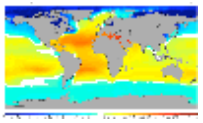
TALKING POINT

Feb 23, 2007

Rescuing data from obscurity

All over the world, data gathered by environmental researchers is gathering dust on 8 inch computer tapes, on 5 inch floppies and in decaying notebooks forgotten at the back of never-again-to-be-opened cabinets. These data, most often paid for by taxpayers, didn't necessarily make it into the publication, was perhaps thought uninteresting or was simply left behind when the investigator moved to a better-paid job in finance.


Isotope ratios

In my experience there are vast treasure stores of data sitting around laboratories that haven't ever been collated because it just didn't seem important, or it was something that one might eventually "get round to doing".

Environmental research today is complex and deals with a vast number of intertwined problems. It's so intertwined that data collected for one purpose may end up playing a key role in some quite unrelated field. Add to that the need for global data sets to compare with global models or remote-sensing information and it's clear that rescuing these individual pieces of information is more important than ever.

As an example, over the last few years some colleagues and I created a global database of seawater oxygen isotope ratios (i.e. the ratio of ^{18}O to ^{16}O in the water molecules). These are used as indicators of water masses and freshwater sources but also influence paleoclimate records derived from carbonates found in ocean sediments. I became interested in this data because as part of my research in palaeoclimate modelling I had included water isotopes in an ocean model and wanted to be able to compare my modelling with observations.

I was aware that some large-scale observational campaigns had included this tracer (notably GEOSECS in the 1980s), and I simply assumed that someone would subsequently have made a gridded data set. I was, however, mistaken. So I began collating observations from the literature in my spare time. I quickly found that the scope of the observational record was much larger than I'd realized – there are more than 100 references – and that other researchers were working along similar lines. We also found that we were collating a great deal of "grey" data – that is, data that might have been associated with a project or publication but that had been recorded separately or was simply unpublished.

Quality matters

Quality control was, of course, vital. We quickly became aware of significant inconsistencies in much of the data. What was nominally the same data from the same investigator would have different transcription errors (typos, missing numbers, sign errors, etc) depending on who we'd got it from. Separate measurements of the same deep water mass, which should be relatively stable, would be wildly out – by much more than the analytical error. Many of these problems were fixed by tracking down the original source, or getting an up-to-date calibration from the laboratory involved, but some remain a mystery.

Nonetheless, the database grew substantially and, as geographical regions were filled in, anomalous data stood out more clearly and could be flagged more easily. Eventually we ended up with a comprehensive data set that is becoming widely used. It has served to highlight many hard-to-find papers and data sets of wider relevance that would previously have been ignored.

Despite this success, there are still substantial amounts of data, both published and unpublished, that are not included in the database. It is worth exploring why that is the case in order to suggest ways in which it can be overcome.

The most important problem is the natural tendency for investigators to want to retain ownership of data and to be part of any analysis. To be clear, I'm not speaking about data sets that have just been gathered and that researchers haven't yet worked on – it is obviously correct that the data gatherers need to have a reasonable time window within which to assimilate their data – but rather I'm referring to the older work that is no

longer at the forefront of their research.

This relates to the need of the original investigators to see themselves as researchers rather than "mere" operational data gatherers. While understandable, this shift is part of an inevitable transition as a scientific field matures. The first few measurements of a quantity are important science, the next few provide vital evaluation, but after a few tens of thousands, the important questions often relate to spatial and temporal variability and are difficult for a single researcher in isolation to address.

An ancillary concern is the worry that the researcher will not be be properly credited. It does take time and effort to collate data and ensure that metadata are appropriate, for which, sadly, there is little recognition. A number of ideas have been proposed. For instance, a data centre could issue a digital object identifier for each data set so that it could then be referenced directly in articles and reports. No solution is yet available, however.

The second significant problem is simply the time factor. If data are not archived in a usable format with appropriate metadata at the time of publication or when the analyses are done, it can be a significant chore to do it a year or more later. There is therefore a paradoxical situation where the more data there are, the less likely it is that they will be submitted to an archive. This is where more institutionalized help may be useful. For instance, research centres and funding agencies could specifically fund short-term data-rescue efforts.

Going out to collect new data is expensive and duplicating previous efforts is rarely worthwhile scientifically. Getting a new decade-long time series will take, well, a decade. Despite the problems mentioned above, data rescue is cheap and effective and does not involve foreign travel – something that should commend itself to programme managers, if not the investigators themselves.

About the author

Gavin Schmidt is a climate researcher at the NASA Goddard Institute for Space Studies and a regular contributor to *RealClimate*. He would be very appreciative of any more data on seawater $\delta^{18}O$ that anyone has lying around.

metadata, culture, and saving work

Excellent article. Even when the original providers intend to provide reusable data, the technological solutions that are at hand often fail to meet the need of later researchers, data managers, and educators. An entire project (shameless plug: the Marine Metadata Interoperability project at marinemetadata.org) has been funded, by NSF and others, to improve the use of metadata in oceanographic science -- much of our work applies throughout the environmental science realm, and we are seeing increasing acceptance of its importance.

You were very delicate in your analysis of the causes of data hoarding -- it seems to me the cultural training inherent to becoming a successful scientist deserves a mention. Changing the scientific reward systems to favor collaboration, data citation/publication/reuse, and data stewardship will have a great effect on the value attached to existing data sets.

Technical solutions *do* exist to support many of these goals -- sometimes a few too many solutions -- but the field is still relatively underfunded. As the priorities change and technical progress continues, a revolution in data and metadata management will follow, probably sooner than most expect.