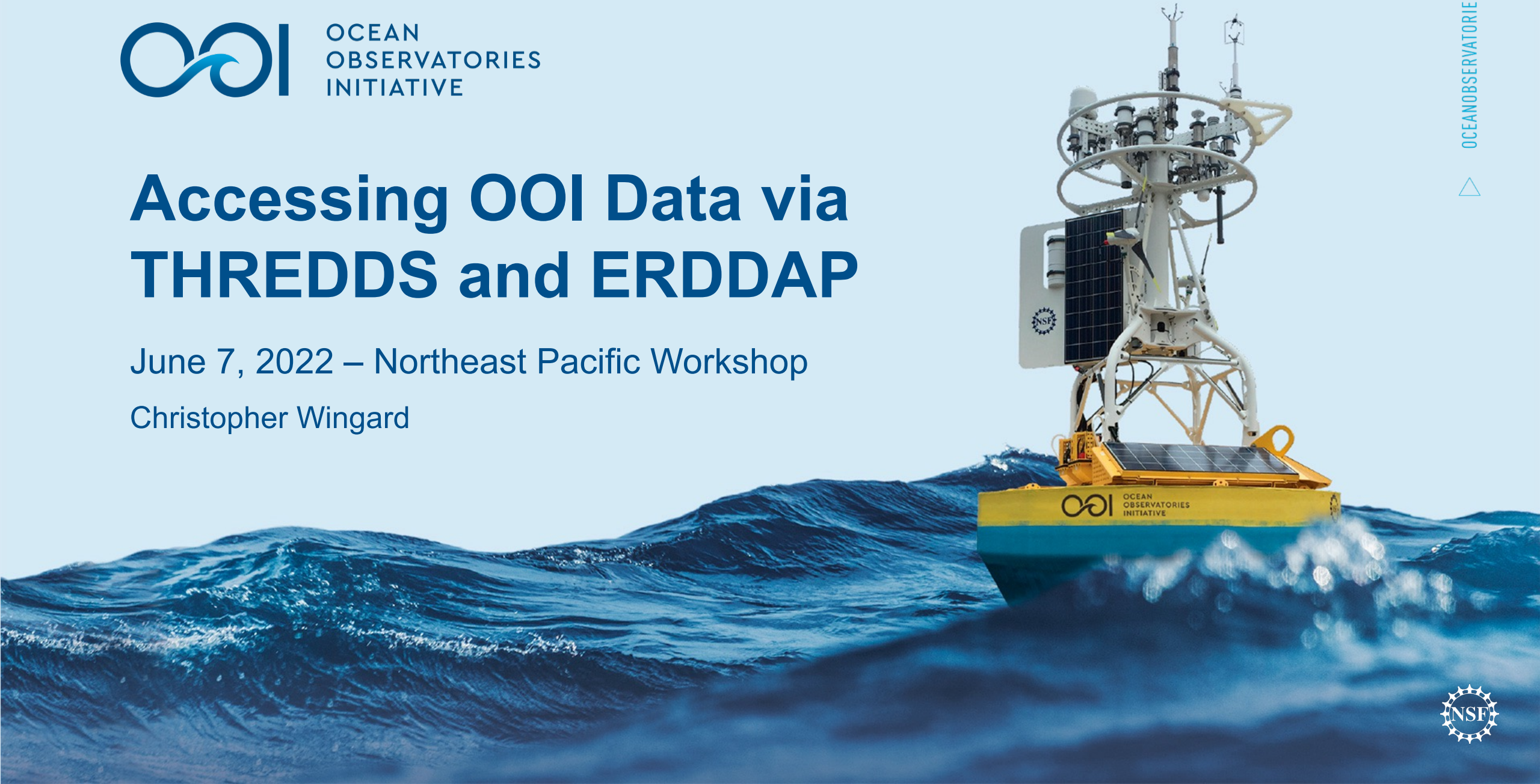
OCEAN OBSERVATORIES INITIATIVE

Accessing OOI Data via THREDDS and ERDDAP

June 7, 2022 – Northeast Pacific Workshop

Christopher Wingard



OCEANOBSERVATORIES.ORG

NSF

Housekeeping

- This is intended to be an interactive session. Questions? Please, do not hesitate to ask!
  - Raise your hand (real or virtual)
  - Post to the chat
  - Use the Breakout Session Worksheet
  - Visit the “Help Desk” during the Poster Sessions
- Introduction
  - OOI Endurance Array Data Lead (formerly Mooring Instrument Lead)
  - Been a member of OOI since Summer 2010
  - Spent some time as part of the UCSD CI Data Team
  - Helped develop, alongside Stuart Pearce, Russell Desiderio and Craig Risien, the python code used to calculate OOI data products
  - Actively work with OOI data (sensor deep-dives, HITL annotations, QARTOD test limits, etc.)

Terminology (Handout: Decoder)

Array

- One of 5 major research regions that make up OOI (e.g., Global Station Papa (**GP**) or Coastal Endurance (**CE**))

Site

- A specific geographic location within an array (e.g., Coastal Endurance Oregon Inshore = **CE01**)

Platform

- An assembly/infrastructure at a site that hosts a complement of integrated scientific instruments. May be fixed (mooring) or mobile (profilers or gliders) (e.g., Coastal Endurance Oregon Inshore Surface Mooring = **CE01ISSM**)

Node

- A section of a platform with one or more data loggers and power controllers. Instruments on platforms are plugged into nodes. (e.g., the data logger on the Near Surface Instrument Frame (NSIF, 7 m), part of the mooring riser, of the Coastal Endurance Oregon Inshore Surface Mooring = **CE01ISSM-RID16**)

Instrument/Sensor

- Terms often used interchangeably. An instrument is a piece of equipment used to collect data. A sensor is a part of an instrument which measures a specific quantity. Each instrument has a Unique ID (e.g., the dissolved oxygen sensor on the NSIF of the Coastal Endurance Oregon Inshore Surface Mooring = **03-DOSTAD000**

Reference Designator

- Identifies a particular instrument on a particular node/platform at a particular site
- Example: **CE01ISSM-RID16-03-DOSTAD000** = Coastal Endurance Oregon Inshore Surface Mooring - Mooring Riser - Dissolved Oxygen Sensor

Data Delivery Method

- **telemetered**: Data returned wirelessly. May be truncated or decimated due to size.
- **recovered_host/recovered_inst**: Data downloaded directly from either the platform computer and/or from the instrument after the system is recovered
- **streamed**: Data accessible in real-time, streamed over the fiber optic network from Cabled Array and select Endurance platforms

