Empty archives

Most researchers agree that open access to data is the scientific ideal, so what is stopping it happening? **Bryn Nelson** investigates why many researchers choose not to share.

ILLUSTRATIONS BY J. H. VAN DIERENDONCK

In 2003, the University of Rochester in New York launched a digital archive designed to preserve and share dissertations, preprints, working papers, photographs, music scores — just about any kind of digital data the university's investigators could produce. Six months of research and marketing had convinced the university that a publicly accessible online archive would be well received. At the time of the launch, the university librarians were worried that a flood of uploaded data might swamp the available storage space.

Six years later, the US$200,000 repository lies mostly empty.

Researchers had been very supportive of the archive idea, recalls Susan Gibbons, vice-provost and dean of the university's River Campus Libraries — especially as the alternative was to keep on scattering their data and dissertations across an ever-proliferating array of unintegrated computers and websites. "So we spent all this money, we spent all this time, we got the software up and running, and then we said, 'OK, here it is. We're ready. Give us your stuff'," she says. "And that's where we hit the wall." When the time came, scientists couldn't find their data, or didn't understand how to use the archive, or lamented that they just didn't have any more hours left in the day to spend on this business.

As Gibbons and anthropologist Nancy Fried Foster observed in their 2005 postmortem[1], "The phrase 'if you build it, they will come' does not yet apply to IRs [institutional repositories]."

A similar reality check has greeted other data-sharing efforts. Most researchers happily embrace the idea of sharing. It opens up observations to independent scrutiny, fosters new collaborations and encourages further discoveries in old data sets (see pages 168 and 171). But in practice those advantages often fail to outweigh researchers' concerns. What will keep work from being scooped, poached or misused? What rights will the scientists have to relinquish? Where will they get the hours and money to find and format everything?

> **"We got the software up and running and said 'Give us your stuff'. That's when we hit the wall."**
> — Susan Gibbons

Some communities have been quite open to sharing, and their repositories are bulging with data. Physicists, mathematicians and computer scientists use arXiv.org, operated by Cornell University in Ithaca, New York; the International Council for Science's World Data System holds data for fields such as geophysics and biodiversity; and molecular biologists use the Protein Data Bank, GenBank and dozens of other sites. The astronomy community has the International Virtual Observatory Alliance, geoscientists and environmental researchers have Germany's Publishing Network for Geoscientific & Environmental Data (PANGAEA), and the Dryad repository recently launched in North Carolina for ecology and evolution research.

But those discipline-specific successes are the exception rather than the rule in science. All too many observations lie isolated and forgotten on personal hard drives and CDs, trapped by technical, legal and cultural barriers — a problem that open-data advocates are only just beginning to solve.

One of those advocates is Mark Parsons at

the National Snow and Ice Data Center at the University of Colorado in Boulder. Parsons manages a global programme to preserve and organize the data produced by the International Polar Year (IPY) that ran from March 2007 to March 2009 and included an estimated 50,000 collaborators from more than 60 countries.

The IPY policy calls for data to be made available fully, freely, openly and on the shortest feasible timescale. "Part of what is driving that is the rapidness of change in the poles," says Parsons. "If we're going to wait five years for data to be released, the Arctic is going to be a completely different place."

Reality bites

But reality is forcing a longer timescale. As soon as they began implementing the data policy, Parsons and his team encountered a staggering diversity of incoming information, as well as wide variations in the culture of data sharing. Fields such as atmospheric science and oceanography, Parsons says, have well-developed traditions of free and open access, and robust databases. But fields such as wildlife ecology and many of the social sciences do not. "What we discovered was that this infrastructure to share the data doesn't really exist, so we need to start creating that," Parsons says.

But his programme lacks the resources required to create that infrastructure on a large scale. So the team has resorted to preserving as much data as it can. It has delegated much of that job to national coordinators, or "data wranglers", as Parsons calls them, who contact investigators and, "get the data branded and put in the IPY corral".

