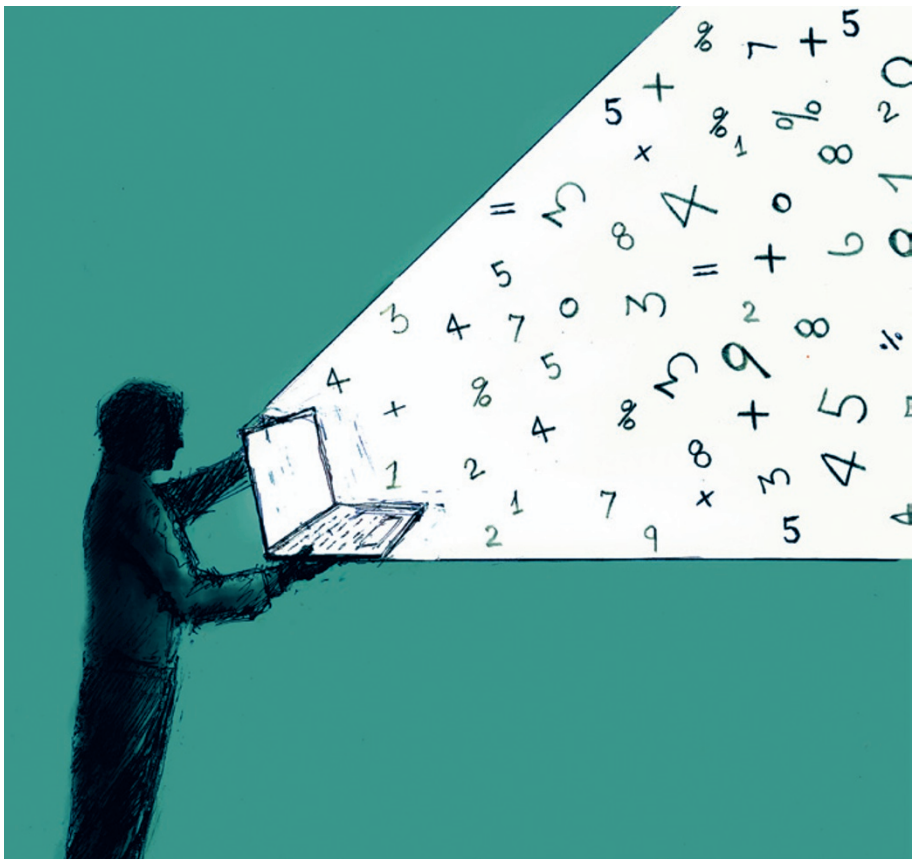
TURNING POINT An interest in climate change bridges science and policy **p.245**

@NATUREJOBS Follow us on Twitter for the latest news and features **go.nature.com/e492gf**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**



GARY WATERS/IKON IMAGES/CORBIS

DATA-SHARING

Everything on display

Researchers can get visibility and connections by putting their data online — if they go about it in the right way.

BY RICHARD VAN NOORDEN

Lizzie Wolkovich always felt she ought to make her research data freely available online. "The idea that data should be public has been in the background through my entire career," she says.

Yet in 2003–09, while she was working on her ecology PhD, there were few incentives for her to share. Sharing would not help her to get grants or publications, and although posting data online was not unheard of, few researchers actually did it, she says. Many preferred to hang on to their hard-won field data, sharing privately if they did so at all.

But after she earned her doctorate, Wolkovich overcame her hesitation, thanks to a combination of helpful colleagues, improved resources and a discernible shift in the research community's attitude. So in 2010, through an online data repository called the Knowledge Network for Biocomplexity, Wolkovich released her doctoral data set — the fruit of thousands of hours spent measuring the diversity of arthropods in 56 experimental soil plots she had set up in the arid scrubscape of southern California. Since then, she has publicized all the data that she has collected, including a meta-analysis of 50 other studies that she examined to see how factors such as rising temperatures affect the life cycles of plants. Wolkovich, now at the University of British Columbia in Vancouver, Canada, says that she herself had never objected to sharing her results — she had just not known how to do so. She likes the fact that her data are now easily accessible to other researchers and anyone else who is interested. "It saves me so much time," she says.

