#### Managing a Community Data Collection with Open Source Software

Roland Schweitzer<sup>1</sup>, Ethan Davis<sup>2</sup>, Sean Arms<sup>3</sup>, Robert Simons<sup>4</sup>, Kevin O'Brien<sup>5</sup>, and David Neufeld<sup>6</sup>

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

## **OUR GOAL**

 Publish the largest possible collection of highquality 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.



# 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 gue.

UPDATE Plot 1	Animate Correlation Viewer Show values Save as LAS for UAF
🕈 Data	GEO-IDE > National Oceanic and Atmospheric Administration (NOAA) > National Centers for Environmental Prediction (NCEP) > Fukushir
O Eastward Water Velocity	
O Grid x Surface Wind Stress	DEPTH (m) : 0 TIME : 10-MAR-2011 00:00
Grid y Surface Wind Stress	Salinity (psu)
Salinity	DATASET: Best Time Series
O Surface Atmospherics Pressure	LAS 8./Ferret 7.504 NOAA/PMEL
O Water Surface Elevation	Subsampled 4 in Y
O Water Temperature	
model surface roughness     model surface collisity flux	
O model surface shortwave flux	33.200 33.41 33.62 33.800 33.900 34.00 34.100 34.200 34.300 34.400 34.500 34.600 34.700 34.800
O model surface temperature flux	
Axes Selections	
Plot Types	39.0°N
Analysis Transformations	38.0°N
Plot Options	
Show graticule	37.0°N
None	
Color palettes	
Mukay (14) - soland schweitzer@s - Muk	36.0°N       NOAA   PMEL   Privacy   Contact Administrator
M Inbox (14) - roland.schweitzer@n ×       ▲ My D         ← → C       ① ferret.pmel.noaa.gov/uaf/la         UPDATE       Plot 1         ▲ Air Temperature       Plot 1         ▲ Air Temperature       Barometric Pressure         ▲ Latitude       Longitude         ▲ Latitude       Air Temperature         ▲ Rain_TOT       Relative Humidity         ▲ Rain_TOT       Relative Humidity         ▲ SIRW_AVG       Time         ▲ WS_MAX       WS_S_WVT         ▲ WindDir_DU_WVT       WindDir_SDU_WVT         ▲ Selections       Plot Types	So.PN NOAA (* MEL   Privacy   Contact Administrator Sine - Google Drive x LAS x + as/ULhtml#%7B*id*:0.%20*targetPanel*:1.%20*operation*:*Timeseries_station_plot*:%20*variableHashes*:%5B*ad63006528b7dbd593b0700fc9100648*%5D.%20 Annual Correlation Viewer Show values Save as LAS for UAF Discrete Geometry Data > Timeseries Data > RTC MetStation real time data > Air Temperature Dataset: RTC MetStation real time data Variable: Time Time: 2018-0911100:00:002 : 2019-09-11100:00:00Z Latitude: 37.6415 : 37.6415 Longitue: -122.6967 : -122.1967
M Inbox (14) - roland.schweitzer@n × ▲ My D   ← → C ③ ferret.pmel.noaa.gov/uaf/la   □ UPDATE Plot 1     ● Air Temperature   ○ Barometric Pressure   ○ Latitude   ○ Longitude   ○ PAR_Den_AVG   ○ Rain_TOT   ○ Relative Humidity   ○ SIrW_AVG   ○ Time   ○ WS_MAX   ○ WS_S_WVT   ○ WindDir_DU_WVT   ○ WindDir_SDU_WVT   ○ station_id     Axes Selections     Plot Types   Line Plots   ○ Timeseries Plot	36.0*N NOAA   * PIEL   Privacy   Contact Administrator  Drive - Google Drive × LAS × +  sx/U.html#%76°id*0.%20°targetPanel*1.%20°operation**Timeseries_station_plot*%20°variableHashes*%58°ad63006528b7dbd593b07001c9100648°%50.%20  Note - Google Drive × O LAS × +  sx/U.html#%76°id*0.%20°targetPanel*1.%20°operation**Timeseries_station_plot*%20°variableHashes*%58°ad63006528b7dbd593b07001c9100648°%50.%20  Note - Google Drive × O LAS × +  sx/U.html#%76°id*0.%20°targetPanel*1.%20°operation**Timeseries_station_plot*%20°variableHashes*%58°ad63006528b7dbd593b07001c9100648°%50.%20  Note - Google Drive × O LAS × +  sx/U.html#%76°id*0.%20°targetPanel*1.%20°operation**Timeseries_station_plot*%20°variableHashes*%58°ad63006528b7dbd593b07001c9100648°%50.%20  Note - Google Drive × O LAS × +  sx/U.html#%76°id*0.%20°targetPanel*1.%20°operation**Timeseries_station_plot*%20°variableHashes*%58°ad63006528b7dbd593b07001c9100648°%50.%20  Note - Google Drive * Show values Save as LAS for UAF  Discrete Geometry Data * Timeseries Data * RTC MetiStation real time data * Air Temperature  Dataset: RTC MetiStation real time data Variable: Time: 2018-09-11T00:00:002: 2019-09-11T00:00:002 Latitude: 37.6415: 37.6415 Longitue: -122.6967: -122.1967
