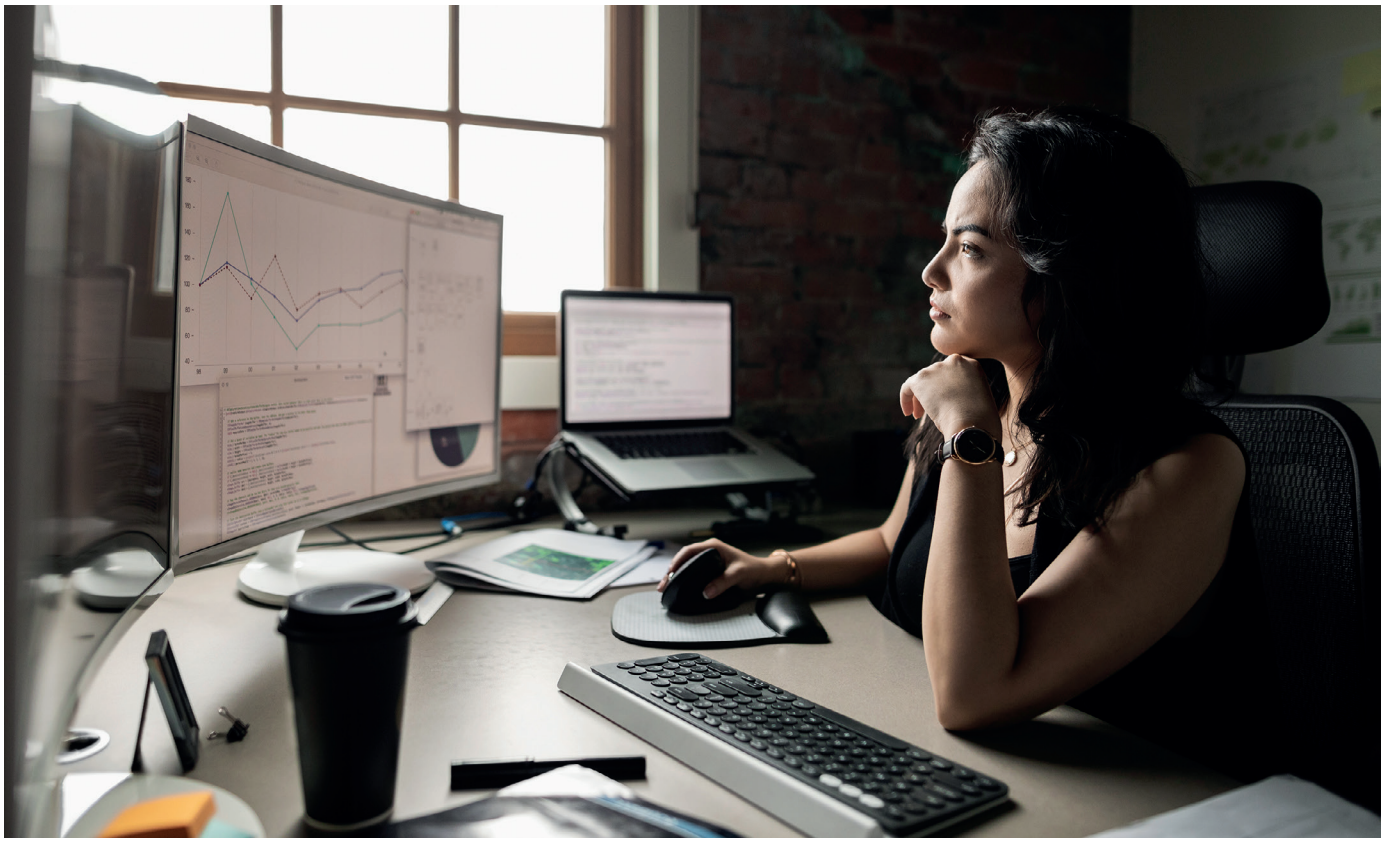
MYSTERY WRITER Postdocs routinely ghostwrite peer reviews **go.nature.com/2pwsjov**

NO EXCUSE Database of female scientists could help end 'manels' **go.nature.com/2px8uyn**

ONLINE Career resources from our community at **nature.com/careers**

HERO IMAGES/GETTY



Data sharing can be complex for scientists to navigate, but the rewards are often career-enhancing.

OPEN SCIENCE

Setting your data free

As science becomes more open, researchers who share data are reaping the benefits.