Wolkovich is one of a number of early-career researchers who are enthusiastically posting their work online. They are publishing what one online-repository founder calls small data — experimental results, data sets, papers, posters and other material from individual research groups — as opposed to the 'big data' spawned by large consortia, which usually employ specialists to plan their data storage and release. The many resources now available give researchers options for where and how to post their data, releasing potentially fruitful data sets that used to be locked up in unpublished paper files, buried in journal-article appendices or hidden away on scientists' hard drives.

OPENING UP

Open data-sharers are still in the minority in many fields. Although many researchers broadly agree that public access to raw data would accelerate science — because other scientists might be able to make advances not foreseen by the data's producers — most are reluctant to post the results of their own labours online (see *Nature* **461,** 160–163; 2009). When Wolkovich, for instance, went hunting for the data from the 50 studies in her meta-analysis, only 8 data sets were available online, and many of the researchers whom she e-mailed refused to share their work. Forced to extract data from tables or figures in publications, Wolkovich's team could conduct only limited analyses.

Some communities have agreed to share online — geneticists, for example, post DNA sequences at the GenBank repository, and astronomers are accustomed to accessing images of galaxies and stars from, say, the Sloan Digital Sky Survey, a telescope that has observed some 500 million objects — but these remain ▶

the exception, not the rule. Historically, scientists have objected to sharing for many reasons: it is a lot of work; until recently, good databases did not exist; grant funders were not pushing for sharing; it has been difficult to agree on standards for formatting data and the contextual information called metadata; and there is no agreed way to assign credit for data.

But the barriers are disappearing, in part because journals and funding agencies worldwide are encouraging scientists to make their data public. Last year, the Royal Society in London said in its report *Science as an Open Enterprise* that scientists need to "shift away from a research culture where data is viewed as a private preserve". Funding agencies note that data paid for with public money should be public information, and the scientific community is recognizing that data can now be shared digitally in ways that were not possible before. To match the growing demand, services are springing up to make it easier to publish research products online and enable other researchers to discover and cite them. There are so many, in fact, that choosing where and how to publish data sets and other supplementary material can be confusing (see 'Abundant options').

"Lots of people are getting into data-hosting, and I think it will be tricky to decide where to put your data," says Heather Piwowar, who studies data-sharing for the US National Evolutionary Synthesis Center in Durham, North Carolina.

SHARE AND SHARE ALIKE

Although exhortations to share data often concentrate on the moral advantages of sharing, the practice is not purely altruistic. Researchers who share get plenty of personal benefits, including more connections with colleagues, improved visibility and increased citations. The most successful sharers — those whose data are downloaded and cited the most often — get noticed, and their work gets used. For example, one of the most popular data sets on multidisciplinary repository Dryad is about wood density around the world; it has been downloaded 5,700 times. Co-author Amy Zanne, a biologist at George Washington University in Washington DC, thinks that users probably range from climate-change researchers wanting to estimate how much carbon is stored in biomass, to foresters looking for information on different grades of timber. "I would much prefer to have my data used by the maximum number of people to ask their own questions," she says. "It's important to allow readers and reviewers to see exactly how you arrive at your results. Publishing data and code allows your science to be reproducible."

Even people whose data are less popular can benefit, adds Piwowar. By making the effort to organize and label files so that others can understand them, scientists become more organized and better disciplined themselves, and can avoid confusion later on. "It is often very hard to find and understand your own work if you are looking at it years from now," says Piwowar. Scientists might be inclined to stuff their data into folders that can get lost and muddled — but if they store the files in an online repository, they are forced to curate and collate the data, she says.

The fear of being scooped is a powerful inhibitor. But scientists can put an embargo on their data, so that only they can see the work until they are ready to make it public. And data sets are becoming increasingly citable, bringing their authors formal recognition: data published in a data journal, on Dryad or on the repository figshare.com are given a digital object identifier (DOI) that can be referenced in other publications. (Figshare is owned by Digital Science, a sister company to Nature Publishing Group.)