Data Stream Name

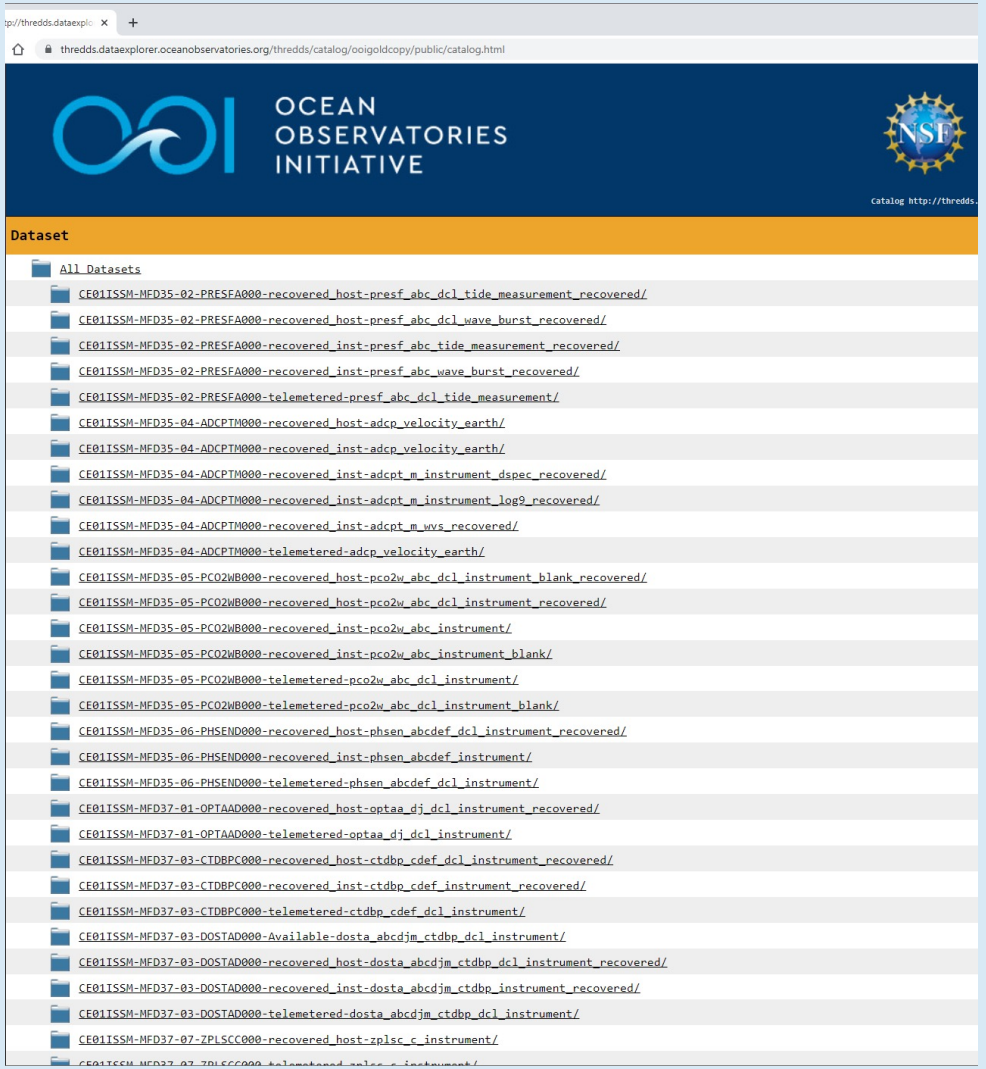
- Data feed from a sensor that has been read, parsed, and separated based on content (e.g., “engineering”, “science”, “metadata”, etc.) into a named dataset, or “stream”. Stream names are often method specific

Parameters

- Also sometimes called data variables
- A particular value returned from a sensor (e.g., practical salinity from a CTD). There are multiple parameters in a data stream, some of which may be identified as an OOI Data Product

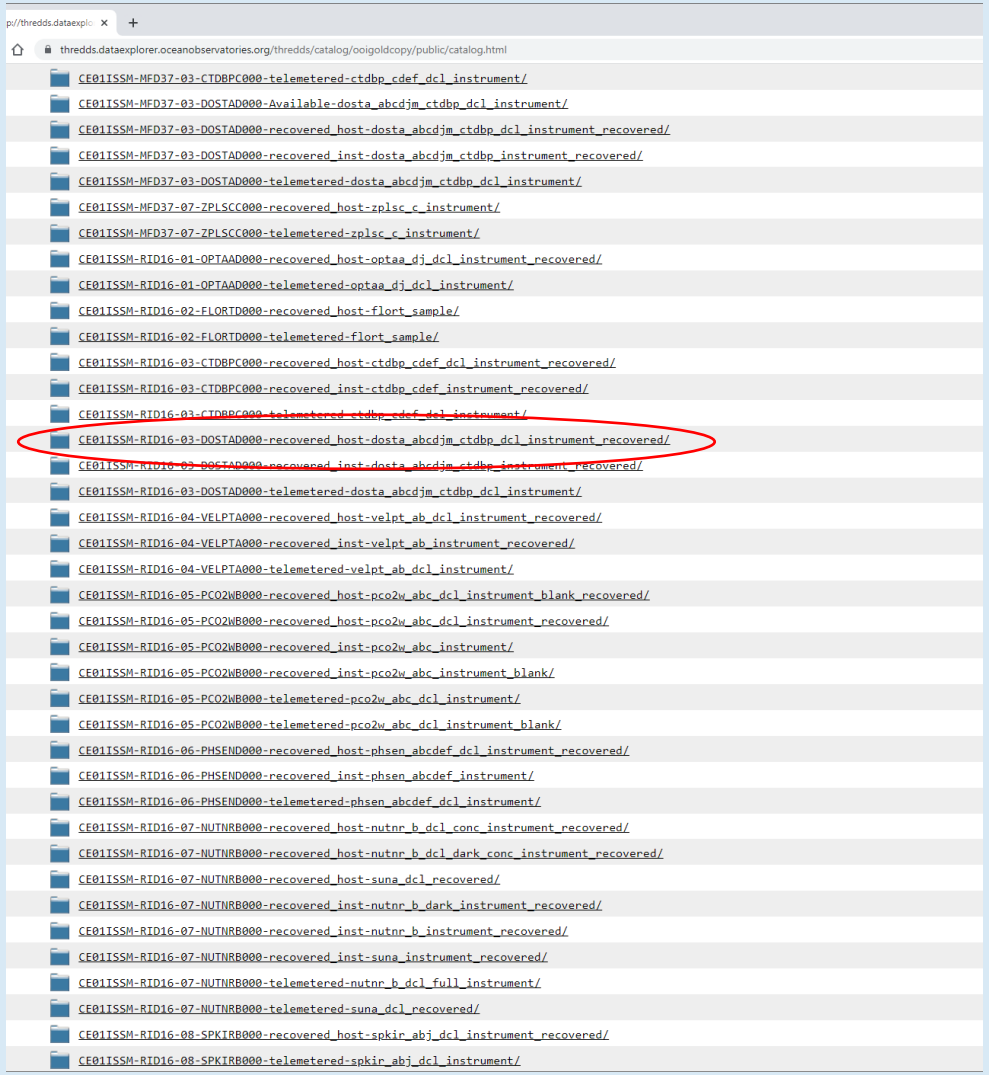
Thematic Real-Time Environmental Distributed Data Services (THREDDS)

