Managing a Community Data Collection with Open Source Software

Roland Schweitzer[1], Ethan Davis[2], Sean Arms[3], Robert Simons[4], Kevin O'Brien[5], and David Neufeld[6]

[1]Self Employed
[2]UCAR Unidata
[3]University Corporation for Atmospheric Research
[4]NOAA/NMFS/SWFSC
[5]NOAA/PMEL UW Joint Institute for the Study of the Atmosphere and Ocean
[6]CIRES

November 25, 2022

Abstract

The Unified Access Framework (UAF) project of the NOAA Global Earth Observation - Integrated Data Environment (GEO-IDE) in an on-going effort to provide access to NOAA-wide data in a way that is FAIR and meets PARR requirements. The first priority of UAF is to copy success. We recognize: data that follows the Climate and Forecast netCDF convention is readily used by working scientists; THREDDS Data Servers and ERDDAP servers are a popular ways to serve such data; these servers can be interrogated by software to determine that the data follows the conventions and the servers can be federated. To make the collection we construct a master "raw" catalog of candidate data set from THREDDS servers around NOAA and other organizations. The raw catalog is examined by custom software to eliminate large data collections which are not aggregated in time and organize the results into a "clean" catalog. The catalog is then examined by ERDDAP to provide ERDDAP GridDAP access and to verify that the data sources follow the CF convention. The gridded data sets are merged into a collection of TableDAP (netCDF Discrete Sampling Geometry) data sources. Currently the UAF ERDDAP server is home to 10,712 data sets. After the UAF ERDDAP server has examined the data collection, a Live Access Server (LAS) is configured to offer data analysis and visualization access to all the data sets. The final piece of the puzzle is to make the data FAIR and to achieve PARR compliance. This requires some tools that have been adapted and developed for this purpose. We resurrected the ncISO tool which can examine the contents of CF netCDF data sources and create ISO metadata and score the data according the the Unidata Attribute Convention for Data Discovery. We can help the centers hosting the data meet their PARR requirements by properly integrating the resulting metadata from ncISO into NOAA's central data catalog. We have recently updated the templates which are used to generate the metadata to insure they are meeting the latest ISO and ADDC specifications. Work is underway at NOAA and Unidata to integrate the ncISO code back into the GitHub repository for the THREDDS Data Server. This will bring together two disparate ncISO implementations. UAF is a few people working a few hours a month to maintain and large and useful data collection and in this talk we'll tell you how we do it.

PRESENTER: Roland Schweitzer

OUR GOAL

- Publish the largest possible collection of high-quality OPeNDAP accessible data sets.

METHODS

1. Create a catalog of THREDDS Data Servers from NOAA and selected partners.
2. Crawl the entire collection, writing a new catalog hierarchy eliminating any candidate catalogs containing un-aggregated time series or non-OPeNDAP data sources.
3. Verify that each of the resulting data sources contains a CF-Compliant netCDF data set.
4. Collect Discrete Sampling Geometry datasets into a central ERDDAP server.
5. Publish the results in a THREDDS Data Server, a Live Access Server and an ERDDAP server (with both the grid and discrete sampling data sets).
6. Collect and publish ISO-19115 metadata for each data set into noaa.data.gov, data.gov and NOAA One-Stop.