BY GABRIEL POPKIN

Ecologist Thomas Crowther knew that scientists had already collected a vast amount of field data on forests worldwide. But almost all of those data were sequestered in researchers' notebooks or personal computers, making them unavailable to the wider scientific community. In 2012, Crowther, then a postdoctoral researcher at Yale University in New Haven, Connecticut, began to e-mail and cold-call researchers to request their data. He started to assemble an inventory, now hosted by the Global Forest Biodiversity Initiative, an international research collaboration, that contains data on more than 1 million locations. Data are stored in CSV files (plain-text files that contain a list of data) on servers at Crowther's present laboratory at the Swiss Federal Institute of Technology in Zurich and on those of a collaborator at Purdue University in West Lafayette, Indiana; he hopes to outsource database storage to a third-party organization with expertise in archiving and access.

After years of courting and cajoling, Crowther has persuaded about half of the data owners to make their data public. The other half, he laments, say that they support open data in principle, but have specific reasons for keeping their data sets private. Mainly, he explains, they want to use their data to conduct and publish their own studies.

Crowther's database challenges reflect the current state of science: partly open, partly closed, and with unclear and inconsistent policies and expectations on data sharing that are still in flux. High-level bodies such as the US National Academies of Sciences, Engineering, and Medicine and the European Commission have called for science to become more open and endorsed a set of data-management standards known as the FAIR (findable, accessible, interoperable and reusable) principles. Government funding agencies in the United States, Europe and Australia require researchers to devise plans for data management and, in some cases, data sharing; some private funders also require them. Many journals, including *Nature*, have adopted policies that encourage or require authors to ▶

make data available. A plethora of open-access repositories host data sets from almost all fields, and scientists have been publicly criticized by colleagues for not sharing data.

Science is moving towards a greater openness, in terms of not just data but also publications, computer code and workflows. Yet researchers who are learning to navigate the open-science arena face a thicket of thorny issues. Many scientists — especially early-career researchers who are building a publication record — worry that sharing their data too early could lead to their getting scooped by a competitor. They must also decide whether to spend valuable time curating and sharing data sets. Some even look unfavourably on open-science practices: a 2016 editorial[1] in *The New England Journal of Medicine* referred to scientists who use data collected by others as "research parasites" and called for data sharing to happen "symbiotically, not parasitically".

Those who want to make their data more open face a bewildering array of options on where and how to share it. They might also lack necessary expertise in data curation and metadata (information that describes a data set). Such expertise can help to ensure that the data they plan to share are useful for others.

However, opening up data can yield benefits: it can catalyse new collaborations, increase confidence in findings and generate goodwill among researchers. Joining Crowther's database enabled Daniel Piotto, a forest ecologist at the Federal University of Southern Bahia in Ilhéus, Brazil, to generate insights into forests not just in Brazil but worldwide. Data sets are becoming easier to cite, and are often accompanied by a digital object identifier (DOI) that makes them independently discoverable. This citability enables researchers to get credit for their data sets and to list them on job, tenure and promotion applications. And there are also the less tangible satisfactions of contributing to the scientific enterprise and giving back something of value to the taxpayers who support basic research.

A MOVE TOWARDS OPENNESS

Before the digital era, sharing data typically required sending them to researchers on request. Now, data can be shared instantly with anyone who has an Internet connection. Moreover, advances in measurement technology in many fields have heralded 'big data' that can form the basis of hundreds or thousands of studies.

CERN, Europe's particle-physics laboratory near Geneva, Switzerland, which was the birthplace of the World Wide Web in 1989, has long been a pioneer in open data. Its Zenodo repository, which hosts data sets, computer code and other resources, and appends them with DOIs, is set to form part of the European Open Science Cloud, an upcoming Europe-wide virtual infrastructure for managing scientific data. Sabina Leonelli, a philosopher who studies open science at the University of Exeter, UK, says that the project, which will bring together existing national and institutional repositories, might be the most ambitious of its type, although exactly how scientists will interact with it remains to be determined.

In the United States, the National Center for Biotechnology Information has been a trailblazer: it launched an open genomics repository called GenBank in the early 1990s that now serves around 30 terabytes of data each day to researchers worldwide. In 1994, NASA enacted a formal policy to share its data publicly, and other space agencies have followed suit. As a result, the data from some publicly funded Earth-observation satellites are free and open. In astronomy, in which a few large and expensive telescopes provide much more data than a single researcher could analyse, open data is also the norm. Many governments' weather data are also open.