- M2M Requests (async)
  - User and request specific
  - Temporal organization
  - Limited retention time (currently 6 months, under review)
- Gold Copy THREDDS
  - https://thredds.dataexplorer.oceanobservatories.org/thredds/catalog/ooigoldcopy/public/catalog.html
  - Created by internal OOI processes, updated daily
  - Organized by reference designator, delivery method and stream
  - Covers most data sets provided by OOI through the M2M system (exceptions are “engineering” or “metadata” streams)
  - Eliminates the need for most M2M requests
  - ***Basis for the data served via Data Explorer***



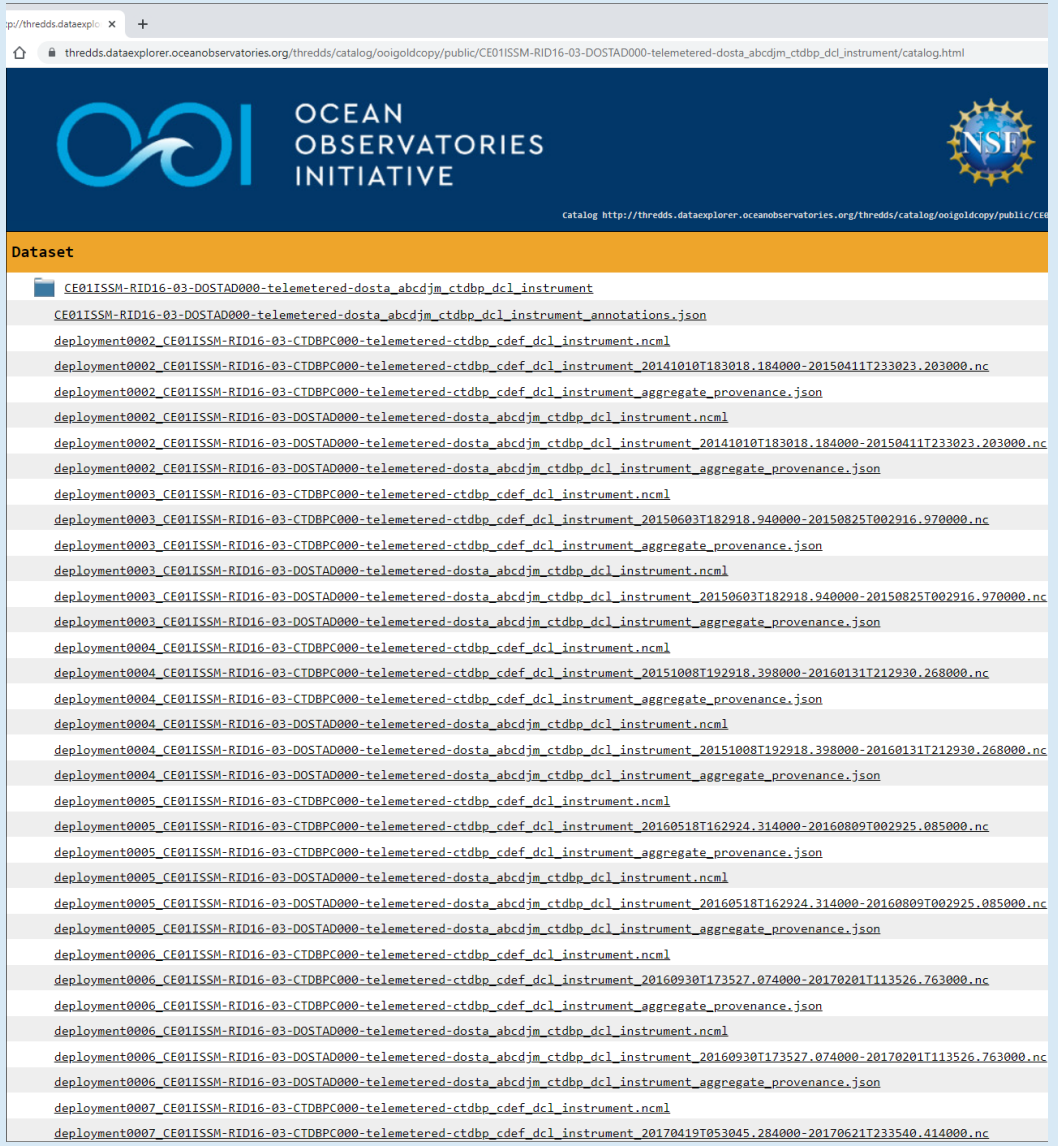
Gold Copy THREDDS: Finding Data

- Browser
  - Scroll/CTRL-F to find the telemetered data from the Coastal Endurance Oregon Inshore Surface Mooring DO sensor on the midwater platform (NSIF, 7m)
    - CE01ISSM-RID16-03-DOSTAD000-telemetered-dosta_abcdjm_ctdbp_dcl_instrument
- Code-based
  - Siphon, python software developed by UCAR/Unidata to access remote data servers (initially THREDDS, but expanding)
  - Web scraping (Beautiful Soup module for python, webread function in Matlab, or readLines in R)
  - URL for an instrument specific dataset can be constructed from the reference designator, data delivery method and stream name



