



National Oceanography Centre
British Oceanographic Data
Centre BODC

Using ERDDAP to expose discrete water measurement data

MEDIN Report – small bid funding

British Oceanographic Data Centre, National Oceanography Centre
Emma Slater, Richenda Houseago-Stokes and James Ayliffe

Summary

BODC have developed a semi-automated workflow to expose unrestricted discrete water measurements data from completed projects. Through this funding BODC have demonstrated that ERDDAP can deliver a one click download of this data from the MEDIN metadata record, with a URL to enable the use of the ERDDAP subset/graph features.

Background

ERDDAP is a technology developed by the National Oceanic and Atmospheric Administration's (NOAA) which has been adopted globally for marine data management ^[1]. Supporting the development and use of ERDDAP infrastructure allows federated data services and improved **F**indability, **A**ccessibility and **R**eusability of data.

BODC currently holds over 4.7 million discrete water measurement data values. These data are stored on internal BODC databases and only a small subset, approximately 14%, of discrete measurements are visible via MEDIN. Discrete water measurement data are regularly requested by scientists, and currently have to be delivered manually, through the BODC Requests Team. Making these data available would enable BODC to meet FAIR principles ^[2].

BODC are keen to adopt the ERDDAP Application Programming Interface (API) to broker discrete water measurements between BODC and MEDIN. ERDDAP is a widely adopted API tool ^[3] which will improve the **A**ccessibility of these data holdings ^[2,4] and enable users to have easier access to the data (within 2 clicks).

Through ERDDAP, a user can choose to subset the data by parameters, time frame, location and download all data which may contain data from several discrete measurements and potentially several cruises. ERDDAP provides dynamic visualisations so a user can see what they have filtered before downloading. The user is also given a large list of file formats to download the data in (e.g. csv, MatLAB, Google Earth kml, NetCDF, html etc.). By exposing BODC discrete measurements data holdings through ERDDAP, MEDIN users can choose to access the whole dataset within 2 clicks, or choose to subset data, providing dynamic access to BODC data holdings.

Planning

The MEDIN funding enabled BODC to start building a new delivery workflow for discrete water measurements. Before the development work started, the workflow was investigated, and we reviewed what users would find most beneficial. During this process, we decided to scale back and concentrate on delivering data from a completed project whose data would not be changed or updated and was completely unrestricted, allowing access for all. This approach would enable a framework to be built that could expose further projects in the future, in a manageable manner. The Liverpool Bay Coastal Observatory Discrete Water Sample Data Set was chosen to be exposed. This dataset contains 38 parameters and 35,767 data values.

Development work

BODC uses agile practices and uses the scrum methodology to complete its development work in a series of sprints.

During the BODC-in-kind development sprint BODC achieved the following:

To improve overall user/developer experience, new more powerful virtual machines (VM) were commissioned. Three in Liverpool to follow good development practice (development, test, production) and an off-site production backup in Southampton, in case of Liverpool server issues. During the server setup, it was decided to upgrade to the latest version of ERDDAP (at time of install) v2.02 from v1.82, a full change list is documented here

<https://coastwatch.pfeg.noaa.gov/erddap/download/changes.html>. Improvements were also made to how data managers access/interact with the server-side ERDDAP software.

During the MEDIN funded development sprint BODC achieved the following:

The Liverpool Bay Coastal Observatory dataset was chosen to be exported into ERDDAP. This required a new ORACLE view which could be securely accessed via a python program (csv_maker.py) using an Application Programming Interface (API). BODC used an API which comes with ORACLE called ORDS (**Oracle REST Data Services**). As mentioned, the csv_maker.py was written to generate a single or multiple csv files using the ORDS endpoint. Another python program (xml_maker.py) was written which semi-automated the creation and updating of an XML file (datasets.xml). Datasets.xml is a key part of the ERDDAP software, as it informs ERDDAP how to interact with the data file. BODC also enriched the datasets.xml file with useful metadata about the data channels, this was done via xml_maker.py connecting to the NERC vocabulary service to pull linked metadata thus reducing human error and time. The MEDIN metadata record for the Liverpool Bay Coastal Observatory Dataset was manually updated, so that it would give direct access to data. These procedures and workflows were documented, so that more datasets could be ingested in the future.

