Managing the UAF data collections: Lessons Learned

Roland Schweitzer (OAR PMEL/SDIG), Ansley Manke (NOAA/PMEL), Karl Smith (JASIO), Kevin O’Brien (JASIO), Eugene Burger (NOAA/PMEL)

The Unified Access Framework is an effort to provide real-world data access to a NOAA-wide collection of quality data sets via THREDDS catalogs, ERDDAP and Live Access Server (LAS). As conceived, the quality of a candidate data set for the collection would be judged on whether the data are aggregated (at least along the time dimension), whether the data access points follow the Climate and Forecast (CF) NetCDF conventions and if the THREDDS servers enable all of the appropriate services.

Since many THREDDS data catalogs and servers are disorganized, not properly aggregated and contain some individual data access points that do not follow the appropriate conventions, the project developed software to produce a “clean” catalog representation of the candidate catalogs by removing unaggregated data and data that fails the conventions check as well as enabling remote versions of any missing services.

We expected that the catalogs would improve over time and the burden of “cleaning” them would be reduced. In fact, it has gotten larger. As currently implemented, the “catalog cleaner” is examining thousands of individual data granules to produce the clean catalog and those exact same data sources are also being examined by the project’s ERDDAP server which serves as the primary access point for discrete geometry data and a secondary way of accessing the gridded data collection.

This is a massive duplication of effort and a burden on the resources of the NOAA centers who are providing the infrastructure to maintain this collection. Work is now underway to streamline this architecture and eliminate the duplicated effort in light of the fact that more catalogs are joining and the quality of the member catalogs is not improving substantially over time. The current cleaning process also produces a rubric displayed via a web page which grades the catalog on the criteria described above.

In the new architecture the cleaning process will eliminate un-aggregated data collections using heuristics while examining the catalogs, it will add missing services as before and provide links to a web service that will produce the clean catalog. Any sub-catalogs in the collection known to contain low quality data will be added to configuration of the cleaner that will skip those catalogs. When adding new catalogs, the Catalog Evaluator will be used to do an initial assessment of the catalog and allow the worst sub-catalogs to be identified and added to the configuration used to skip catalogs. This will allow a new “clean catalog” to be generated in a matter of minutes, rather than the days required to examine the entire catalog space currently. The heavy lifting of examining the contents of each data source in the candidate catalogs and assessing whether it follows the convention will be shifted to the projects ERDDAP server. If after all there are still data sources with issues found by ERDDAP those can be marked for elimination. Finally, since the burden of examining the individual data sources and been taken up by ERDDAP, the configuration for the UAF Live Access Server can be generated directly from the ERDDAP server an order of magnitude more quickly.

Using Google Sheets and App Scripts to Generate netCDF File Headers

Meg Tilton (NESDIS CIRES/NOAA NCEI)

NOAA’s National Centers for Environmental Information (NCEI) is developing the high-level, derived products for space weather data from the GOES-R satellite series. NOAA’s Space Weather Prediction Center (SWPC) and other customers will use these derived products to develop forecasts and models. These NetCDF-4 products use inputs from four different instruments and have been created by multiple scientists and developers. To ensure that these products’ header metadata are consistent, complete, and compliant with ACDD standards, we created templates for them using an automated Google Sheet. This spreadsheet stores header metadata and uses a Google Apps script (similar to JavaScript) to create Common Data Language (CDL) files, which can quickly be turned into master NetCDF files. The spreadsheet allows for easy cross-product comparison of NetCDF dimensions, variables, and attributes.

This presentation focuses on the development of NetCDF headers but also demonstrates the widespread utility of Google Apps scripts when working with shared documents.

NCEI NetCDF templates - A Walkthrough

Ajay Krishnan (NESDIS Science & Technology Corp/ NCEI)
The NOAA National Centers for Environmental Information (NCEI) have developed NetCDF templates based on what are called "feature types" by Unidata and CF. These templates conform to Attribute Convention for Dataset Discovery (ACDD) and NetCDF Climate and Forecast (CF) conventions. Adding to these established conventions, NCEI also provides several recommendations for both NetCDF variables and attributes. These best practices capture NCEI's experience in providing long-term preservation, scientific quality control, product development, and multiple data re-use beyond its original intent. In conjunction with the templates, NCEI has provided 'gold standard' example files which follow the templates and a collection of reports which document the results of testing the files in various compliance checkers. This presentation will review the NCEI templates and provide information regarding the gold standard examples and subsequent reports.

**7A.04 An Introduction to ERDDAP**

*Bob Simons (NMFS SWFSC ERD)*

This talk will give an introduction and overview of ERDDAP for people considering using or installing ERDDAP, and describe some of the features added to ERDDAP in the last year.

**7A.05 Efficient data archiving of CO-OPS’ real-time water level data at hundreds of locations using ERDDAP**

*Armin Pruessner (NOS CO-OPS)  Peter Stone (CO-OPS), Shawn Maddock (CO-OPS), Val Soika (CO-OPS)*

The Center for Operational Oceanographic Products and Services (CO-OPS) collects, quality-controls and disseminates real-time water level data at hundreds of locations along US coastal water ways and the Great Lakes. In addition, CO-OPS processes these data monthly to create verified data products. This talk will highlight how CO-OPS is using ERDDAP to efficiently create archive files of real-time data and verified data products. We will discuss some of the challenges when dealing with large volumes of time series data, how ERDDAP assists with archiving, and discuss the future use of ERDDAP for archiving all the other coastal observational data sets that CO-OPS collects.

**7A.06 ERDDAP and the International Ocean Community**

*Kevin O’Brien (OAR University of Washington/JISAO, NOAA/PMEL)*

Currently in the ocean observing community, there are a wide variety of platforms that provide measurements on a variety of timescales, both in real time and through delayed mode processes. These distinct data streams also tend to have disparate methods of managing and serving the data and information. As the ocean community evolves to support the concepts of Essential Ocean Variables (EOV), or Essential Climate Variables (ECV), it becomes critical to provide access to integrated collections of well documented observations, which can often originate from different platform observing networks.

NOAA’s Observing System Monitoring Center (OSMC), in partnership with the Unified Access Framework (UAF) project, is utilizing ERDDAP as the framework to improve integration of data and information across the global ocean observing networks. ERDDAP has been deployed to serve delayed-mode, quality controlled data from several observing networks, and efforts continue to include even more networks. In addition, ERDDAP is a critical component in a pilot project currently underway to improve the ways in which observations can be contributed to the real time ocean observing system.

In this presentation, we will provide an update on the projects involved in implementing ERDDAP as well as how ERDDAP use is growing within the international ocean community.