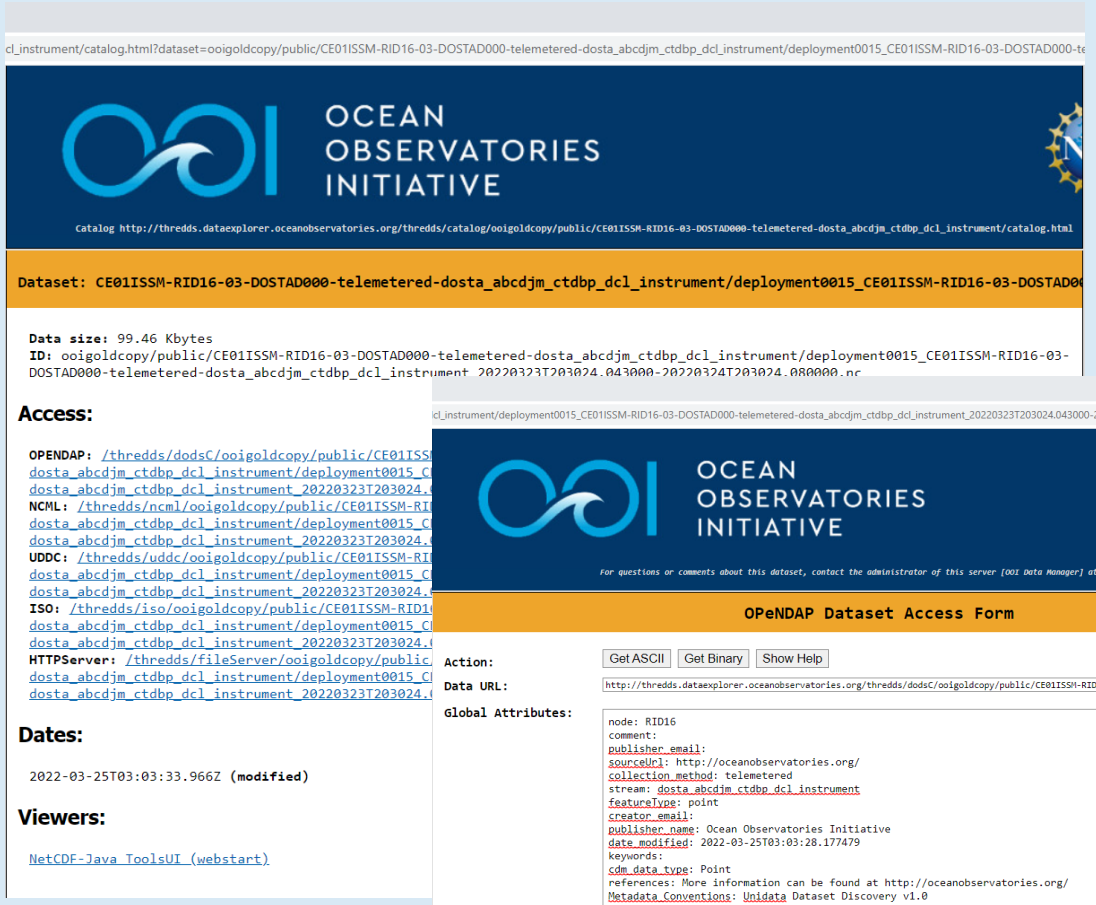
Gold Copy THREDDS: Datasets

- Files named by deployment number, RefDes, method, stream, file contents:
  - *Annotations: Should include all annotations associated with the reference designator but does not. Users will need to use the M2M system to obtain the full set of annotations*
  - *NCML: Bug in stream engine used to generate these files, don't use them (other issues as well)*
  - **NC: netCDF files time stamped based on time coverage in the file. These contain the instrument data**
  - Provenance: Firehose of metadata. Includes information about every deployment and every calibration coefficient for every instance of an instrument deployed at the site. In addition, contains information about the request, the parameters and files used in any processing as well as the functions called
- Co-located instrument(s) if required by processing



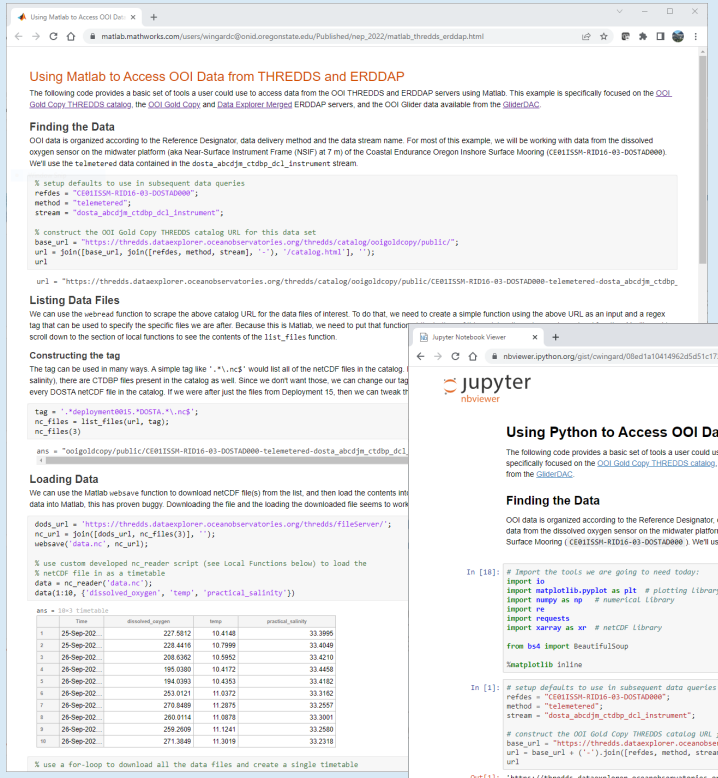
Gold Copy THREDDS: Downloading

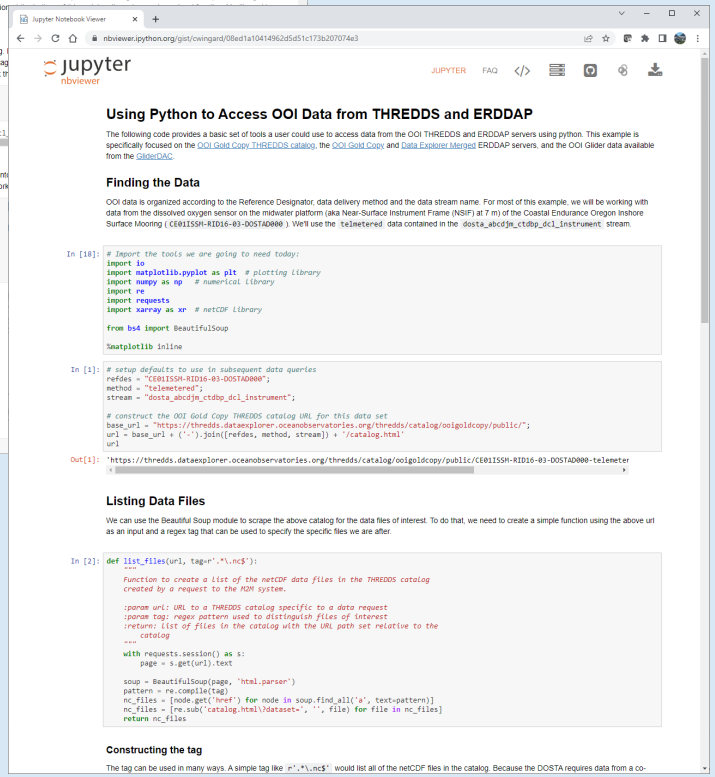
- Direct downloads via browser
  - OPenDAP: An OPenDAP dataset access form with the OPenDAP link to the data as well as the metadata used to describe the data file
  - NCML: an NcML representation of the dataset
  - UDDC: an evaluation of how well the metadata contained in the dataset conforms to the netCDF Attribute Convention for Data Discovery (NACDD)
  - ISO: an ISO 19115 metadata representation of the dataset
  - **HTTPServer: Direct download of the file to your computer**
- Via Code



Gold Copy THREDDS: Code Examples

- Matlab: https://matlab.mathworks.com/users/wingardc@onid.oregonstate.edu/Published/nep_2022/matlab_thredds_erddap.html
- Python: https://nbviewer.ipython.org/gist/cwingard/08ed1a10414962d5d51c173b207074e3



