ILLUSTRATION BY THE PROJECT TWINS

IN PURSUIT OF DATA IMMORTALITY

Data sharing can save important scientific work from extinction, but only if researchers take care to ensure that resources are easy to find and reuse. **By Michael Eisenstein**

Between 1969 and 1972, the United States landed six crewed spacecraft on the Moon as part of the Apollo programme. The missions retrieved priceless samples. But for more than four decades, the data from those samples remained stashed away at a handful of US laboratories — until Kerstin Lehnert came along.

A geoinformatician specializing in data rescue and preservation, Lehnert set out in 2014 to transform these data sets into a usable resource. Her team at Columbia University's Lamont–Doherty Earth Observatory in Palisades, New York, pored through old conference abstracts, scanned reams of publications and debriefed the senior researchers who first studied those lunar samples to collect, organize and annotate as much information as possible. One scientist, Lehnert says, "came with a half-metre-high pile of old, folded printouts and we spent a whole summer typing those data into Excel spreadsheets". Thanks to their efforts, these one-of-a-kind data are now freely available in the Astromaterials Data System.

Countless other laboratories, and their precious, irreplaceable data, are not so fortunate.

Lost to the ages

'Big science' efforts led by international consortia typically have data-management and sharing plans built in. But many labs doing small- to medium-scale studies in more specialized areas — such as analysing the biological contents of a single lake, or tracking the physiology of specific animal models — have no such systems. Their data often remain siloed in the labs that generated them, fading from memory as project members leave.

For the scientific community, that's a tragedy of wasted effort, lost collaborative opportunities and irreproducibility. "Things don't have to be really popular in order to be still very valuable," says Erik Schultes, international science coordinator for the GO FAIR International Support and Coordination Office in Leiden, the Netherlands. Established in 2018 to develop best practices for data preservation and sharing, GO FAIR is one of several efforts engaging with researchers in almost every scientific discipline to secure today's data for posterity. But success will require a concerted effort — and a shift in lab culture.

Digital data might be more convenient and shareable than the paper notebooks and printed photographs of yore, but they won't last forever. Physical storage media degrade; file formats and the software that produced them become obsolete. Most importantly, scientists can lose track of data when they stop being immediately useful. Even if retrieved, archival files often lack the context needed to interpret them.

"I've gone back and tried to make sense of data that I collected 10 or 15 years ago," says Dominique Roche, an ecologist at Carleton University in Ottawa who also studies data reuse and reproducibility. "I'm particularly knowledgeable about proper data management, and it was almost impossible." The difficulty only grows when researchers seek older data from other groups. In 2013, Timothy Vines, a data scientist then at the University of British Columbia in Vancouver, Canada, and his colleagues tested the limits of this accessibility by requesting data from 516 studies published between 2 and 22 years earlier. They managed to retrieve fewer than one in 5 data sets, and found that the likelihood of data being available and usable dropped by 17% each year after publication[1].

In recent years, researchers have taken to uploading their data to open-access repositories. This is an important step towards preservation and access, but it doesn't ensure reusability. In a survey of 100 data sets on the repository Dryad, Roche and his colleagues found that more than half lacked data needed to reproduce the work, and more than one-third were either not machine-readable or essentially unusable in other ways[2].

This is assuming that one can even find a particular data set: shared data can be scattered among multiple repositories, and it can be challenging to search across them, says Schultes.

A FAIR solution

The good news is that more sophisticated solutions are emerging. In 2016, a multinational team coordinated by Barend Mons, a specialist in biosemantics at Leiden University Medical Center, and including Schultes, published a framework known as the FAIR Data Principles[3]. The acronymic title describes its objective: that scientific data should be findable, accessible, interoperable and reusable.

Many of the framework's goals can be met through careful data curation and metadata creation. Metadata consist of documentation that describes a data set in a format that is both human- and machine-readable. They might, for example, describe the cell types and imaging parameters used in a microscopy experiment. That's essential information for third-party analyses, but also for finding the data. Other tools that can aid findability include re3data, developed by data-preservation organization DataCite, based in Hanover, Germany, which can help users to quickly narrow down which repositories are most likely to contain data relevant to their research. Google also offers a Dataset Search service, which can search across thousands of repositories to uncover specific data sets.

