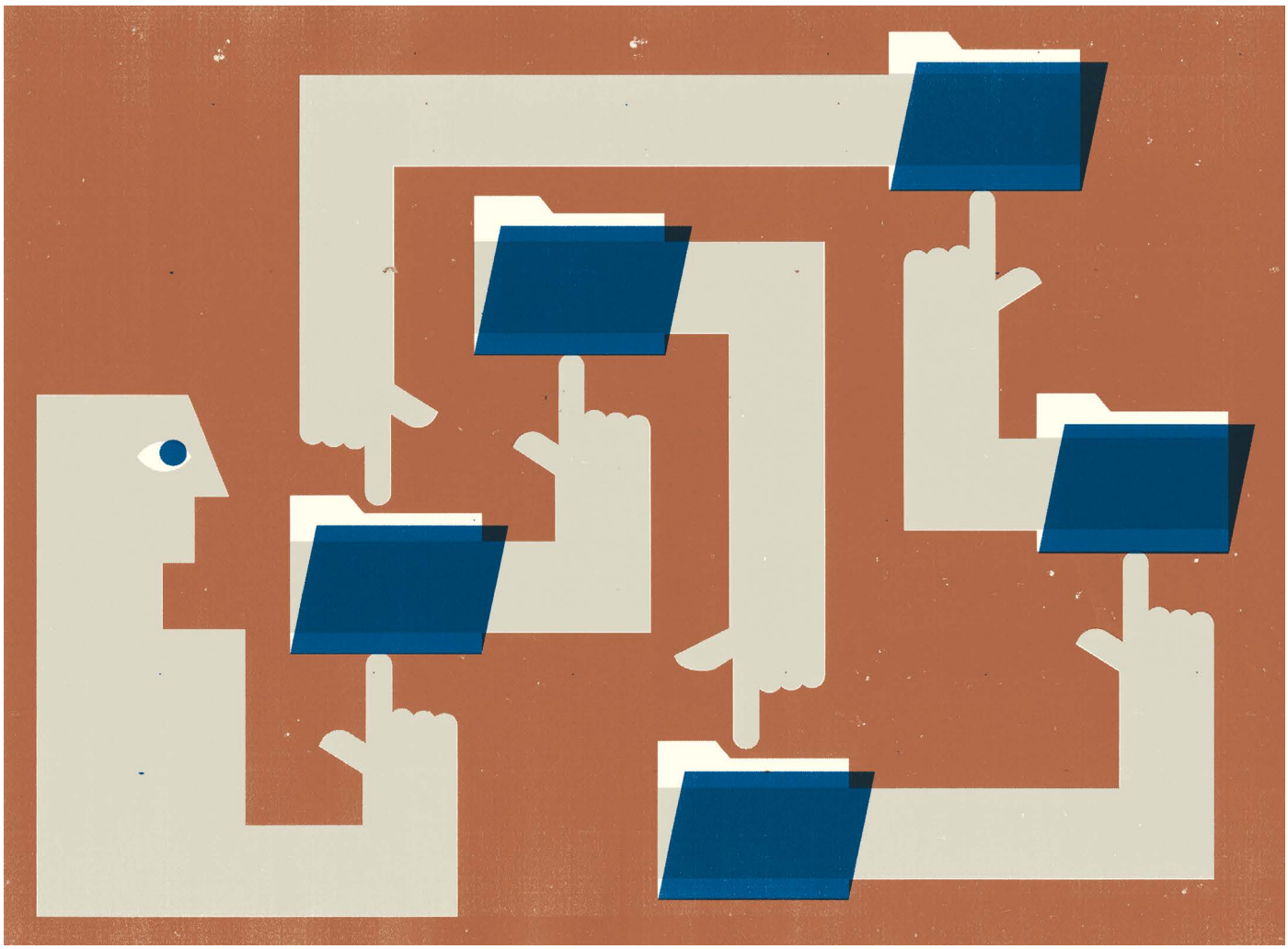

ILLUSTRATION: THE PROJECT TWINS

NEED TO UPDATE YOUR DATA? FOLLOW THESE FIVE TIPS

Researchers who run long-term projects often need to update their data sets. Here's how to do it while maintaining reproducibility. **By Sarah Wild**

Each week since 1977, researchers at the Portal Project have monitored how rodents, ants and plants interact with each other and respond to their climate on plots of land in Arizona. At first, the team shared those data informally. Then, beginning in the 2000s, the researchers would publish a data paper, wait several years and then publish a new one with combined old and new data to keep the information current.

"Data collection is not a one-time effort," says Ethan White, an environmental data scientist at the University of Florida in Gainesville, who began collaborating with the project in 2002. New tools have allowed the team to automate and modernize its strategy. In 2019, White and his colleagues developed a data workflow based on the code-sharing site GitHub, the data repository Zenodo and the software automation tool Travis CI, to keep their data current while preserving earlier versions (G. M. Yenni *et al*. *PLoS Biol*. **17**, e3000125; 2019); so far, the Zenodo repository holds around 620 versions. "We wanted an approach that would let us update things more consistently, but in a way that if someone ever wanted to replicate a past analysis, they could go back and find the precise original data that we used."

Long-term ecological research is not the only area that needs to maintain and update data for future use. Many researchers add to, revise or overhaul their data sets over the course of their projects or careers, all while continuing to publish articles.

But despite the need to update and preserve versions of data, there is little guidance for how to do so, says Crystal Lewis, a freelance data-management consultant in St. Louis, Missouri. "There are no standards for repositories; the journals are not telling you how to correct a data set or how to cite new data, so people are just winging it."

