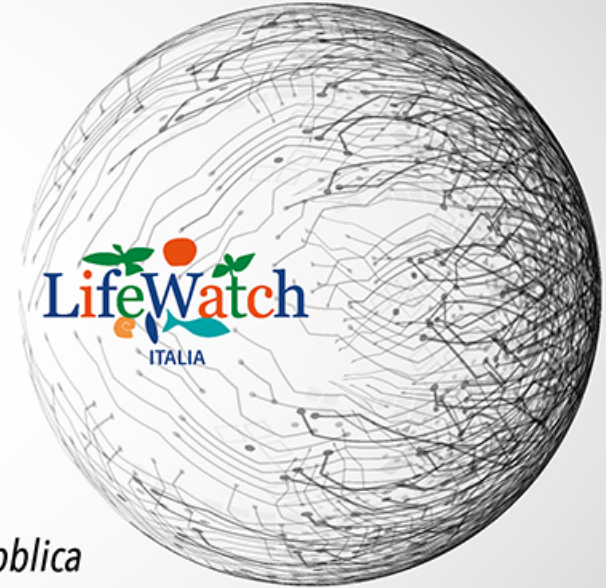


# Conferenza Annuale di LifeWatch Italia Roma, 25-27 giugno 2018



*in collaborazione con il Segretariato Generale della Presidenza della Repubblica*

**Processing of occurrence biodiversity datasets to be distributed in the  
European infrastructures  
EMODnet Biology and SeaDataCloud**

**M. Eliezer, M.Lipizer, B. Cataletto, A. Giorgetti**



# Outline

- Introduction
- QC
- Format conversion
- DOI minting and Data publishing
- EMODnet Biology data product based on LTER-C1 station
- SeaDataCloud new VRE and online QC

# Introduction

- OGS is recognised as the Italian National Oceanographic Data Centre (OGS-NODC) within the International Oceanographic Data Exchange (IODE) System.
- Since 2013, OGS-NODC has extended the range of parameters managed, including also marine biodiversity data. This new data type required specific Quality Control procedures and conversion to standard formats.



# Process

- Data arrives in spreadsheets.
- QC and gathering extended metadata.
- Conversion to two different standard data formats required by EMODnet Biology and SeaDataCloud, and DOI minting.
- More QC is done by these European infrastructures.
- Long term availability is guaranteed by both OGS-NODC and the two European clouds. Data is provided also by web services.

# Initial QC

- Duplicates?
- Taxon match using WoRMS online tool.  
Originators contact for resolving ID or name problems.
- Basic checks: lifeStage conforms to BODC vocabulary S11? Coordinates are decimal?...
- Missing metadata? e.g., the sampling process

# Data format

- Data usually arrives in tables designed for an easy compilation, but they are not ready for data consumption.
- In case of current data originators, we've agreed to use a template which is still very originator oriented.
- However, in case of “archaeological” data, data templates are different and metadata gathering is also challenging.

# Data format

- In most cases the data arrives in a one row per each species: A very wide row with values at different date, depth, and even measurement type.
- Reading from the beginning of the row, the related information is the one last mentioned, e.g., the author may not repeat the eventDate if it remains unchanged.

# Data format

- The format which is the base of both EMODnet Biology and SeaDataCloud has only one measurement value in a row.
- A manual conversion is error-prone and time-consuming. So, we've developed a Java program.
- Naturally this program is also doing some Quality Control.



# Automatic data conversion

- Input: data.csv and different complementary CSV files.
- Output files, DwC terms:
  - event.csv for the Event Core
  - occurrence.csv for the Occurrence Extension
  - extendedmeasurementorfact.csv for the eMoF Extension
- Upon need the program creates size.csv to be mapped as eMoF, or other eMoF files.

# Additional QC

- Another duplicates test.
- Match between term and ID in the sampling descriptors of the eMoF.
- Missing information.

# Publishing

- These output files are unified to a flat table to be used in the conversion to SeaDataCloud ODV and CDI.
- Same output files are also mapped using GBIF IPT web application for publishing DwC-A file that is harvested by EMODnet Biology providing also OBIS standard format.
- The data object for the DOI minting includes both flat table and the DwC-A file.

# Publishing

- Dataset is also imported to NODC database, and SeaDataCloud publication includes both CDI for metadata and biological variant of ODV for the data file.
- The dataset can be found via SeaDataCloud and via the search form in OGS-NODC web site.

# FAIR data

- As a result of this process the data corresponds to FAIR principles. It's Findable, Accessible, Interoperable, and Reusable.
- Its long term availability is provided by both OGS-NODC and two important European infrastructure.

# EMODnet Biology data product based on LTER-C1 station (Trieste, Italy)

- LTER-C1 station is a long term ecological research site (LTER-Italy) starting at 1986.
- LTER North Adriatic plankton series web interface can be found at: <https://pherman.shinyapps.io/shiny/>
- This data product allows end user to filter and to plot in order to have some insights even before downloading data.
- The web application was written in R programming language by Deltares: Peter.Herman@deltares.nl



# SeaDataCloud new VRE and online QC

- SeaDataCloud will storage data provided by its partners on a cloud, providing also a Virtual Research Environment (VRE).
- In addition, an online biodiversity QC developed by VLIZ, based on the same concept of LifeWatch Belgium data services.



# Connecting to related environmental data

- Physical and chemical data published via SeaDataCloud arrive also to EMODnet Chemistry which has specific themes about nutrients etc. There are different ways to help end user connect biotic and abiotic data: link inside the abstract, DOI relations, SeaDataCloud directories such as EDMED and CSR, new OBIS-ENV-DATA.