RESULTS

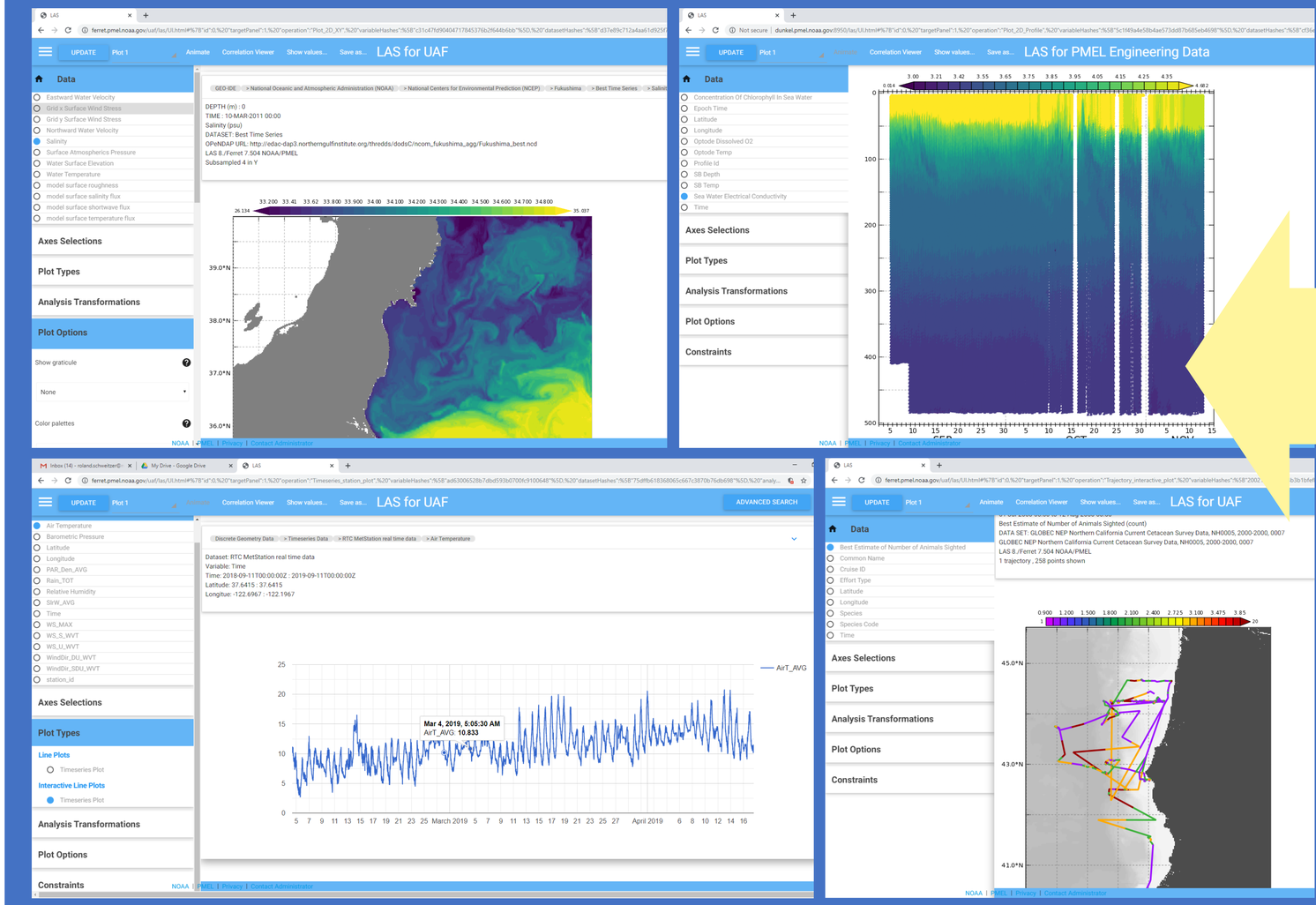
- 12420 data sets available from your favorite server using your favorite scientific analysis and visualization software.
- Searchable and discoverable via the standard NOAA and federal data repositories.

DISCUSSION

- With the right approach (copying existing successful standards and data services) and the right software (big data servers like THREDDS OPeNDAP, ERDDAP and data catalogs) and a little software glue (the UAF Catalog Cleaner and ncISO) a few people can produce and maintain a large high-quality data service filled with useful and discoverable data sets.

About LAS

Find out how a few nerds manage a data collection of 12,000+ data sets with open-source software, ingenuity and a bit of custom software glue.



AGU100 ADVANCING EARTH AND SPACE SCIENCE

Managing a Community Data Collection with Open Source Software
IN33B-0826

Roland Schweitzer, Ethan Davis, Sean Cody Arms, Robert Simons, Kevin O'Brien, David Neufeld

THE FOSS and the GLUE

FOSS:

THREDDS Data Server
https://github.com/Unidata/thredds

ERDDAP
https://github.com/BobSimons/erddap

Live Access Server (v8)
https://github.com/NOAA-PMEL/LAS

Live Access Server (v9)
https://github.com/NOAA-PMEL/las9
(coming soon)

GLUE:

UAF Catalog Cleaner
https://github.com/NOAA-PMEL/FastClean

ncISO
https://github.com/NOAA-PMEL/uafnciso

Try LAS 9
(not optimized for mobile)

NOAA National Oceanic and Atmospheric Administration U.S. Department of Commerce

PMEL Pacific Marine Environmental Laboratory