EMSO ERIC Data Services: managing distributed data through an ERDDAP federation

Antoine Queric, Ifremer (France), antoine.queric@ifremer.fr
Rob Thomas, Marine Institute (Ireland), rob.thomas@marine.ie
Maurice Libes, CNRS (France), maurice.libes@osupytheas.fr
Enoc Martinez, Universitat Politècnica de Catalunya (Spain), enoc.martinez@upc.edu
Claudia Fratianni, INGV (Italy), claudia.fratianni@ingv.it
Tania Morales, PLOCAN (Spain), tania.morales@plocan.eu
Helen Snaith, BODC NOC (UK), h.snaith@bodc.ac.uk
Maria Sotiropoulou, HCMR (Greece), marsot@hcmr.gr
Sylvie Van Iseghem, Ifremer (France), sylvie.van.iseghem@ifremer.fr
Raluca Radulescu, GeoEcoMar (Romania), raluca.radulescu@geoecomar.ro
Paulo José Relvas de Almeida, UAlg (Portugal), prelvas@ualg.pt
Raul Bardaji, UTM-CSIC (Spain), bardaji@utm.csic.es
Ivan Rodero, EMSO ERIC CMO (Italy), ivan.rodero@emso-eu.org

Bringing together distributed data

EMSO is a consortium of partners sharing a common strategic framework of scientific facilities (data, instruments, computing and storage capacity). Formally, it is a European Research Infrastructure Consortium (ERIC), a legal framework created for pan-European large-scale research infrastructures. EMSO ERIC consists of a system of regional facilities/observatories placed at key sites around Europe, from the North East Atlantic, through the Mediterranean, to the Black Sea. The observatories are platforms equipped with multiple sensors, placed through the water column and on the seafloor. These observatories constantly measure different biogeochemical and physical parameters that address natural hazards, climate change and marine ecosystems.

Central to the EMSO ERIC mission is the collection, curation and provision of high-quality oceanographic measurements for the assessment of long-term trends. EMSO ERIC regional facilities collect a variety of data spanning oceanographic measurements through video and acoustic data types. In developing an integrated data management ecosystem, EMSO ERIC faced the common challenges where data sources that are of different volume, velocity and variety are hosted across many partners. EMSO ERIC is committed to ensuring that datasets fulfil the FAIR principles of being Findable, Accessible, Interoperable, and Reusable. The primary goal of this ecosystem is to deliver data and products from the aggregation of sources from the regional facilities in a reliable and integrated manner. It facilitates the user with a single access point to all EMSO ERIC observatories through harmonization processes. It also offers tools that enable users to easily find and access data assets, including data portals, application programming interfaces (APIs), dashboards, and a virtual research environment. This presentation outlines experiences for delivering a federated ERDDAP system, which complements other data access mechanisms of the EMSO ERIC data management ecosystem.

EMSO ERIC ERDDAP federation

ERDDAP is a data server created by NOAA in the United States that provides a simple, consistent way to serve data on the web. ERDDAP is free and open source. It uses Apache-like licenses, so it can be adapted or enhanced to fit a user's requirements. Users can download subsets of gridded and tabular scientific datasets in common file formats and make graphs and maps, which can be embedded in web pages and can be configured to update with the latest data available. In addition to a web interface, ERDDAP also provides a RESTful API that allows users to programmatically interact with the data using

scripting languages such as Python, R or Matlab, and can be used to provide data to dashboards and other applications. ERDDAP is well suited to the distributed data requirements of EMSO ERIC as data do not need to be transformed from the local storage format of choice (e.g. flat files, relational database, noSQL database) to one “master” format in order to be served. This means EMSO ERIC is not forcing data architecture decisions onto each partner who may not fit with their existing architecture. ERDDAP can be set up to serve data from the storage structures already in place at each organization and it automatically aggregates files of the same XML model within a dataset. This is a useful feature for users interested in a subset of a time-series who don’t want to have to stitch together a series of files themselves before working with the data. ERDDAP also facilitates interoperability and data reuse since it is able to take a variety of formats as input and output them into user preferred file formats.

Architecture and implementation

In order to provide a single ERDDAP endpoint for EMSO ERIC end-users, a number of architecture choices had to be evaluated. The relative benefits and costs for three solutions were considered: EMSO ERIC data harvested centrally and served from one single ERDDAP server; each partner serves data through their own ERDDAP server and references to other servers; a distributed/federated network of ERDDAP servers. The ERDDAP federation model was agreed to be the optimum solution, and it was decided to investigate the ERDDAP federation functionality for its implementation, i.e., ERDDAP provides the ability to reference datasets served from other ERDDAP servers. The configuration is based on a simple URL that leads to the desired dataset hosted by a remote ERDDAP server. The central EMSO ERIC ERDDAP server then provides those datasets as if they are locally hosted (Figure 1).

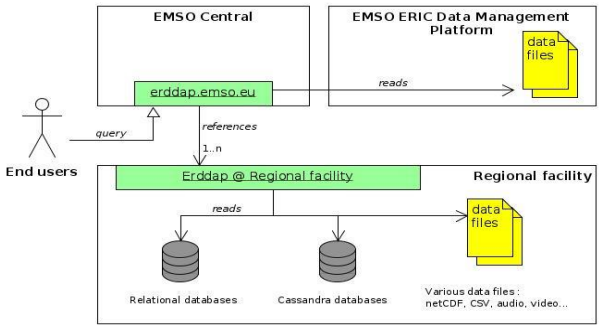

Figure 1: EMSO central ERDDAP server (http://erddap.emso.eu) references remote datasets

Whereas some EMSO ERIC partners already have experience setting up and delivering data through ERDDAP servers, not all EMSO ERIC regional facilities are currently participating in the ERDDAP federation. Undergoing efforts include the deployment of dedicated ERDDAP servers at EMSO ERIC regional facilities and the integration of datasets from the EMSO ERIC data management platform into ERDDAP (see Figure 1). It gives the user the appearance that all data are being sourced from one single location. A further goal is to provide end-users with meaningful examples for the usage of the datasets within the EMSO ERIC virtual research environment (https://jupyter.emso.eu/). Default queries are configured to show the users what they can do with the data for each dataset.

Application of community standards and formats

While ERDDAP provides a technical architecture to achieve the Findable and Accessible components of the FAIR principles, the content still needs to be well managed and marked up in order to achieve Interoperability and Reusability. Ensuring consistent metadata markup from community vocabularies and populating the metadata (or attributes) of community standards, for example, CF, OceanSITES, and SeaDataNet, is an ongoing process. Rather than bespoke code and processes being maintained locally to produce a variety of community formats, the open-source nature of ERDDAP opens possibilities for EMSO ERIC, as well as other communities across the marine domain, to contribute enhancements to future releases of ERDDAP to deliver these formats.