One of the most successful data-wrangling countries has been Sweden, which formed a subcommittee to correct its early lag in collecting and then received national funding for its own IPY data archive. National coordinator Håkan Olsson, a specialist in remote sensing at the Swedish University of Agricultural Sciences in Umeå, says that the country's archive is helping to house data from smaller, independent projects that would never reach large international databanks.

Nevertheless, he says, many Swedish researchers still don't archive their data, or don't put data in formats that make them easily searchable and retrievable. He faults the funding agencies too. "Unlike some other countries," he says, "the research councils in Sweden do not yet have a practice to grant funds with the condition that data from the project is sent to a data centre."

Even when wranglers can identify the data, it is not always obvious where the data should go. For example, says Parsons, "you would think that any snow and ice data would go into the National Snow and Ice Data Centre". But the centre's funding is generally tied to specific data streams, he says, which means it can find itself in the position of accepting glacial data from a programme it has money for, while being forced to turn away similar glacial data from programmes where it does not.

Despite the launch earlier this year of the Paris-based Polar Information Commons to make polar data more accessible, Parsons says, that with all the "naive assumptions", the lack of planning and other unanticipated obstacles, properly managing the IPY data will require another decade of work.

In other fields, however, the main barriers to data sharing are concerns about quantity and quality. The US National Science Foundation's (NSF's) Laser Interferometer Gravitational-Wave Observatory (LIGO), for example, uses giant detectors in Louisiana and Washington to search for gravitational waves that might indicate the presence of rare phenomena such as colliding black holes or merging stars. LIGO is also working with the Virgo consortium, which operates a similar detector near Pisa, Italy.

> **"We don't just have to analyse the data, we need to make sure the data are right."**
> — Szabocs Márka

Neither team has detected the signal they are looking for yet — but that's not surprising: gravitational waves are expected to be extraordinarily faint. The key to detecting them is to eliminate every possible source of spurious vibration in the detectors, whether from seismic events, electrical storms, road traffic or even from the surf on distant beaches. It requires what Szabolcs Márka, a physicist at Columbia University in New York and the university's lead scientist for LIGO, calls "a really paranoid monitoring of the environment".

The question of what data should be shared has provoked strong debate within the LIGO and Virgo teams. Should they open up all their terabytes of data to outside scientists, including the torrents of environmental data? Or should they release just the cleaned-up data stream most likely to reveal a gravity wave? Would naive outsiders fail to process the raw data adequately, leading to premature announcement of gravitational wave 'discoveries' that would hurt everyone's credibility? Or would the extra eyes bring fresh perspective to the search?

"I'm torn," says Márka, who says that the precise terms of data sharing are being negotiated with the project's funders. "We don't just have to analyse the data, we need to make sure the data are right."

How data should be shared is also a substantial problem. A prime example is the issue of data standards: the conventions that spell out exactly how the digital information is formatted, and exactly how the contextual information (metadata) is listed.

In some disciplines it is comparatively easy to agree on standards, says Clifford Lynch, executive director of the Coalition for Networked Information based in Washington DC, which represents academia on data and networking issues. "If you look at something like the sequencing of a genome, there's a whole lot of tacit stuff that's already settled," he says. "Sequencing one genome is very similar to sequencing another." But for other groups — say, environmental scientists trying to understand the spread of a pollutant — the choice of common standards is far less obvious.

The all-too-frequent result is fragmented and often mutually incomprehensible scientific information. And that, in turn, stifles innovation, says James Boyle, a law professor at Duke University in Durham, North Carolina, and a founding board member of Creative Commons, a non-profit organization that supports creative content sharing.

Always somebody smarter

"Researchers generally create their own formats because they believe that they know how their users want to use the data," says Boyle. But there are roughly a billion people with Internet access, he says "and at least one of them has a smarter idea about what to do with your content than you do". For example, web users are using applications such as Google Earth to plot the spread of pandemics[2] or to collect information on the effects of climate change. All that is needed, says Boyle, are common languages and formats for data.

Perhaps not surprisingly, data-sharing advocates say, the power to prod researchers towards openness and consistency rests largely with those who have always had the most clout in science: the funding agencies, which can demand data sharing in return for support; the scientific societies, which can establish it as a precedent; and the journals, which can make sharing a condition of publication.