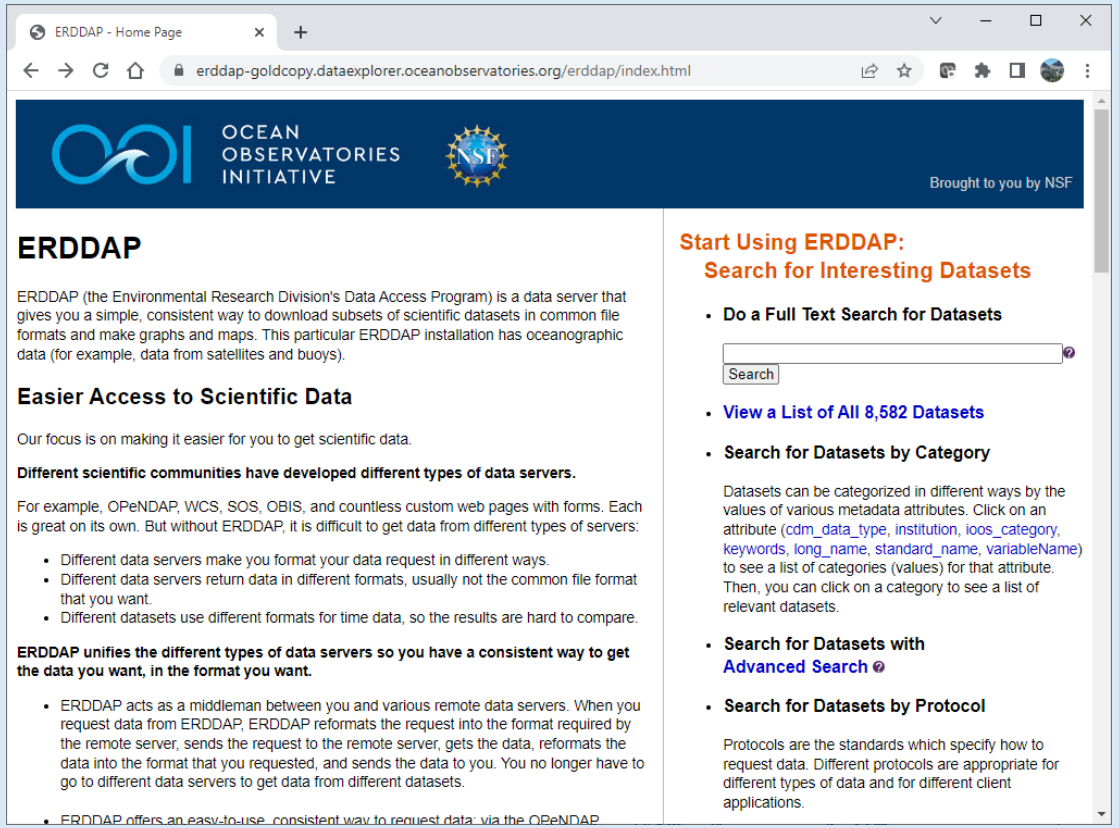


OCEANOBSERVATORIES.ORG

OCEAN OBSERVATORIES INITIATIVE

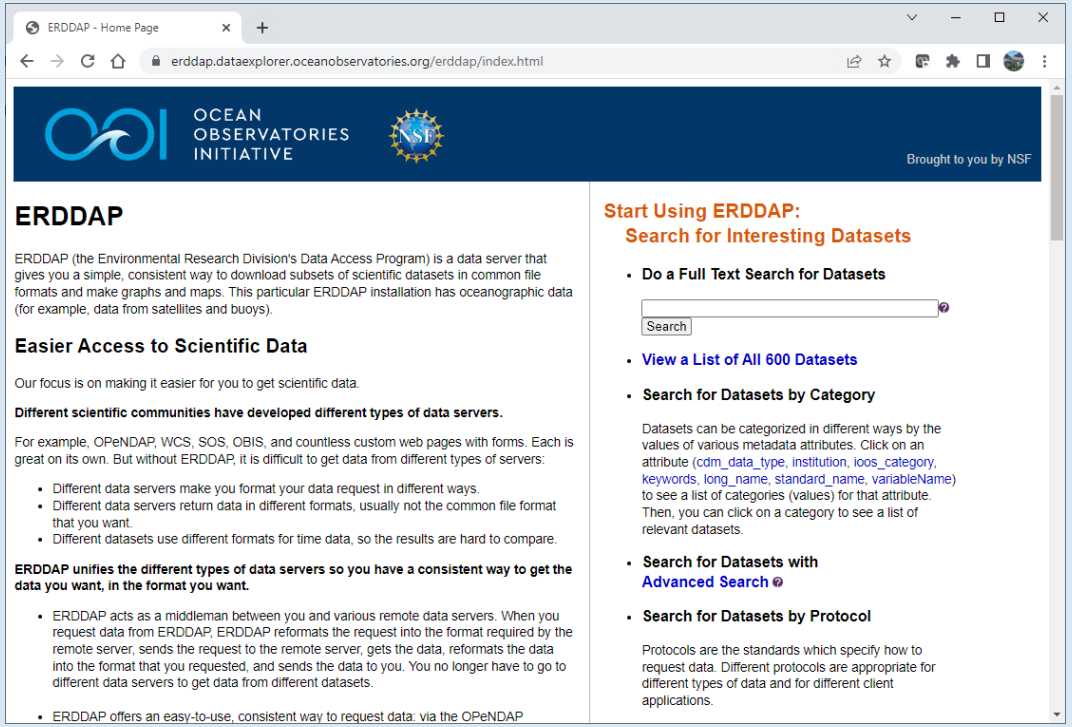
Environmental Research Division's Data Access Program (ERDDAP)

- Gold Copy ERDDAP
  - https://erddap-goldcopy.dataexplorer.oceanobservatories.org/erddap/index.html
  - netCDF files from the Gold Copy THREDDS catalogs loaded into ERDDAP datasets
  - Datasets are organized by reference designator, delivery method, stream and deployment number
  - *Based on the coded examples above, our dataset ID would be …. missing …*
    - *Some 26+ instances of ERDDAP used to represent the GC ERDDAP.*
    - *Memory and management issues*
    - *Rethinking it*
- Gold Copy THREDDS == full resolution data



