Data for eternity

Unlike accountants, scientists need to store their data forever. This expanding task requires dedication, expertise and substantial funds.

Data are at the heart of scientific research. Therefore, all data and metadata should be stored — forever, and accessibly. But it would be naïve to think that such a 'gold standard' of preservation could be achieved. In one spectacular example of the failure of science to save its treasures, some of NASA's early satellite data were erased from the high-resolution master tapes in the 1980s (*Science* **327,** 1322–1323; 2010). The lost data could now help extend truly global climate observations back to the 1960s — had they not been taped over.

At the time, the storage capacity of the tapes seemed more valuable than the data they contained. The story involved the preservation of analogue tapes with whale oil and the need for tape players the size of a large fridge. It vividly illustrates just how much the technology of data storage has changed since the 1960s.

But even in the past 20 years, standards of data documentation and preservation have been revolutionized. This has been largely ignored in the attacks on Phil Jones of the University of East Anglia's Climatic Research Unit over the loss of metadata regarding Chinese station locations used in his 1990 study.

When Jones and colleagues assessed what influence the 'urban heat island' effect had on the global warming signal (*Nature* **347,** 169–172; 1990), *Nature* was publishing hardly any colour figures and content was not fully available online. More importantly, the option of adding supplementary information to a paper — in hardcopy — had only just been introduced as "a scheme for assisting with the publication of data that would otherwise be buried in people's desk drawers" (*Nature* **346,** 215; 1990). At the time, the long-term vision of *Nature* was clear, but distant: "Eventually, of course, supplementary information will be distributed electronically, through an electronic database. But that is light-years away."

Until the introduction of full-scale supplementary information, ensuring that accessible records were kept was down to the authors. Of course, the loss of important information, such as the exact station locations used in the Jones *et al.* paper, is unacceptable (as Phil Jones himself put it; *Nature* **463**, 860; 2010) from a scientific point of view. But it is hardly surprising and probably widespread: scientists are not well-placed to guarantee continuity of data storage, especially while they are still in their vagabond years of PhD and post-doc work.

Nature Geoscience requires that authors make their data available on publication. The easiest way of ensuring that all the relevant information is accessible, and will remain so in the long term, is to use professionally run databases, which are now available for all sorts of Earth science data.

The creative push in science will always be for the production of better-resolved, more complicated data sets. Ingenious ways of storing and releasing these data are invariably developed with considerable lag. But this is not an excuse to neglect the issue. The preservation of valuable data sets and their distribution on demand is of utmost importance for the progress of science. The continuous attention of dedicated professionals — and substantial funds — is needed for database development to keep up with the science. ❐

Publishing ambiguity

Online publishing has blurred the boundary between accepted and published articles.

The world of science has accelerated. With the advent of online publication over the past 10 years, it no longer needs to take months or years for an accepted paper to become available to journal subscribers, and the term 'monthly journal' is losing its meaning. Articles are published online weeks to months before publication in print, with benefits all round: authors can make their peer-reviewed results available to the scientific community quickly, readers can keep abreast of the latest developments and publishers can provide a continuous stream of content in an increasingly competitive market.

But the downside of early online publishing is a confusing array of publicly available article types, awaiting print publication in various stages of editorial preparation. Adding to the confusion, interactive journals such as *Atmospheric Chemistry and Physics* place papers online first for peer review, and then in their final form. As the focus of scientific journals is moving from print to electronic publication, each publisher makes their own decision regarding the balance of speed versus the completeness of published work. But when papers go online before they are in final form, uncertainty arises regarding the canonical publication date.

Publisher's policies regarding the accessibility of online articles are equally piecemeal. *Science Express* — where *Science* papers are posted online up to six weeks ahead of publication in print — is available to site licence subscribers only as a premium add-on. And when journals of the American Geophysical Union publish 'in press' papers before their print version, only the titles of these papers are available to non-subscribers. On publication in print, abstracts are also free to access.

Nature Geoscience papers are published online in their final, definitive form — fully proofread and formatted — and the date of online publication is the date of record. However, we consider papers elsewhere as published as soon as the scientific content is fully available online, with a Digital Object Identifier (DOI name). That is, we are happy to highlight 'in press' articles, whatever format they are in. We also count them as part of the body of existing literature when assessing the advance of a submitted paper over existing knowledge.

Given the way the publishing industry is moving, it seems unlikely that print publication dates will play a role in the long run. And as the demand for print subscriptions wanes, unified payment models for accessing papers online and in print are likely to evolve. What needs to be decided is how much a preliminary paper published online should be allowed to change before it constitutes a new paper. ❐