Metadata generation can create considerable work, but there are resources to expedite it. The Center for Expanded Data Annotation and Retrieval (CEDAR) at Stanford University in California runs a platform that generates simplified forms to produce FAIR-compliant metadata. These can be uploaded to repositories alongside the data they describe. GO FAIR also regularly runs Metadata for Machines workshops, at which data specialists and domain-specific experts help researchers to generate well-crafted metadata.

"Things don't have to be really popular in order to be still very valuable."

Fleshing out the record

Other efforts aim to preserve historic data sets. For example, Canada's nationwide Living Data Project trains and supports junior scientists to work with labs that have precious archival data from ecology or environmental science but lack the skills or resources to preserve them adequately. Roche, one of the project's coordinators, says the goal is to "organize the data, manage them properly and create the metadata so that then the data can be made public and are going to be understandable and reusable". The group has taken on more than 40 projects since 2020, salvaging one-of-a-kind research material, including 20 years of records of flora from Canada's Yukon tundra, and observations of bird populations from Tanzania's Serengeti region dating back to 1929.

But however old the data, preservation isn't a one-time task: to remain usable, raw scientific data must be maintained in formats that are compatible with contemporary hardware, software and operating systems. "You have to continue migrating data forward," says Christine Borgman, an information scientist at the University of California, Los Angeles. "As each new technology comes along, you've got to keep on upgrading every time."

That's a burdensome process, acknowledges Klaus Rechert, a computer scientist at the University of Freiburg in Germany. "For every data format, you need a migration tool," he says, "and the number of data formats is exploding." As an alternative, Rechert's team focuses on emulation — using software to replicate the hardware and operating system required to run old programs. This means that researchers can interact with old data sets using the original software. It has the added benefit of preserving the software itself, which is an important component of the scientific record.

But emulation can be technically challenging. So Rechert and his colleagues at the University of Freiburg have developed the Emulation-as-a-Service Infrastructure (EaaSI) — a cloud-based system that researchers can use to boot up antiquated systems. For example, a user who needs to run software originally designed for an old Apple or PC — or even older systems such as those produced by Commodore — can replicate that computing environment on any modern machine running Linux. The emulator's complexity is hidden behind a user-friendly interface, with technical components managed by the EaaSI team. "We currently do everything to automate it," says Rechert. "We are able to analyse the data set and try to figure out what is the most appropriate software environment."

A culture of preservation

With better tools available, the trick now is to give researchers incentives to put in the extra effort — a task that entails overcoming long-entrenched views on how scientific effort is credited and rewarded. This is especially true in academia, where publications remain the coin of the realm. Even with the advent of services such as DataCite, which provide ways to cite data sets, funders and hiring committees tend to gloss over those contributions in a scientist's CV. "Institutions don't really care whether your data sets get cited," says Roche.

Some major funders — including the US National Institutes of Health and Wellcome in London — have formal requirements for data management and sharing, and a number of journals make repository use a precondition. This can be a big incentive: Lehnert notes that when several major geoscience journals adopted the FAIR principles in 2019, submissions to the EarthChem Library data repository tripled. But there is little close oversight, and few teeth for punishing non-compliance; and researchers are rarely given the resources to support preservation efforts. "It keeps getting pushed down to the principal investigator as their responsibility," says Borgman.

Remedying this will require structural changes in the infrastructure for scientific funding and support. But the rising generation of scientists — born into an era of open-access, open-source and automated science — might be more amenable to the effort than their predecessors. "Nobody wants to hear that they might die tomorrow, but maybe your computer dies tomorrow and you don't have a good back-up," says Lehnert. "The data has to go into the repository so that 20 years from now, we're not suddenly saying, 'We need to invest again in rescuing these data.'"

Michael Eisenstein is a freelance writer based in Philadelphia, Pennsylvania.

1. Vines, T. H. *et al. Curr. Biol.* **24**, 94–97 (2014).
2. Roche, D. G., Kruuk, L. E. B, Lanfear, R. & Binning, S. A. *PLoS Biol.* **13**, e1002295 (2015).
3. Wilkinson, M. *et al. Sci. Data* **3**, 160018 (2016).