A case study: Transitioning the Bio-Oracle dataset version 3 to ERDDAP

Salvador Jesús Fernández Bejarano[1] (salvador.fernandez@vliz.be), Vinícius Salazar[2] (viniws@gmail.com), Jorge Assis[3] (jorgemfa@gmail.com), Bart Vanhoorne[1] (bart.vanhoorne@vliz.be), Frederic Leclercq[1] (frederic.leclercq@vliz.be), Lennert Tyberghein[1] (lennert.tyberghein@vliz.be), Olivier De Clerck[4] (olivier.declerck@ugent.be), Ester Serrão[3] (eserrao@ualg.pt), Heroen Verbruggen[2] (heroen.verbruggen@gmail.com), Lennert Schepers[1] (lennert.schepers@vliz.be)

[1]VLIZ, Flanders Marine Institute (Belgium)
[2]University of Melbourne (Australia)
[3]University of Algarve (Portugal)
[4]Ghent University (Belgium)

Understanding habitats changes is key in the context of climate change and global warming. Restoration strategies are optimized based on the identification of biodiversity hotspots and potential species niche shifts. However, the techniques that allow to gather this policy-relevant information require of harmonised, interoperable, high-quality environmental datasets, which are not always readily available in the marine realm [Tyberghein et al., 2012; Assis et al., 2017]. Such datasets are essential for species distribution modeling and detecting habitat changes. Initially released in 2012, with a major update in 2017 incorporating benthic data layers and future climate scenarios aligned with IPCC predictions, Bio-ORACLE now introduces Version 3. This version includes substantial updates not only in data but also in access methodologies, transitioning from a file-based system to an ERDDAP server. Our experience illustrates a case study of transitioning an existing database into ERDDAP and, from that, we offer a concise set of recommendations for teams considering to do the same.

Bio-Oracle version 3

The Bio-Oracle database (https://bio-oracle.org/) was traditionally disseminated as raster files, downloadable from the Bio-ORACLE website or retrieved using the R package sdmpredictors. Bio-Oracle encompasses 18 environmental variables such as ocean temperature, salinity or concentration of phytoplankton. Version 3 enhances the resolution of the dataset on multiple levels: spatial resolution increases to 0.05 degrees, and future time predictions expand to ten decades per variable, encompassing the years 2000 - 2010, 2010 - 2020, and so forth until 2090 - 2100 [Assis et al., under review]. Future climate change scenarios are updated to include the latest Shared Socioeconomic Pathways (SSP) from the IPCC Sixth Assessment Report.
To address the challenge of heavy data files resulting from increased resolution, Bio-Oracle Version 3 adopts NetCDF grids. This binary file format allows subsetting smaller portions of the dataset, essential for efficient data handling and analysis. Each NetCDF accommodates environmental variables with dimensions of longitude, latitude, and time, aligning with the Climate Forecast conventions. This also facilitates integration into an ERDDAP instance, freeing the client from dealing with large data files by outsourcing the slicing and other capabilities to the server side. ERDDAP offers a consistent way to download subsets of scientific datasets, make graphs, and maps, with a focus on simplifying scientific data retrieval.

Publishing via ERDDAP

A dedicated ERDDAP server was installed at the Flanders Marine Institute Data Centre to host Bio-Oracle Version 3 (https://erddap.bio-oracle.org). Our ERDDAP instance is encapsulated

within a Docker container, ensuring a self-contained and reproducible deployment environment. Its configuration is available at <https://github.com/bio-oracle/bio-oracle-erddap>. We built upon the IOOS 'gold standard' ERDDAP configuration (https://github.com/ioos/erddap-gold-standard). Custom scripts for automated data processing are integrated in our ERDDAP server configuration, facilitating seamless updates and maintenance. Automation reduces manual intervention, minimizes errors, and ensures the dataset's timely availability for data consumers.

Accessing Bio-Oracle v3

ERDDAP provides both a web interface and a REST API, allowing server-side operations on datasets. Users can retrieve specific slices of data, optimizing bandwidth use. ERDDAP supports up to 20 file types for downloading data, including NetCDF, CSV, GeoTiff, and PNG plots. This flexibility is advantageous for users interested in specific regions, realms, or climate change scenarios and decades. ERDDAP API's consistency across datasets simplifies integration with other data sources. For instance, the Irish Marine Institute meticulously curates a registry of ERDDAP projects and deployments at <https://github.com/IrishMarineInstitute/awesome-erddap>. This registry, systematically accessed by versatile clients like rerddap in R, facilitates comprehensive data searches across diverse ERDDAP deployments [Chamberlain S., 2023]. The rerddap R package offers functionalities for listing and downloading data within the R environment, while Erddapy serves as its Python counterpart [Fernandes F., 2017]. In addition to these general-purpose clients, we have developed dedicated R and Python extensions, namely 'pyo_oracle' and 'biooracler,' tailored for seamless exploration and retrieval of Bio-Oracle data. These dedicated packages streamline the assimilation of Bio-ORACLE data into prevailing bioclimatic modeling workflows. They are available at <https://github.com/bio-oracle/pyo_oracle> and <https://github.com/bio-oracle/biooracler> respectively. The development of these API clients underscores our commitment to accommodating a broad spectrum of scientific frameworks and programming languages, nurturing collaborative and interdisciplinary research endeavors. Furthermore, the dataset is poised for dissemination across novel data platforms and initiatives, including the European Digital Twins of the Ocean, the Microsoft Planetary Computer, and Google Earth Engine. Leveraging a renowned standard server like ERDDAP will be instrumental in facilitating data sharing with these initiatives.

References

Assis J. et al., (2017). *Bio-ORACLE v2.0: Extending marine data layers for bioclimatic modelling*. Global Ecology and Biogeography, 27, 277–284.

Assis J. et al., (under review). *Bio-ORACLE v3.0. Pushing marine data layers to the next-generation scenarios of climate change research*. Global Ecology and Biogeography.

Chamberlain S., (2023). *rerddap: General Purpose Client for 'ERDDAP' Servers*. https://docs.ropensci.org/rerddap/, https://github.com/ropensci/rerddap

Fernandes F., (2017). *Python interface for ERDDAP*. https://ioos.github.io/erddapy/, https://github.com/ioos/erddapy

Tyberghein L. et al., (2012). *Bio-ORACLE: A global environmental dataset for marine species distribution modelling*. Global Ecology and Biogeography, 21, 272–281.