M Inbox (14) - roland.schweitzer@n × ▲ My D   ← → C ① ferret.pmel.noaa.gov/uaf/la   □ UPDATE Plot 1     ● Air Temperature   ○ Barometric Pressure   ○ Latitude   ○ Longitude   ○ PAR_Den_AVG   ○ Relative Humidity   ○ SIrW_AVG   ○ Time   ○ WS_MAX   ○ WS_S_WVT   ○ WindDir_DU_WVT   ○ WindDir_SDU_WVT   ○ station_id     Axes Selections     Plot Types   Line Plots   ○ Timeseries Plot	36.0*N NOAA I PIEL Privacy I Control Administrator bine - Google Drive X LAS x + sc/ULhtml#%76"id*0.9201*argetPanel*1.9201*operation**Timeseries_station_plot*.9201*anableHashes*%58"ad63006528b7dbd593b0700fc9100648*%50.9201 Administration Viewer Show values Save as LAS for UAF Discrete Geometry Data > Timeseries Data > RTC MetStation real time data > Air Temperature Dataset: RTC MetStation real time data Variable: Time Time: 2018-09-111700.000.002 : 2019-09-111700.00.002 Latitude: 37.6415 : 37.6415 Longitue: -122.6967 : -122.1967
M Inbox (14) - roland.schweitzer@n ×       ▲ My D            ← → C           ① ferret.pmel.noaa.gov/uaf/la             □ UPDATE Plot 1           Plot 1          ● Air Temperature          Plot 1          ● Air Temperature          Plot 1          ● Air Temperature          Plot 1          ● Barometric Pressure          Latitude         ● Latitude          Q         Longitude         ● PAR_Den_AVG          Relative Humidity         ● SIrW_AVG          Time         ○ WS_MAX          WS_S_WVT         ○ WS_U_WVT          WindDir_DU_WVT         ○ WindDir_DU_WVT         ○ WindDir_SDU_WVT         ○ station_id         Axes Selections         Plot Types         Line Plots         ○ Timeseries Plot         Interactive Line Plots         ● Timeseries Plot	Add + KEL Privacy I Contact Administrator NAA + KEL Privacy I Contact Administrator Write - George Drive X O LAS X + acyULINtmH%77B'rd*0.0%20*targetPanel*15.%20*operation*? Timeseries_station.plot*%20*variableHashes*%58*ad63005528b7dbd533b0700f;9100648*%5D.%20 Autmate Correlation Viewer Show Values Save as LAS for UAF Discrete Geometry Data > Timeseries Data > RTC MeiStation real time data > Alit Temperature Dataset: RTC MetStation real time data Variable: Time The: 2018-00-11100.000.02 : 2019-09-11100.00.002 Latitude: 37.6415 : 37.6415 Longitue: -122.6967 : -122.1967
M Inbox (14) - roland.schweitzer®n × ▲ My D   ← → C ③ ferret.pmel.noaa.gov/uaf/la   □ UPDATE   Plot 1     ● Air Temperature   ○ Barometric Pressure   ○ Latitude   ○ Longitude   ○ PAR_Den_AVG   ○ Rain_TOT   ○ Relative Humidity   ○ SIrW_AVG   ○ Time   ○ WS_MAX   ○ WS_LU_WVT   ○ WindDir_DU_WVT   ○ WindDir_SDU_WVT   ○ station_id     Axes Selections   Plot Types   Line Plots   ○ Timeseries Plot   Interactive Line Plots   ● Timeseries Plot	Accord For the provided and the series of th
M Inbox (14) - roland.schweitzer@r × ▲ My D <td< td=""><td>store Score Drive x 0 L3 x + nx/U.html#X7B*df-0.%20*LargetPanel*1.%20*operation*Timesenes_station_plot*%20*variableHashes*%58*ad63006528b7dbd593b0700td9100648*%59.%20 xr/U.html#X7B*df-0.%20*LargetPanel*1.%20*operation*Timesenes_station_plot*%20*variableHashes*%58*ad63006528b7dbd593b0700td9100648*%59.%20 xr/U.html#X7B*df-0.%20*LargetPanel*1.%20*operation*Timesenes_station_plot*%20*variableHashes*%58*ad63006528b7dbd593b0700td9100648*%59.%20 xr/U.html#X7B*df-0.%20*LargetPanel*1.%20*operation*Timesenes_station_plot*%20*variableHashes*%58*ad63006528b7dbd593b0700td9100648*%59.%20 Xr/U.html#X7B*df-0.%20*LargetPanel*1.%20*operation*Timesenes_station_plot*%20*variableHashes*%58*ad63006528b7dbd593b0700td9100648*%59.%20 Decrete Geometry Data * Timeseries Data * Att Temperature Dataset: RTC MetStation real time data Variable: Time Time: 2018.09-11700.00.002 : 2019-09-117100.00.002 LargetUie: -122.6967:-122.1967 25 25 25 25 25 25 25 25 25 25</td></td<>	store Score Drive x 0 L3 x + nx/U.html#X7B*df-0.%20*LargetPanel*1.%20*operation*Timesenes_station_plot*%20*variableHashes*%58*ad63006528b7dbd593b0700td9100648*%59.%20 xr/U.html#X7B*df-0.%20*LargetPanel*1.%20*operation*Timesenes_station_plot*%20*variableHashes*%58*ad63006528b7dbd593b0700td9100648*%59.%20 xr/U.html#X7B*df-0.%20*LargetPanel*1.%20*operation*Timesenes_station_plot*%20*variableHashes*%58*ad63006528b7dbd593b0700td9100648*%59.%20 xr/U.html#X7B*df-0.%20*LargetPanel*1.%20*operation*Timesenes_station_plot*%20*variableHashes*%58*ad63006528b7dbd593b0700td9100648*%59.%20 Xr/U.html#X7B*df-0.%20*LargetPanel*1.%20*operation*Timesenes_station_plot*%20*variableHashes*%58*ad63006528b7dbd593b0700td9100648*%59.%20 Decrete Geometry Data * Timeseries Data * Att Temperature Dataset: RTC MetStation real time data Variable: Time Time: 2018.09-11700.00.002 : 2019-09-117100.00.002 LargetUie: -122.6967:-122.1967 25 25 25 25 25 25 25 25 25 25









NOAA

trajectory, 258 points shown





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)