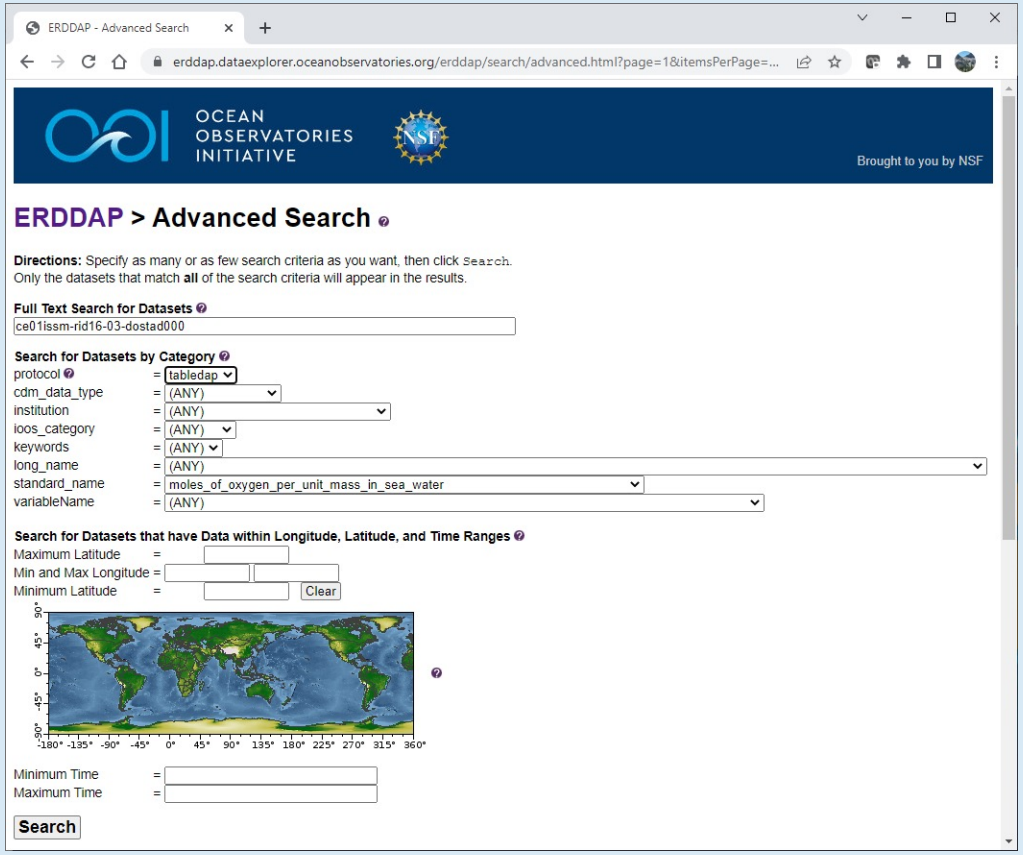
Data Explorer (Merged) ERDDAP

- https://erddap.dataexplorer.oceanobservatories.org/erddap/index.html
- Datasets are organized by the reference designator
- Data from different data delivery methods (and streams) is combined into a single, merged time-series
- Select parameters available:
  - CF Standard Name (e.g., moles_of_oxygen_per_unit_mass_in_sea_water)
  - OOI Data Product (e.g. DOCONCS)
  - *Data Explorer generated QC flags based on stock QARTOD style tests*
- High resolution data resampled to max 1 minute resolution
- Overlapping deployments are combined to create a single timeseries
- Not fully 1:1 with the Gold Copy THREDDS catalogs. Some of the more complex data sets are still being worked on



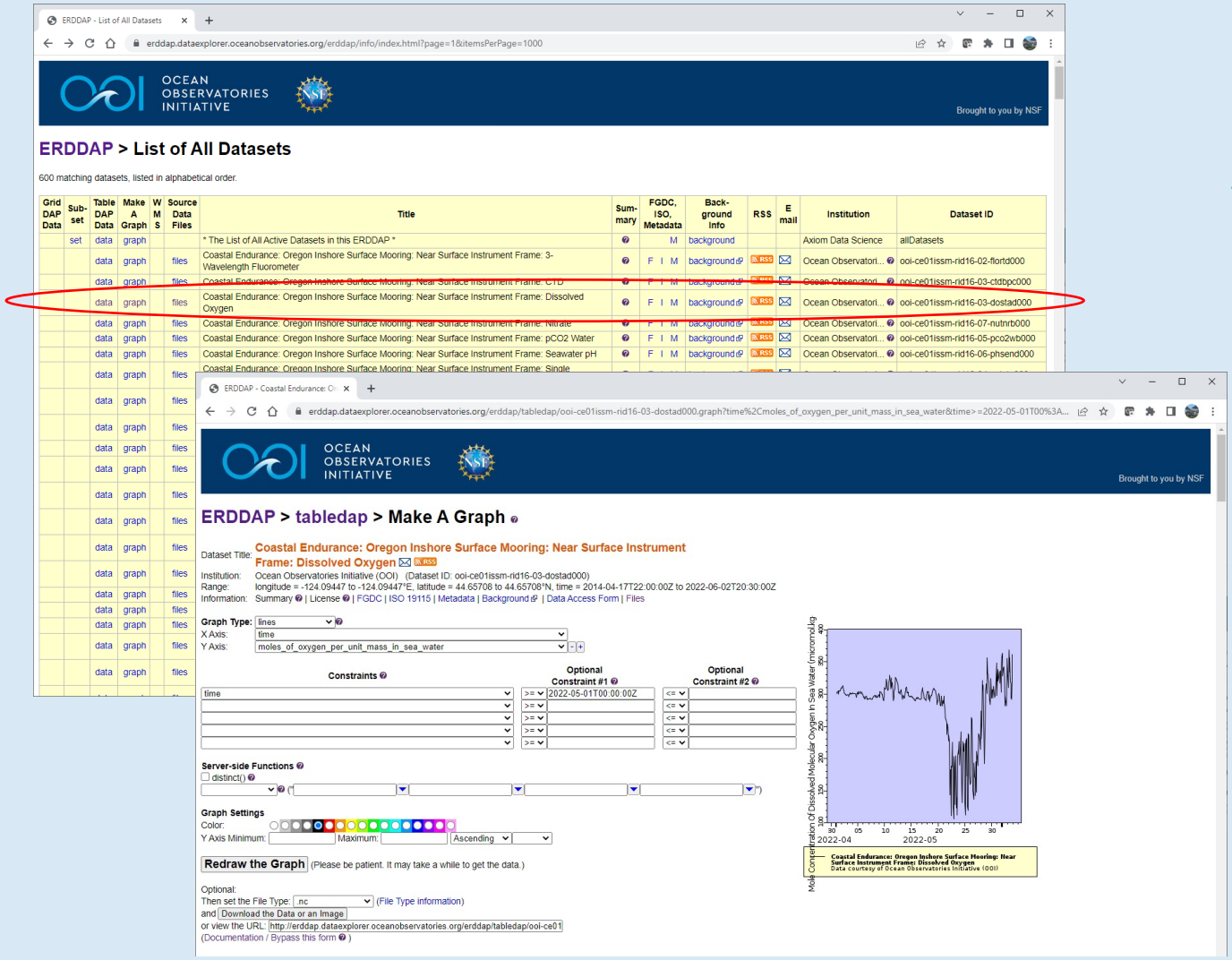
Data Explorer ERDDAP: Finding Data

- Browser
  - Advanced Search: https://erddap.dataexplorer.oceanobservatories.org/erddap/search/advanced.html
  - Datasets: https://erddap.dataexplorer.oceanobservatories.org/erddap/info/index.html
    - Scroll/CTRL-F
- Code
  - RESTful API with extensive documentation
  - Matlab: User needs to construct requests (combination of RESTful API, browser queries or OOI Decoder)
  - Python: erddapy (https://github.com/ioos/erddapy)
  - R: rerrdap (https://github.com/ropensci/rerddap)