Good data-science practice can make the process more methodical. Here are five tips to help alter and cite data sets.

Choose a repository

Although it's easy to place data on personal websites or in the cloud, using a repository is the simplest way for researchers to store,

share and maintain multiple versions of their data, says Kristin Briney, a librarian at the California Institute of Technology in Pasadena, who helps researchers to manage their data. "It'll get it out of the supplemental information; it'll stop being shared upon request; it'll stop being shared on personal websites," on which it can be lost.

By the end of this year, US federal funding agencies will require researchers to put data in a repository, with some agencies, including the National Institutes of Health, already implementing the policy. Some journals also require authors to use data repositories. *PLoS ONE*, for example, recommends several general and subject-specific repositories for its authors, including the Dryad Digital Repository and Open Science Framework.

A repository, or data archive, is more than just cloud storage. Repositories provide long-term storage with multiple backups. Zenodo, for example, says that data will be maintained as long as Europe's particle-physics laboratory CERN, which runs the site, continues to exist. Generally, repositories also promise that archived data will remain unaltered and assign a persistent identifier to data sets so that others can find them.

Briney suggests that researchers check whether their funding agency has specific recommendations. There might also be a particular repository for the type of data, such as GenBank for genetic sequences; or a discipline-specific repository for the field of study. Some universities offer institutional options, which usually have the added benefit of technical support. When there is no specific repository available, the non-profit organization the Gates Foundation in Seattle, Washington, recommends generalist repositories, such as Zenodo, Dataverse, Figshare and Dryad.

Create multiple versions

For transparency and accessibility, making a new version when data are added is essential. The alternative — overwriting the old data with the new — makes it impossible to repeat previous analyses or to see how the data have changed over time. Although best practice around creating versions and data alterations tends to focus on future users and scientific reproducibility, the real beneficiary is the researcher, says Lewis. "Three months from now, you will forget what you did — you will forget which version you're working on, what changes you made to a data set. You are your biggest collaborator."

This is when data repositories come into their own, because many create new versions by default when data are added. Some repositories, such as Zenodo, also mint a digital object identifier (DOI) for each version automatically. "Since the very beginning, Zenodo has provided versionable data with individual DOIs that will take you to a specific version of the data, and also an overarching DOI that will link together all of those versions," says White. That creates an umbrella link, as well as a mechanism to cite specific versions of the data.

Managing versions without a repository is also possible. Researchers who store their data on GitHub, for instance, can use automation to create new 'releases' whenever they update their data. They can also create a version of the data set manually, using distinct file names, to differentiate these files from the earlier set, Briney says.

Define file names and terminology

Briney regularly helps researchers to wrangle their data. Her favourite tips for data management are to establish a file naming convention, which includes the date (often given as YYYYMMDD or YYYY-MM-DD), and to store files in their correct folders (see go.nature.com/44fggza). This is true whether you're storing data locally or in remote repositories. "It takes 10 minutes to come up with a file-naming convention, everything gets organized, and that way you can tell related files apart," she says. "It's like putting your clothes away at the end of the day."

"There are no standards for repositories, so people are just winging it."

Briney also recommends documenting metadata, explaining the different variables used, and the location of data in the various files and folders. These practices "help you, but are also good for data sharing, because somebody else can pick up your spreadsheet" and understand it.

Sabina Leonelli, who studies big-data methods at the Technical University of Munich in Germany, says that researchers should also explicitly document the terminology and queries used to generate and analyse their data. She gives an example of research using a biomedical database: "When you access certain databases, you frame your query" based on current definitions, she says. As knowledge develops, definitions shift and change, and if the specific definitions you used aren't captured, she says, you might forget the query that originally shaped your data.

Write a change log

As data evolve, it can be difficult to remember how they changed. Change logs can detail those differences.

Sometimes, the alteration is simply the addition of data. In other cases, the changes are more complex. For example, researchers might add variables to the data set or document an error in a previous version, says Lewis.

Each log entry can include the date and version number, starting with the first version. Data scientists might also want to note new techniques or even different software used in the data collection and analysis processes. "Any changes in the data files, the structure of the metadata, or the programming language you're using will have immediate repercussions on the readability and the usability of the data," says Leonelli.

Some researchers, such as White, also maintain an up-to-date data paper, in which they detail how their data have changed over time. White and his colleagues first uploaded their data paper to the bioRxiv preprint server in 2018 and have updated it twice since. "It's a little awkward in the academic credit space, but we have not found any journals that are publishing articles that can be continuously versioned through time," he says.

Update your technology

Even if all these steps are followed, future researchers might still struggle to reproduce results. That's because data that are older than about five years become increasingly difficult to access, owing to both the degradation of physical media and the inevitable obsolescence of their data formats. Compact-disc drives, are now a rare sight, for instance.

"Data standards have changed, the way data is stored has changed, and it actually becomes difficult to retrieve older data," says Leonelli. "Ask any biologist who has worked for more than 20 years, and they would immediately recognize this problem."

One good idea, then, is to make your data available in multiple formats, and if possible, in non-proprietary formats (such as a comma-separated values file instead of a Microsoft Excel file).

Some data repositories will do this for you, transforming data into common formats and, potentially, migrating them to different formats in the future as standards change. The Library of Congress, the de facto US national library, regularly updates its list of recommended formats to ensure long-term access to data. For those who do not use repositories, the solution is to regularly export old files into more current formats.

Failure to take this step, Leonelli says, could mean your data end up extinct. "We run the risk of living in a time when the only data available to us, and that we think exists, is the data created in the last five years."

Sarah Wild is a freelance journalist in Canterbury, UK.