The trick is to wield that power effectively. The NSF, for example, has funded groundbreaking research into digital archiving, search and networking technologies. But its data-sharing policies for standard research grants, for example, have come under fire for being scattered and ad hoc; they are often stipulated on a per-project basis. Gibbons says she is especially disappointed with a 2003 mandate by the US National Institutes of Health (NIH), which could have dramatically changed the culture of data sharing. The mandate does require a

data-sharing plan for any grant worth $500,000 or more in direct annual costs or an explanation of why sharing isn't possible. But details about how to make the data available were so vague, says Gibbons, that researchers soon stopped paying attention, content to sit back until someone got in trouble for not playing by the rules.

Officials at the NIH Office of Extramural Research reply that the data-sharing policy's 'vagueness' is, in fact, flexibility, an attempt to avoid forcing every research programme into a one-size-fits-all straightjacket. They note that the policy also recognizes that there may be valid reasons for not sharing, including concerns about patient privacy and informed consent.

The chicken or the egg?

Nonetheless, until data sharing becomes a requirement for every grant, says Daniel Gardner, a physiologist and biophysicist at the Weill Medical College of Cornell University, "people aren't going to do it in as widespread of a way as we would like". Right now, he says, "you can't ask large numbers of people to do it, because it's a lot of work and because in many cases the databases don't exist for it. So there is kind of a chicken and egg problem here."

One solution would be for agencies to invest in the infrastructure necessary to meet their archiving requirements. That can be difficult to arrange, says Boyle. "Infrastructure is the thing that we always fail to fund because it's kind of everybody's problem, and therefore it's nobody's problem." Yet some agencies have been pioneers in this area. One often-cited example is the Wellcome Trust, the largest non-governmental UK funder of biomedical research. Since 1992, its Sanger Institute near Cambridge has been developing and housing some of the world's leading databases in genomics, proteomics and other areas.

Another prominent example is the NIH's National Library of Medicine, which in 1988 established the National Center for Biotechnology Information (NCBI) to manage its own collection of molecular biology databases, including the GenBank repository. James Ostell, chief of the NCBI's Information Engineering Branch, likes to show a colour-coded timeline of contributions to GenBank since its founding in 1982 — a progression that dramatizes the fast-evolving history of genetic sequencing. Ostell points out thick waves of colours flowing from the left side of the chart. Representing traditional sequence divisions such as viruses, rodents, primates, plants and bacteria, they dominated GenBank's contents for years. Other sequences, produced by faster techniques, began to put in appearances in the mid 1990s. Then in late 2001 a sudden surge of green, representing DNA snippets derived from whole-genome shotgun sequencing, quickly took over. By 2006, the green accounted for more than half of the database's contents.

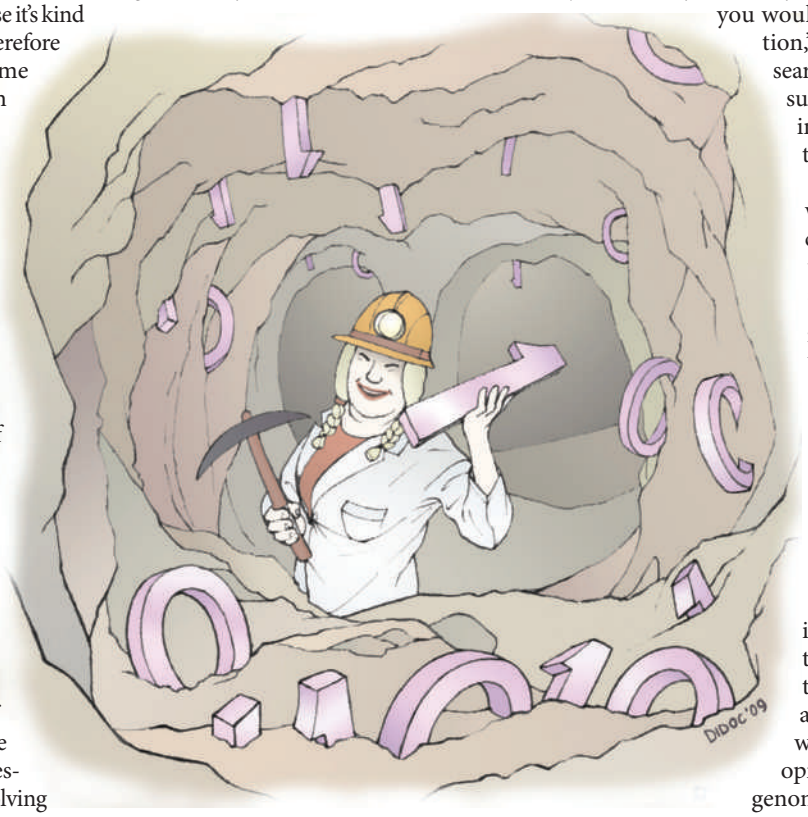
> **"At least one of the people out there has a smarter idea about what to do with your content than you do."**
> — James Boyle

