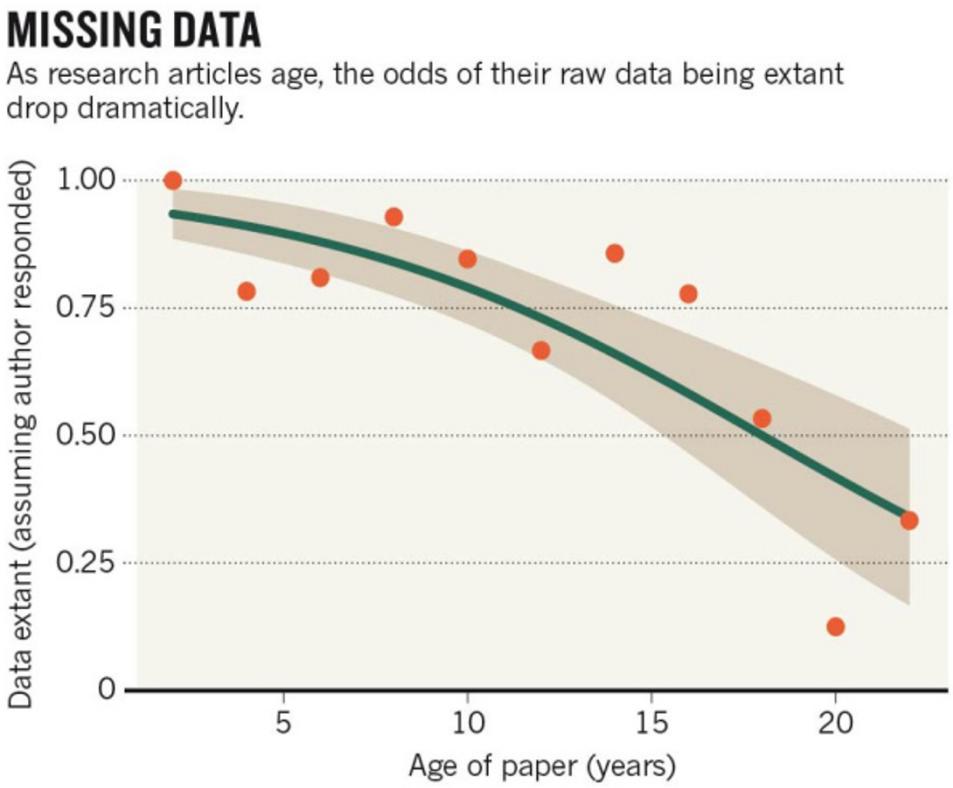
Scientists losing data at a rapid rate

Decline can mean 80% of data are unavailable after 20 years.

Elizabeth Gibney & **Richard Van Noorden**

19 December 2013



In their parents' attic, in boxes in the garage, or stored on now-defunct floppy disks — these are just some of the inaccessible places in which scientists have admitted to keeping their old research data. Such practices mean that data are being lost to science at a rapid rate, a study has now found.

The authors of the study, which is published today in *Current Biology*[1], looked for the data behind 516 ecology papers published between 1991 and 2011. The researchers selected studies that involved measuring characteristics associated with the size and form of plants and animals, something that has been done in the same way for decades. By contacting the authors of the papers, they found that, whereas data for almost all studies published just two years ago were still accessible, the chance of them being so fell by 17% per year. Availability dropped to as little as 20% for research from the early 1990s.

"Most of the time, researchers said 'it's probably in this or that location', such as their parents' attic, or on a zip drive for which they haven't seen the hardware in 15 years," says Timothy Vines, the lead author on the study and an evolutionary ecologist at the University of British Columbia in Vancouver. "In theory, the data still exist, but the time and effort required by the researcher to get them to you is prohibitive."

Another challenge was simply tracking down authors and receiving a response, something at which the team was successful in just 37% of cases. The likelihood of being able to find a working e-mail address, even after an extensive online search, declined by 7% per year. Meanwhile, only around half of the authors with valid addresses responded to the requests, however old the paper.

A role for journals

Matthew Woollard, director of the UK Data Archive in Colchester, cautions that the analysis does not take into account the size of the individual data sets, nor whether the data were held by institutions. "In the late 1990s or even early 2000s, much larger data sets would be more unlikely to end up in personal collections and so, possibly, have a higher chance of being kept institutionally," he says.

But overall, Woollard says, the results are broadly what he would expect across disciplines. The study's authors argue that journals are in the best position to do something about this. Demanding that authors submit their data to a public archive as a condition of publication could have a huge impact, says Vines, who is a managing editor of *Molecular Ecology,* a journal that introduced this policy two years ago. “It’s a very easy thing for journals to do, and I think it would dramatically improve the quality and quantity of data that are archived."

Nature requires authors to make data “promptly available to readers without undue qualifications” and to disclose restrictions upon submission. Some types of data, such as DNA sequences, must be submitted to a community-endorsed public repository. For other kinds of data, where public repositories are less developed, this is “strongly recommended”.

Although discipline-specific archives are making it easier for scientists to preserve and share their data, they are currently used by only a small number of eager early-adopters, says Michael Hildreth, a physicist at the University of Notre Dame in Indiana and leader of the US-government-funded Data and Software Preservation for Open Science. But as the tools to explore them and link data together are developed, they could become powerful ways to both organize and preserve data, he adds.

However, a survey presented at the International Congress on Peer Review and Biomedical Publication in Chicago, Illinois, in September found that researchers might be becoming more reluctant, not less, to share their data — at least in medical research. A survey of authors publishing in the *Annals of Internal Medicine* between 2008 and 2012 found that their willingness to share their data decreased from 62% to 47% over the period.

Irretrievable

Despite Vines' concern that allowing valuable data to disappear is crazy, tales of research data being lost to history are all too common. Agricultural researcher Melvin McCarty, for instance, spent 15 years between 1958 and 1973 recording the life cycles of plants and grasses near Lincoln, Nebraska. Forty years later, ecologist Lizzie Wolkovich went searching for McCarty’s data as part of an effort to tie together experiments exploring how rising temperatures affect plant life cycles. But McCarty had died, and his raw data could not be found. “There is nothing we can replicate now. The loss of the long-term data set is very sad,” says Wolkovich, who works at the University of British Columbia in Vancouver.

A similar fate befell the raw data collected in the 1980s by Otto Solbrig, a biologist at Harvard University in Cambridge, Massachusetts, on species of violets in New England. Plant biologist Sydne Record at Michigan State University in East Lansing wrote to him in 2009 asking for the original data, to test out a mathematical analysis of population viability that she was developing — but Solbrig didn’t have them. “We had at least 20 big folders with those data, but nobody was interested in them so we threw them away,” he says.

Nature | doi:10.1038/nature.2013.14416

References

1. Vines, T. H. *et al*. *Curr. Biol.* http://dx.doi.org/10.1016/j.cub.2013.11.014 (2013).

Nature ISSN 0028-0836 EISSN 1476-4687

SPRINGER NATURE