Would-be sharers often worry that their data are too disordered or shoddy to release into the world. "I make my data available, and it can be a pain. I'm also scared and embarrassed about errors — most of us are, especially early-career scientists," says Piwowar. "We don't yet have a culture of forgiveness around that, unlike in computer programming, where everyone knows there are bugs in code." She advises researchers to look into repositories to get a sense of the quality standard for experimental data. "It doesn't have to be perfect," she says. "It's probably less thorough than you think."

HEATHER PIWOWAR

"Lots of people are getting into data-hosting, and I think it will be tricky to decide where to put your data."
Heather Piwowar

As sharing grows more common, scientists may worry less about posting data sets. "Ultimately, data will be so ubiquitous that we will no longer be in a world where researchers are so scared," says Carl Boettiger, an ecologist at the University of California, Santa Cruz, who keeps his entire laboratory notebook open online (see *Nature* **493,** 711; 2013). "At the end of the day, science is a social process. You will never get there hiding yourself and your work," he adds.

THE RIGHT PLACE

Depositing data on a personal website is unlikely to be the best way to get it reused and cited. For a start, the website may not be around in five years, says William Michener, director of e-science initiatives at the University of New Mexico in Albuquerque. Michener is principal investigator for a multinational programme called DataONE, which is funded by the US National Science Foundation and promotes best practices to scientists as part of its aim to make data more discoverable. Journal publishers back up their research papers with the help of non-profit archiving services such as Portico and CLOCKSS, which are financed by participating libraries and publishers, and which store material on a number of servers so that it will not disappear if a publisher goes bankrupt. Some data publishers have similar contingency plans, and Piwowar recommends looking into them. If no back-up plans are in place, she says, "it suggests they haven't prioritized well enough how to steward their data".

Just as important as sharing data publicly is making sure that other researchers can understand them. Susanna Assunta-Sansone, associate director of the Oxford e-Research Centre at the University of Oxford, UK, says that putting out data without noting what it means will ensure that "it's not really reusable". To avoid this, researchers must choose appropriate metadata: descriptions of the data's content and how they are arranged and set up. This type of curation is useful not just for human readers, but also for computer programmes that might be used to search

WHERE AND HOW

Abundant options

Online data repositories are proliferating: the searchable catalogue Databib lists 594 websites. Hundreds are specialists, devoted to particular kinds of data. But general-purpose repositories do exist: they include Dryad, which many scientists use to store the data underlying their publications; GitHub, which is usually used to host software code and to collaborate on developing it, but also hosts other data; European Commission repository ZENODO; and figshare.com, a general repository for posters, papers and data sets that welcomes negative results that would otherwise never be published.

Publishers have started to launch journals dedicated to data sets and descriptions of data, such as BioMed Central's *GigaScience*. Some scientists post data on social networks such as ResearchGate or Academia.edu.

Each discipline is evolving its own ways to structure data and metadata. In biology alone, biosharing.org lists some 530 standards, including MIAME (Minimum Information About a Microarray Experiment) and PDB (Protein Data Bank format). To avoid confusion, researchers should familiarize themselves with the best practices in their fields. R.V.N.

through or connect data sets. Intelligent searches often rely on whatever descriptive metadata researchers have attached to the data. The metadata are read by an application programming interface (API), a set of commands that computer programmes use to interact with data stores and pull information from them. Not all data repositories use APIs; those that do not may not be the best places to store or release information, because it could be hard for anyone to find.

Sites that are dedicated to hosting particular types of data, such as DNA sequences, usually tell submitters what format is appropriate. They may require data to be entered using an online form or following specific instructions. By contrast, generalist sites — such as institutional repositories, data journals or ventures similar to figshare.com — may have looser requirements. This has the potential to result in a blizzard of different formats and descriptive tags, which could make discovering and reusing data more difficult, so researchers should pay close attention to the norms in their fields.

Decisions about metadata standards should be made early in a research project, says Michener. DataONE has provided a primer on best practices, as has a tool called DataUp, run through the University of California Curation Center in Oakland to help researchers to create data packages that are good enough to put online. Other aspects of data-sharing to consider early on include the information's sensitivity and whether some parts must be stripped out to avoid, for example, identifying human study participants or the locations of endangered species. Researchers also need to be clear about whether they will allow their data sets to be used for any purpose, or whether they would like to limit reuse to, for example, non-commercial applications. One widely understood way of documenting reuse rights is by giving the data one of several different Creative Commons licences.

