Optimized data logistics

Accessible storage of scientific data is usually mandated, but not often achieved. The task needs people who are interested in information technology and regard it as their primary focus.

Data that underpin published research must be accessible. There is no disagreement with this statement in the geosciences. Unfortunately, reality has not caught up with good will. A streamlined way of ensuring that data are made accessible, and in user-friendly formats, is sorely missing. As a result, evasion strategies and excuses for keeping data in closed, local archives proliferate.

A Commentary on page 575 of this issue argues that the concept of data archiving in science — currently an afterthought following analysis, processing and publication — needs to be turned on its head. That is, the question of how to store scientific data must be solved before any data are even collected, and further manipulations need to be automated as far as possible.

The basic idea is familiar. In information technology — of which scientific data storage is a small part — manual processes are to be avoided. They reduce efficiency, introduce error and create bottlenecks. Therefore, once measurements have been made and entered into a software interface, subsequent manipulations ought to be automated where possible. The complete workflow can then be captured in a replicable manner. And when the time for deposition comes, the workload for the scientist should be minimal.

Note that the path towards more efficient data management suggested on page 575 of this issue has been devised by information scientists who work with geoscientists, rather than the Earth or planetary science researchers themselves. Geoscience data have long become complex and voluminous, not only where climate models and satellite-based instruments are involved (see *Science* **331,** 700–702; 2011). The problem of managing these data sets efficiently, and ensuring that they are organized in an accessible manner, has reached a dimension that requires dedicated professionals.

But the information technology infrastructure of universities and research institutions is usually ill equipped to undertake the management of huge data sets, let alone come up with creative solutions for handling and storage. This is particularly true if the data are generated through external project-based funding that includes only limited technical support.

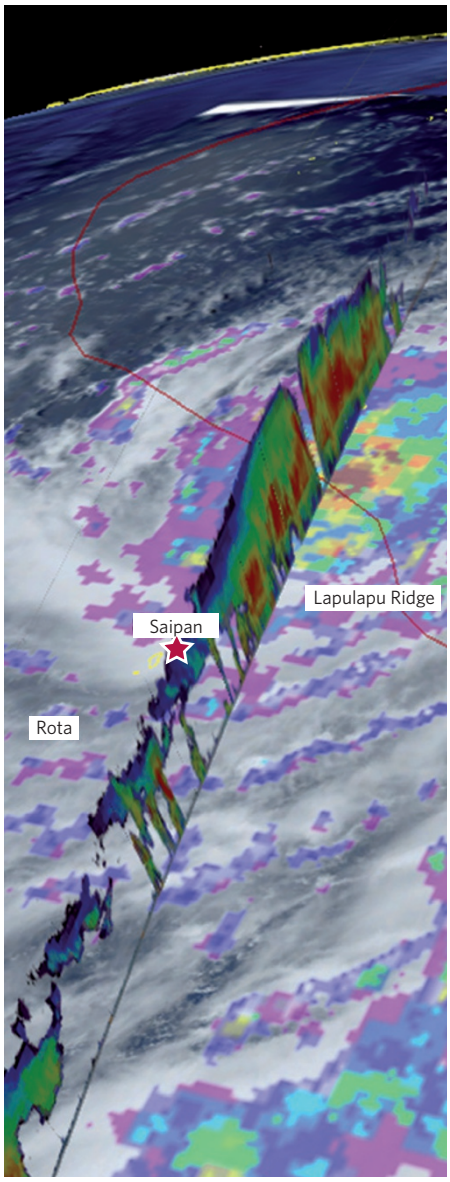
Funding agencies do not usually cover this particular budget gap either, although most of them insist on data deposition as a matter of course. But if a proposal that includes a salary for a data manager is sent back to the applicants with a request to cut the budget, the non-scientific positions will be the first to go. It would make sense for funders who value open-source data to insist that data management positions are ring-fenced, or that information technology departments are asked to collaborate in data-intensive projects.

Indeed, a survey of referees for the journal *Science* revealed that more than 80% of the participants felt that there was insufficient funding for data curation. At the same time, only a miserly fraction of 7.6% of the participants stated that they archive most of their data in community repositories (*Science* **331,** 692–693; 2011), in line with the observation that existing archives are empty (*Nature* **461,** 160–163; 2009).

Yet data storage does not have to be a headache. There are pockets in the Earth and planetary sciences where accessible data storage is already up and running. Archives for satellite data such as the data depot that stores the information obtained by NASA's 'A-Train' satellites (http://go.nature.com/noNVEA), and the repository for the climate model simulations that were used in the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (http://go.nature.com/YdpfQL) are two such examples. But a much more widespread recognition of the scale of the task of data archiving in the geosciences is needed.

As a first step, the community of Earth and planetary scientists needs to develop a set of clear guidelines that detail the types and extent of project- or publication-related data that should be accessible. For the systematic implementation of such guidelines, geoscientists should then seek the support of experts in information technology.

Funding agencies that would like the data obtained with their money to be used to their full potential need to ensure that the data archiving is not left to the PhD students and postdocs who — quite rightly — have other priorities. ❐



Out of the archive. Measurements of reflectivity, cloud-top pressure and rain rate made during super typhoon Choi-wan (15 September 2009) by a combination of satellite instruments on the A-Train, and stored in the A-Train Data Depot.