But not all fields are equal when it comes to data sharing. Neuroscience and biomedicine experiments, for example, often produce only limited amounts of data, says Jack Gallant, a neuroscientist at the University of California, Berkeley. Researchers might invest considerable time in generating a unique data set, such as the seven functional magnetic resonance imaging brain scans that formed the basis of his research group's 2016 paper[2] in *Nature*. In such cases, it can make sense, Gallant says, to get a return on that investment in the form of several publications, before releasing the data to others. And scooping does happen. In 2015, one of Gallant's graduate students lost the potential fruits of about one-and-a-half years' worth of work when a competing research group published a study using a data set that Gallant's team had made public. "That person was pretty devastated," Gallant says. "It made me more aware of the dangers to graduate students."

Gallant, who considers himself a pioneer of open data, was surprised to be on the receiving end of criticism in July for not sharing data. After he opined on Twitter about the drawbacks of a non-open-source computer programming language, a researcher tweeted to ask why he hadn't shared the data from his *Nature* paper. Gallant replied that his team was preparing further papers based on the same data, and that he would release the data "very soon" — he has done so since. But that didn't satisfy his critics. "'We still want exclusivity to publish more papers' isn't a great excuse," another scientist tweeted. "Did you note data restrictions in the manuscript?" he added — a requirement for submitting a manuscript to *Nature* (see go.nature.com/2fd7mnz).

Data sharing can benefit not just the recipients of data, but also the sharers. A 2018 study of such practices in neuroscience revealed that sharers who used data released by others had larger sample sizes in their studies — achieved by using those open data — than did non-sharing scientists[3]. Papers that were based on openly shared data were published in journals of equal impact as often as were those based on non-shared data.

Crowther offered everyone who shared at least a certain volume of data with his forest initiative the chance to be a co-author of a study that he and a colleague led. Published in *Science* in 2016, the paper used more than 770,000 data points from 44 countries to determine that forests with more tree species are more productive[4].

For Piotto, who met Crowther when they were both at Yale, sharing forest data has led to better recognition for his research group. "You don't really get good press with a local study," Piotto says. When he is listed as an author on papers in high-impact journals such as *Nature* or *Science*, he and his colleagues receive a slew of press enquiries. "We get not just more citations, but also TV and magazines and donors know what we're doing here," he says. "It has a global impact. That's super cool."

Those who decided not to share their data missed out on a chance to co-author a high-profile publication — and to contribute to a project that is larger in scope than a typical researcher's solo efforts, Piotto and Crowther say. But they still have trouble selling the idea to some colleagues. "They go, 'I've spent my life collecting it, why should I share it with you?'" Crowther says, adding that he understands their concerns.

GOOD INTENTIONS BUT MIXED RESULTS

Support for open science is growing among researchers and across disciplines, says Carol Tenopir, an information scientist at the University of Tennessee, Knoxville. In the past decade, she has led three surveys of more than 2,000 scientists worldwide, who were asked about their data-sharing practices as part of the Data Observation Network for Earth project, which is funded by the US National Science Foundation. Researchers are now more aware of good data practice than when she started the surveys, she says.

Certain fields have developed a culture of openness. Up to 96% of environmental scientists and ecologists say that they are "willing" to share data, Tenopir has found. By contrast, psychologists and educational researchers share their data less often, although more than half say that they are willing to make at least some of their data available. But fewer than half of the scientists surveyed actually deposit data in open-access repositories. "There is a mismatch between attitude and behaviour," says Tenopir. "You feel good about [data sharing] but you don't actually do it."

A major hindrance is concern about the legality of sharing data, especially when the research subjects are people, Tenopir has found. Researchers should also consider ethical issues before making data available on, for instance, rural villages or local environmental factors in low-income countries, which could compromise the privacy or well-being of residents, adds Leonelli.

But there are ways to share data gathered

from people safely and legally. For many psychology studies, such as those that involve people answering surveys, de-identifying data can be straightforward, says Simine Vazire, a psychologist at the University of California, Davis. Researchers need only to remove participants' names, e-mail addresses and any other personal information, and then to share only the survey responses.

For data that could be used to identify even anonymous study participants, there are techniques to perturb data sets that ensure that useful information is still accessible. Another option is to use a secure repository that restricts access to qualified requesters.

Scientists who work with data gathered from people or clinical samples should explicitly tell their institutions' ethics committees that they plan to make their data open, says Tenopir. For a previous project, she failed to do so, and was therefore unable to share the data. Without the protection of a dedicated archive, she ended up losing the data.

To avoid that happening again, Tenopir includes her data-sharing plans in proposals for experiments and archives data in Dryad, a non-profit digital repository run by scientific institutions and publishers.

OPEN SCIENCE BY DESIGN

Inadequate resources and training also inhibit data sharing, says Alexa McCray, a researcher in knowledge representation at Harvard Medical School in Boston, Massachusetts. "We're still lacking good tools," she says. "People who want to participate sometimes find it difficult." For example, when scientists record data in paper notebooks or in spreadsheets, they must then choose whether to invest further time and energy in curating and sharing the data at the end of a project, or to start a fresh project.