Keeping up with ever-shifting technology has created its own set of challenges, says Ostell. "Nobody has infinite resources. And storing electronic information over time is a dynamic process. If you try to look at a file that you wrote with a word processor 20 years ago, good luck." In the same way, if a data set isn't readable by the latest version of a database, it isn't usable. So an archive may well have to choose between tossing old data out, and paying to preserve the out-of-date software required to make sense of them.

Even more challenging are the legal minefields surrounding personal data and privacy. The need to protect human subjects has led to starkly different approaches. Some projects openly share data, whereas others require researchers to navigate a labyrinthine approval process before granting access. The NCBI has tried to build such requirements into its newer databases. A case in point is its database of Genotype and Phenotype (dbGaP), which archives and distributes the results of genome-wide association studies, medical DNA sequencing, molecular diagnostic assays and almost anything else that relates people's traits and behaviours to their genetic makeup. The dbGaP allows open access to summaries and other forms of information that have been stripped of personal identifiers. But it grants controlled access to personal health information only after a researcher has been approved by a formal review committee.

Novel meaning

Such measures can be cumbersome, says Ostell. Yet the benefits of sharing far outweigh the costs. Some of GenBank's early sequences, for example, included genes from yeast and *Escherichia coli* labelled as DNA repair enzymes. Years later, researchers studying human colon cancer made a link between mutations in patients and those same enzymes[3]. "If you just did a literature search, you would never make that connection," Ostell says. "But when you search on the basis of their genes, suddenly you connect meaning in a way that's novel, which is the basis of discovery."



Sharing is obviously easier when the expectations are clear, and many scientists point to a 1996 meeting in Bermuda as a defining moment for genomics. At the meeting, leaders working on the Human Genome Project hammered out a set of agreements known as the Bermuda principles. Chief among them was the stipulation that sequences longer than 1,000 base pairs be made publicly available, preferably within 24 hours.

The Bermuda principles, in turn, built on the foundations laid a decade earlier by the editors of journals such as *Nucleic Acids Research*, who spurred the early development of GenBank and other genomic repositories by requiring

researchers to deposit their data there as a precondition for publishing. Newer journals, such as the open-access Public Library of Science journals, have made publication contingent on making the data "freely available without restriction, provided that appropriate attribution is given and that suitable mechanisms exist for sharing the data used in a manuscript". The journal *Neuroinformatics* devoted its September 2008 issue to data sharing through the NIH Neuroscience Information Framework. *Ecological Archives* publishes appendices, supplements and data — related to studies appearing in other ecology journals — which include the metadata needed to interpret them. (*Nature* journals require authors "to make materials, data and associated protocols promptly available to readers without preconditions".)

Yet the journals' power to compel data sharing and scientific culture change is not absolute. In March 2009, for example, the journal *Epidemiology* felt able to call only for a "small step" towards more openness. "We invite our authors to share their data and computer code when the burden is minimal," said an editorial[4] in that issue.

