Thinking about data management

Think back to the beginning of modern oceanography. Investigators went out on ships to ask new questions, they tested new instruments, and they came home with the first-ever measurements of oceanic processes. After a bit of rumination, data quality control, and so forth, they published papers on their results. And then what happened to the data?

For the first oceanographers, getting the papers published may have seemed like the end goal of the observations. In any case, data sharing was difficult, with no mass storage devices and no internet. The final sharable data might have ended up in a series of tables in a technical report, and that seemed perfect.

Fast-forward in time to the present. Now we often ask complicated research questions that extend beyond understanding an oceanographic process over a brief window of time. We would like to understand how oceanic processes change over seasons, years, or decades, and we'd like to compile data spanning large geographic regions. All of this requires retrospective study of data collected by multiple investigators. Suddenly (or perhaps not so suddenly) there is research value in all of the old data. But old data might be stuck on an 8-track tape or a floppy disk that is no longer readable. New research questions might also require returning to the original data—the raw voltages measured by an instrument for example: as an example, retrospective studies of turbulent mixing have sometimes reprocessed raw CTD data to remove ship roll (e.g. Gargett and Garner, 2008).

Under current guidelines, how is data management supposed to work? For NSF-funded projects, all data are supposed to become public two years after collection (and some data types are available much faster than that). Investigators sometimes drag their feet a tiny bit, as they try to wrap up final data quality control and finish their own papers. The primary way to make data public is to send them to the National Ocean Data Center (NODC). NODC produces the World Ocean Data Base (and the World Ocean Atlas), and they have a well-established procedure for archiving hydrographic data. Data can also go to specialty data archives. For example, CLIVAR (and now GO-SHIP) repeat hydrography is handled by the CLIVAR & Carbon Hydrographic Data Office (CCHDO), which is located at Scripps. Our colleagues at CCHDO do a terrific job chasing down missing data and converting PI-provided files to consistent format.

But what about data that don't look like hydrographic profiles? Some data, types have well-established archives (for example, underway meteorological data and shipboard Acoustic Doppler Current Profiler (ADCP) measurements). For others, such as microstructure data and lowered ADCP data, the oceanographic community has needed time to agree on data formats, and data archiving is spottier. All of that is gradually getting better, with substantial prodding from the program managers; even with microstructure data, there has been a tremendous effort to rescue old data, agree on a common data formats, and make the collected data public (e.g. Waterhouse et al., 2014).

For a model of how to think about data management, look to the space agencies. Satellites generate gigabytes of data fairly quickly, and before the modern high-speed internet, in the 1980s and 90s, NASA spent a lot of time thinking about how to make data accessible. NASA defines a hierarchy of data levels:

Level 0. Engineering variables without physical meaning. These are curated, but realistically nobody wants them.

Level 1. Raw data with geophysical units. These are also rarely useful for users, but might be essential for anyone with a radically new idea that would require reprocessing from scratch.

Level 2. Data with geophysical corrections applied. For example, in the case of satellite altimetry, this would represent the sea surface heights measured along a satellite track. The data are useful for people who are thinking about satellite sampling problems, but can be confusing for other users.

Level 3. Value-added products. (These are often gridded but are based on a single satellite.) For sea surface temperatures (SSTs), for example, a level 3 product might be released at one-day intervals with data on a regular grid and blanks in places where no satellite measurements were collected that day.

Level 4. Multi-satellite gridded products represent a consistent mapping of data, combining measurements from multiple satellites to make a clean product with well-characterized uncertainties. For example, in the case of altimetry, the AVISO fields are considered a level 4 product, and in the case of SST, GHRSST products or NOAA's Optimum Interpolation SST would be level 4. Level 4 products are popular with users, because they are easy to use and they facilitate first-look investigations and cross-comparison with other data types.

Although Level 4 products are great for preliminary investigations, many decisions go into producing Level 4 products, and careful investigators may find that Level 4 products suppress aspects of a signal that they would like to study, which can send people searching for Level 2 or Level 3 data. Archives for in situ data have not adopted a clearly defined categorization of data levels, but the framework is useful. Think of CTD data. For most purposes, most of us are happy to look at atlas data (a level 4 product), sometimes we want to confine ourselves to data from a single cruise (level 3?), or single profiles that have been cleaned up and quality controlled (level 2). In rare cases, we need to reprocess the CTD data from the raw instrument files (level 0 or level 1) in order to extract some aspect of the data that is normally overlooked.

If you were setting up a data archive you'd want to make level 4 data be easy to find, and you'd want to provide direct access to level 2 or level 3 data. But ideally, somewhere in the bowels of your data system, you'd store a string of raw level 0 or level 1 data files to allow future expert users to explore breakthrough science questions that might extend far beyond what anyone has imagined today. The costs of a hierarchical data system are mostly disk space, and that is ever cheaper.

References

Gargett, A. and T. Garner, 2008: Determining Thorpe scales from ship-lowered CTD density profiles. *J. Atmos. Oceanic Technol.*, **25**, 16571670.

Waterhouse, A. F., et al., 2014: Global patterns of diapycnal mixing from measurements of the turbulent dissipation rate. *J. Phys. Oceanogr.*, **44**, 1854–1872.