The key, McCray says, is to practise "open science by design" — the theme of a 2018 National Academies of Sciences, Engineering, and Medicine report, for which McCray chaired the committee. For example, many researchers now keep data, computer code and other materials in web-based, interactive tools such as the popular Jupyter electronic notebook, which makes online archiving much easier. Some enthusiastic practitioners of open science even share data in real time as it is collected.

Choosing an appropriate database is important. Colleagues might be more likely to search discipline-specific repositories, and some repositories provide sophisticated data structuring, whereas others are simply holding places for spreadsheets, interview transcripts or other documents. McCray recommends that researchers talk to grant-programme officers about choosing a repository that complies with their funders' regulations. Many funders, including the National Science Foundation, have made data-curation and repository fees allowable expenses on grants.

Ecologist Daniel Piotto gained insights into forests worldwide through data sharing.

Scientists should also learn how to curate data so that they are more useful to others — for example, by including metadata that make clear what data sets comprise. That skill is not typically covered in graduate-training programmes, notes Leonelli. "Most researchers are absolutely not trained in open data and how to curate data."

For researchers who wish to learn more, the OpenAIRE and FOSTER Plus projects, both funded by the European Union, provide training resources online. An EU project called ORION, coordinated by the Centre for Genomic Regulation in Barcelona, Spain, is also developing a set of open-science training materials. Many university libraries offer data-stewardship training; researchers who want to learn about open-data practices should start there, McCray recommends.

RECOGNITION NEEDED

For the open-data movement to progress, institutions must recognize and reward the production of data by considering it when hiring, offering tenure to and promoting researchers, say advocates for open science. "As long as we have the academic system set up the way we have, it's really difficult to share data," says Luiza Bengtsson, a biochemist who works in communications at the Max Delbrück Center for Molecular Medicine in Berlin. "Right now, it's about competition; sharing data is about collaboration."

Incentive structures that promote data sharing are starting to appear. Altmetric, an online platform that tracks data on the impact of research, is helping to provide researchers with quantitative measures that could lessen the influence of journal impact factors on evaluations of researchers' productivity. (Altmetric is in the portfolio of Digital Science, which is part of Holtzbrinck, the majority shareholder in *Nature*'s publisher, Springer Nature.)

Journals and funding agencies are also playing a part. Last year, *Nature* began to provide statements on data availability in a non-paywalled section of its papers online, according to a spokesperson, and, from this year, requires the authors of papers in Earth, space and environmental sciences to make supporting data available to others through community repositories. Other journals have crafted similar policies.

The Center for Open Science, a non-profit organization in Charlottesville, Virginia, has created a set of web badges that researchers can affix to papers and data sets to highlight that their data are open. The journal *Psychological Science* introduced the badges in 2014; since then, sharing of data from its papers has increased by a factor of ten[5]. More than 50 journals now offer the badges.

Ultimately, data-sharing responsibilities could shift from individuals to their institutions. As data sets continue to grow in size and complexity, universities and research institutions will need to take responsibility for curating and sharing them, says Barend Mons, a molecular biologist at Leiden University Medical Center in the Netherlands, who advises the EU on open science.

"The biggest mistake people are likely to make is trying to train every young researcher to be a half-baked data steward," Mons says. Instead, he suggests that universities should hire one specialist to curate and share data for every 20 researchers. And because the importance of big data will only continue to grow, scientists with data skills will be in demand. Crowther, for example, employs a full-time data manager, although he acknowledges that not all researchers can afford this luxury.

For early-career scientists who prefer producing data to managing it, Mons has this advice: "Go to a university that takes data stewardship seriously." ■

Gabriel Popkin *is a freelance writer in Mount Rainier, Maryland.*

1. Longo, D. L. & Drazen, J. M. *N. Engl. J. Med.* **374**, 276–277 (2016).
2. Huth, A. G. *et al. Nature* **532**, 453–458 (2016).
3. Milham, M. P. *et al. Nature Commun.* **9**, 2818 (2018).
4. Liang, J. *et al. Science* **354**, aaf8957 (2016).
5. Kidwell, M. C. *et al. PLoS Biol.* **14**, e1002456 (2016).

CORRECTION

The Spotlight article 'Scientists get political' (*Nature* **568**, S1–S3; 2019) erroneously located Maxime Gingras at the Canadian Science Policy Centre. He is, in fact, at the Professional Institute of the Public Service of Canada.