"We believe that data sharing is a matter of time," says Miguel Hernán, an epidemiologist at Harvard University and a co-author of the editorial. But prematurely forcing a sharing requirement on authors "would be suicidal", he warns, especially with unresolved concerns over patient confidentiality. They would simply submit their papers somewhere else.

Another issue facing journals and data banks is how to ensure proper citations for data sets. "The one thing that people clearly care about in the sciences is attribution," says Boyle. Without an agreed-on way of assigning credit for original data falling beyond the parameters of a publication, however, it's no wonder that scientists are reluctant to share: their hard work may never be recognized by their employers or by granting agencies. Worse yet, it could be poached or scooped.

This is one place that technology might help, says Boyle. He points to a music site associated with Creative Commons known as ccMixter, in which users can upload an a capella chorus, a bass line, a trumpet solo or other musical samples. Users are free to remix the samples into new tracks. But when they do, the program automatically keeps a continuous credit record.

So why not implement a similar system that would add a link back to a database every time a researcher repurposed some data? It wouldn't necessarily solve the problem of scooping, Boyle says, "but it aligns the social incentives with the individual incentives". It could also provide a feasible way for universities or funding agencies to track the value of a researcher's data.

International agreement

Other Creative Commons tools are already making their way into international scientific agreements. In May, for example, Creative Commons' CC0 licence was endorsed by participants at a meeting in Rome on resource and data sharing within the mouse functional genomics community. The licence, which allows its users to "waive all copyrights and related or neighbouring rights" and thereby share more of their work, has been translated into dozens of languages.

As welcome as such developments are, however, Boyle points out that the creation of the legal and technical infrastructure to accommodate researchers' data-sharing concerns is a huge task, and should not be left solely to non-profit organizations and individual universities. Nor should it be left to the funding agencies' grant-by-grant allocations for data sharing. It will require major government investments, starting with demonstration projects to explore how sharing can best be done. "What we need is a working example that you can point to," he says.

If William Michener has his way, a virtual data centre funded by the NSF and hosted by his university will be one of those examples. DataONE (Data Observation Network for Earth) exists only on paper, but a five-year, $20-million grant through the NSF's DataNet programme will help to turn it into an open-access database focusing on biology, ecology and environmental science data. Four other $20-million archives are planned under DataNet's first phase.

Michener, director of e-science initiatives for University Libraries at the University of New Mexico, Albuquerque, and a leader of DataONE, says that the archive is designed to accommodate many of the orphan data sets that have yet to find a home, and will target resource-strapped colleges, field stations, and individual or small teams of scientists. In the longer term, the DataONE consortium, which encompasses two dozen partner institutions in the United States, the United Kingdom, South Africa, Australia and Taiwan, will explore business models that could sustain the archive well beyond its initial grant and potential five-year renewal. Among the plans under consideration are a fee-for-service set up, a membership requirement for participating entities and the solicitation of external grants for education and outreach.

DataONE's success, however, may depend on overcoming the same ambivalence among researchers that has bedevilled the University of Rochester and other builders of public databases. Although a strategy is still being worked out, Michener envisions a combination of workshops, seminars, websites and other educational tools to help clarify the how and why of sharing. But one archive can only do so much. Larger efforts will be required to tackle what Michener sees as the overriding challenge: "Changing the culture of science from one where publications were viewed as the primary product of the scientific enterprise to one that also equally values data."

"We need to change the culture of science to one that equally values publications and data."
—William Michener

Without that cultural shift, says Gibbons, many digital archives are likely to remain little more than stacks of empty shelves. ■

Bryn Nelson is a freelance science and medical writer based in Seattle, Washington.

1. Foster, N. F. & Gibbons, G. *D-Lib Magazine* doi:10.1045/january2005-foster (2005).
2. www.nature.com/avianflu/google-earth/index.html
3. Marra, G. & Boland, C. R. *Gastroenterol. Clin. North Am.* **25,** 755–772 (1996).
4. Hernán, M. A. & Wilcox, A. J. *Epidemiology* **20,** 167–168 (2009).

See Opinion, pages 168 and 171, and online special at: http://tinyurl.com/dataspecial.