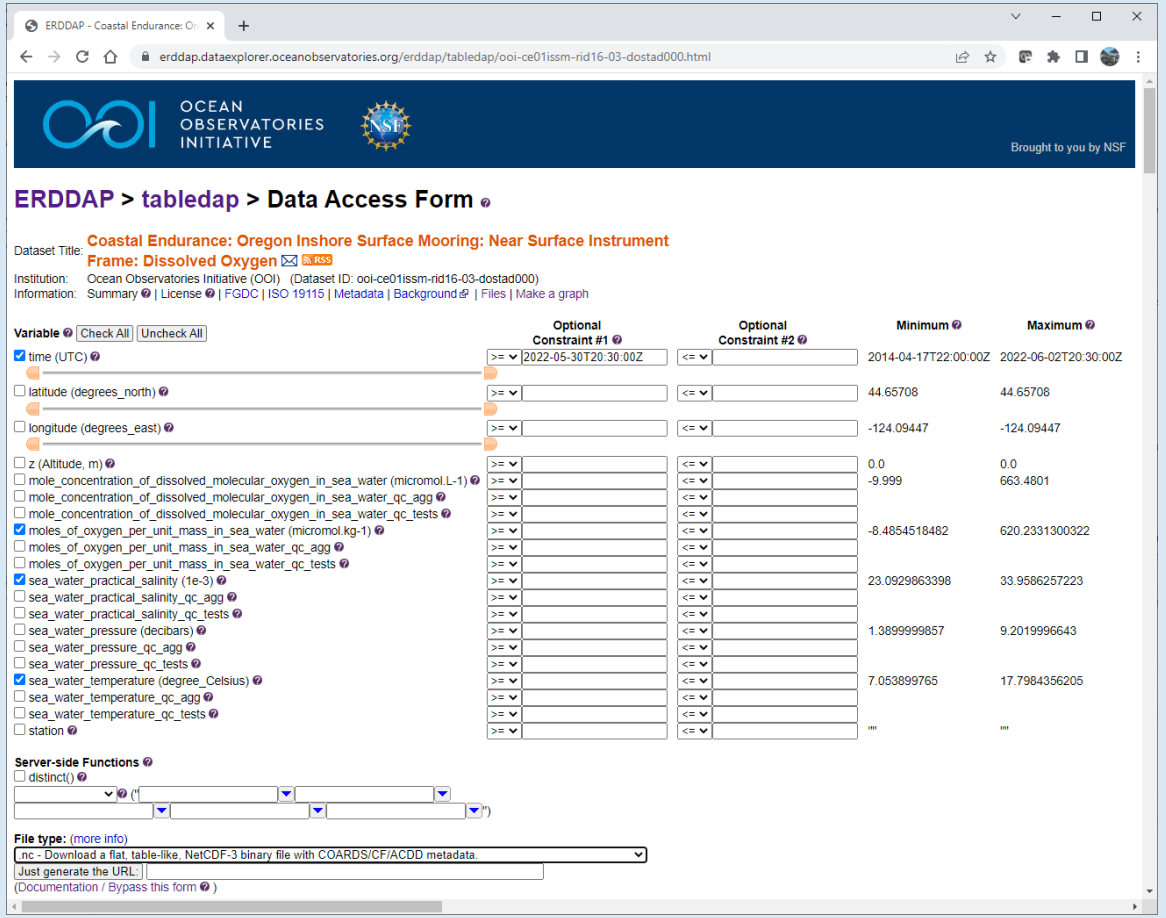
Data Explorer ERDDAP: Datasets

- Datasets are currently all represented as tabular, time-series (TableDAP) data
- Organized by Dataset ID (based on Reference Designator)
- Dataset Title human readable version of the ID
- Data access either through the “data” or “graph” links
- Datasets are updated daily from the Gold Copy THREDDS catalog