"At the end of the day, science is a social process. You will never get there hiding yourself and your work."

Ultimately, says Michener, early-career researchers need to pay attention to new and developing ways to share data, and to the standardized formats that are emerging to make data easier to search and discover. Those who do not, he says, should rethink why they are doing research. "I think we are just now reconnecting with what science is all about — not just creating new knowledge, but also sharing the information and data that underpins those discoveries." ■

Richard Van Noorden *is a senior reporter at* Nature.

TURNING POINT
Kevin Gurney

Sustainability scientist Kevin Gurney has been studying climate change for 27 years. He has worked in academia, public policy, non-governmental organizations (NGOs) and think tanks, and is currently at Arizona State University in Tempe. He describes how he navigates the science–policy divide.

What convinced you to do a graduate degree?
As an undergraduate, I worked at the Lawrence Berkeley National Laboratory in California, taking spectroscopic measures of greenhouse gases. Working with wonderful mentors who were excited about the science was infectious. Later I did a master's in atmospheric science at the Massachusetts Institute of Technology in Cambridge and my focus shifted to chemistry and chlorofluorocarbons (CFCs) — greenhouse gases that also deplete Earth's ozone layer, and so have science and policy implications.

How did you become active in policy?
Regulation was ramping up to stop production of fully fluorinated CFCs, and industry was looking for alternatives. In 1986, I found that compounds called HCFCs, which contained less chlorine and thus caused less ozone depletion, still had the heat-trapping properties of CFCs. The policy implications were huge and there was so much misinformation. I was thinking, people need to know about this. I got more involved with policy at that point.

Why not go on immediately to pursue a PhD?
I wanted to work on the political implications first. In 1992, I started working with the Institute for Energy and Environmental Research in Takoma Park, Maryland. We sued the US Environmental Protection Agency to get it to regulate HCFCs, and we spread the word that HCFCs were not as environmentally friendly as manufacturers claimed. I also got involved in discussions on the Montreal Protocol, the treaty to regulate ozone-depleting chemicals. I realized how ineffectively science and policy interacted. I got a master's in public policy at the University of California, Berkeley, then a PhD in ecology at Colorado State University in Fort Collins. These days it is easier to get an interdisciplinary degree, but I tell my students that some degrees lack a rigorous science foundation. There is no substitute for a solid mathematics and physics background — it gives you credibility.

How did you move from CFCs to carbon?
I attended the negotiations in London and Copenhagen to amend the Montreal Protocol, laying out a plan to manage CFC phase-out. Once the treaty was set, I began to see that rising carbon dioxide levels were an interesting problem. I maintained a personal network of contacts in NGOs, and many organizations were shifting to carbon dioxide and climate research for exactly the same reasons I was — it was quickly gaining traction. NGOs, including the US branch of the conservation group WWF in Washington DC, paid for me to go to Kyoto Protocol negotiations, and I worked pro bono as a science consultant. I told the NGOs I was not going to give anyone just a line they wanted to hear. My PhD adviser let me take vacation to attend negotiations every four months.

What is climate-change negotiation like?
It is the most intense, pressure-filled world you can imagine. I was very involved with language in the Kyoto Protocol about the missing carbon sink — the carbon dioxide absorbed on land, which is not fully understood — and how to account for it. I learned a lot about law during my policy degree, which made me effective in crossing the divide between policy and science. You don't have to dumb down; you have to learn how legislators and policy-makers view science.

You won a Faculty Early Career Development award from the US National Science Foundation in 2009. How are you using it?
I'm doing a risky thing and getting involved with citizen science to use Google Earth to identify power plants (see *Nature* http://doi.org/nb3; 2013). Normally I would be too worried that it would fail to use funding dollars. But we have thousands of people involved and are adding hundreds of power plants to an emissions database that is part of NASA's pilot carbon-monitoring system. It is of interest to climate scientists, social scientists and policy-makers. ■

INTERVIEW BY VIRGINIA GEWIN