Results

Users now have two options for downloading this dataset, which was previously only available through a manual request.

- 1) One click of the URL on the MEDIN metadata record downloads the entire Liverpool Bay Coastal Observatory dataset, in csv file format (Figure 1).
- 2) A click to the ERDDAP instance of this dataset. Here users can subset (Figure 2) or graph (Figure 3) and subsequently download a subset of the dataset and choose from a range of file formats ERDDAP provides.

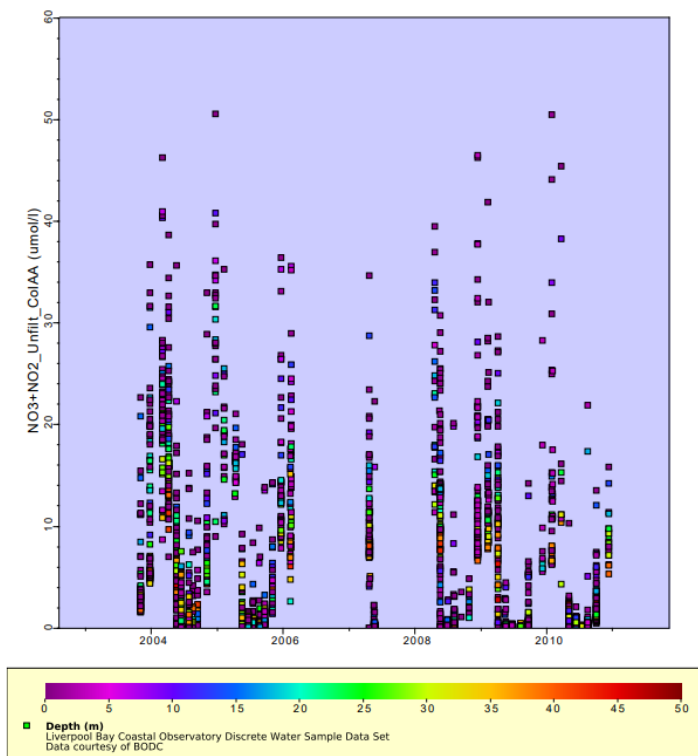


Figure 3 ERDAPP Interface graph

(https://linkedsystems.uk/erddap/tabledap/samples_cobs_20200901.largePdf?time,CPHLFLP1&time%3E=2002-08-07T11%3A34%3A39Z&time%3C=2011-11-10T00%3A00%3A00Z&.draw=markers&.marker=6%7C5&.color=0x000000&.colorBar=LightRainbow%7C%7C%7C%7C%7C&.bgColor=0xffccccff)

Future work

The workflow used in this study works well for complete projects whose data are linked to a MEDIN metadata record. This exercise could now be repeated for other projects, with the only requirement being on the database administrator and data managers' time, saving developer resource.

The workflow does not cater for updates so if there are changes in the database this would require manual intervention to bring the ERDAPP dataset up to date. Further investigation is needed before live projects, where updates to data and restrictions will occur, can be exposed using this ERDAPP procedure.

Citations

- [1] <https://coastwatch.pfeg.noaa.gov/erddap/index.html>
- [2] Benway, H., *et al.* (2020). February 2020. NSF EarthCube Workshop for Shipboard Ocean Time Series Data Meeting Report. doi: 10.1575/1912/25480
- [3] Buck, J., *et al.* (2019) Ocean Data Product Integration Through Innovation-The Next Level of Data Interoperability. *Front. Mar. Sci.* 6:32. doi: 10.3389/fmars.2019.00032

[4] Wilkinson, M., *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018. doi: 10.1038/sdata.2016.18