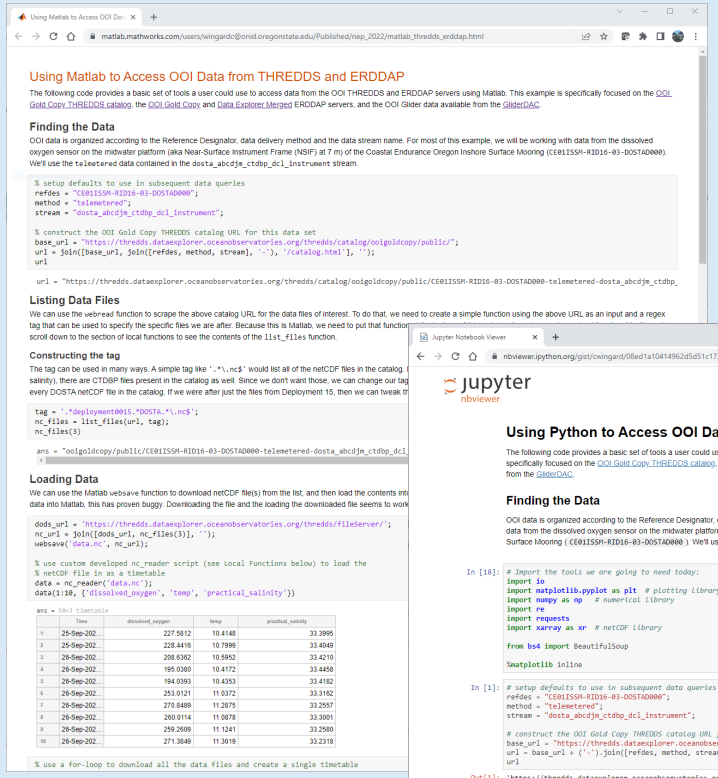
Data Explorer ERDDAP: Downloading

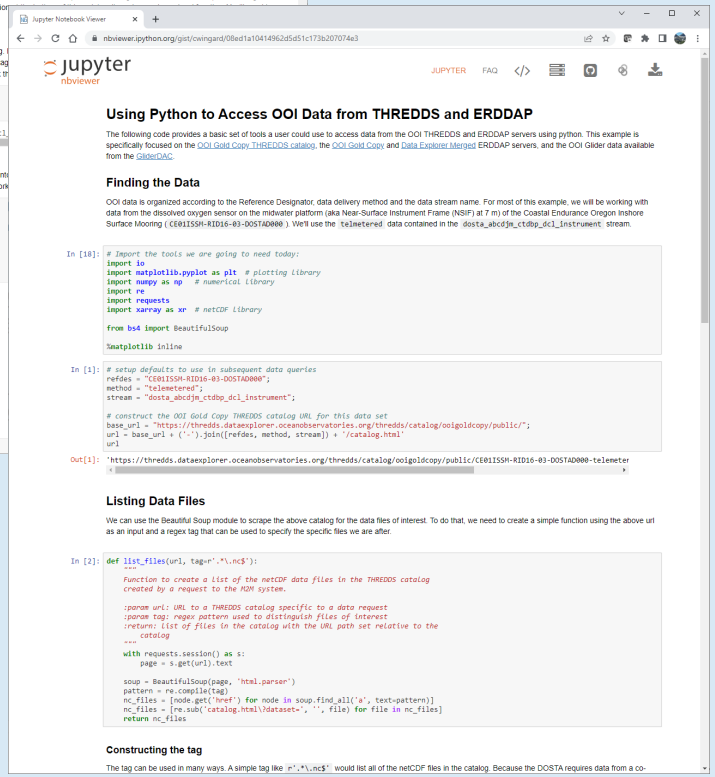
- Direct downloads via browser
  - Use the Data Access form to determine list of variables to request, set constraints (e.g., time range) and construct the URL for the file type to download.
  - ERDDAP supports 30+ file formats for downloading (in particular: csv, json, netCDF, mat)
  - Copy/paste the generated URL, or click on the Submit button (not shown, just below the File type)
  - Can also generate a URL for downloading from the Make a Graph form
- Via Code



Data Explorer ERDDAP: Examples

- Matlab: https://matlab.mathworks.com/users/wingardc@onid.oregonstate.edu/Published/nep_2022/matlab_thredds_erddap.html
- Python: https://nbviewer.ipython.org/gist/cwingard/08ed1a10414962d5d51c173b207074e3



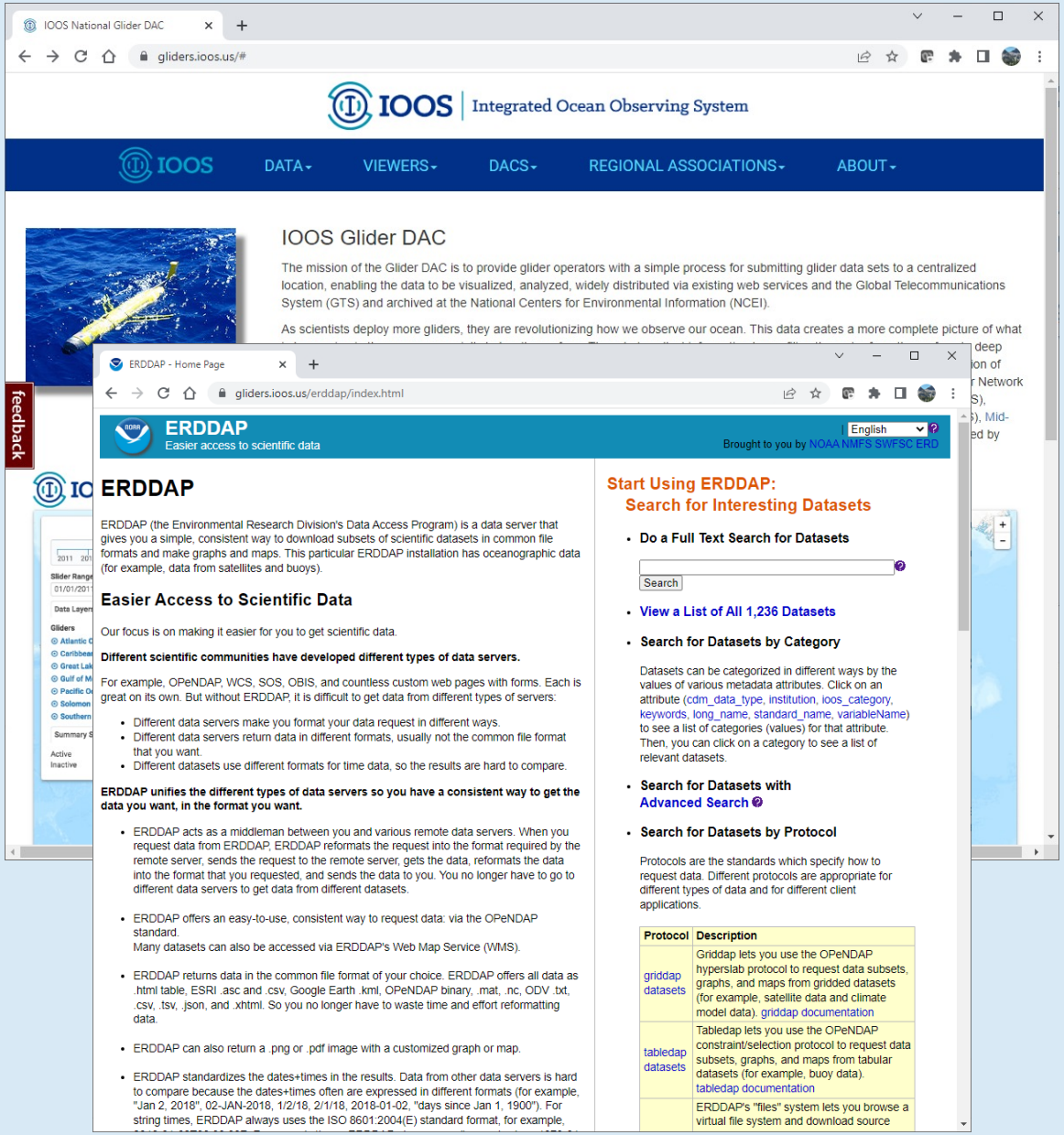


OCEANOBSERVATORIES.ORG

OCEAN OBSERVATORIES INITIATIVE

GliderDAC ERDDAP

- https://gliders.ioos.us/erddap/index.html
- Includes both the full-resolution, recovered data, as well as the decimated, telemetered data
- All OOI data is pushed to the GliderDAC on a regular schedule (actively deployed gliders are updated regularly throughout the day)
- Dataset IDs are organized by glider name and deployment date/time ('-delayed' appended to recovered datasets)
- Use 'OOI Coastal Endurance' or 'OOI Global Papa' as search terms
- For python users, see this nice example on using erddapy to access data from the GliderDAC



OCEANOBSERVATORIES.ORG

OCEAN OBSERVATORIES INITIATIVE

THREDDS versus ERDDAP?

- Not really a true cross-comparison
- They both represent a way of accessing OOI data via a users preferred set of tools (GUI, python, Matlab, R … mix and match, you choose)
- They both have well documented APIs and are commonly used in the bio-geosciences (resources available)
- They offer different variants of mostly the same data
  - THREDDS data is full resolution, covering all the different data delivery methods. User needs to stitch it all back together, but they get everything the system produces. Can be slower, more complex processing required
  - ERDDAP data has been reduced (number of variables, and temporally if data rates greater than 1 minute) but is in a more user-friendly format (already combined, standardized variable names). Data downloads are generally faster
  - Neither has the full set of OOI data, although that is the end goal

OCEAN OBSERVATORIES INITIATIVE

Questions?

OOI Discourse

https://discourse.oceanobservatories.org/

OOI HelpDesk

helpdesk@oceanobservatories.org