



**SeaDataNet**

PAN-EUROPEAN INFRASTRUCTURE  
FOR OCEAN & MARINE DATA  
MANAGEMENT

## ***Development of online Biology Data QC***

***Simon Claus***

***VLIZ- Flanders Marine Institute***



**SeaDataNet**

PAN-EUROPEAN INFRASTRUCTURE  
FOR OCEAN & MARINE DATA  
MANAGEMENT

- Developing a Virtual Research Environment with a packaged set of advanced downstream services for users: Services will include:
  - MySeaDataCloud
  - Sub-setting
  - Online version of the Ocean Data View (ODV) software
  - Online version of the Data-Interpolating Variational Analysis software
  - High level visualisation
  - Upgraded Oceanotron
  - SOS viewing services
  - **Online Biology Data Quality control**

## WP10.2.4: development of online Biology Data quality checks (QC)

This task will develop online services and tools to analyse the quality and completeness of biology data. Records will be reviewed through a series of QC steps dealing with the

- 1) data format
- 2) completeness and validity of information
- 3) quality and detail of the used taxonomy and
- 4) geographic indications and
- 5) whether or not the record is a (possible) outlier.

## WP10.2.4: development of online Biology Data quality checks (QC)

The QC procedures will be developed **as online web services** that can be used by potential data providers and researchers to:

Assess the quality and completeness of their own data prior to  
1) use or 2) submission.



PAN-EUROPEAN INFRASTRUCTURE  
FOR OCEAN & MARINE DATA  
MANAGEMENT



- Large-scale European research infrastructure
- Virtual laboratory for study of biodiversity
- Integrates observatories, data bases, web services and modelling tools distributed throughout Europe.
- Keywords: E-science, **web services**, **data services**, ICT infrastructure, HPC, GRID, BIG data, workflow
- Goals: increase data generation, real time monitoring data, biosensors



# WoRMS

World Register of Marine Species

Home

About

Search taxa

Taxon tree

Literature

Distribution

Specimens

**Match taxa**

Editors

Statistics

Users

Webservice

Photogallery

Info downloads

Sponsors

Glossary

Manual

Log in



RSS



Add provider



@WRMarineSpecies

## WoRMS Taxon match

You can use the WoRMS Taxon Match Tool ([credits](#)) to automatically match your species list or taxon list with WoRMS. After matching, the tool will return your file with the AphiaID's, valid names, authorities, WoRMS classification and/or any other output you selected. [[View manual](#)]  
For performance reasons, the limit is set to 5,000 rows. You can send larger files to [info@marinespecies.org](mailto:info@marinespecies.org) and we will return the results to you as soon as possible.

File\*

Allowed filetypes: Plain text [TXT], Comma Separated [CSV] & Excel Sheet [XLS, XLSX]

Row delimiter   First row contains column names

Column delimiter

Match authority

Match upto  Higher taxa only possible if a full classification is given in additional columns

Limit to taxa belonging to

Output  AphiaID  LSID  TSN  ScientificName  Authority  Accepted name  Classification  Qualitystatus   
 Taxon status  Environment  Citation

This tool uses the following components:

- ✓ TAXAMATCH fuzzy matching algorithm by Tony Rees
- ✓ PHP/MySql port of TAXAMATCH by Michael Giddens
- ✓ Scientific Names Parser by Dmitry Mozzherin

Example: taxonomic qc:

### 1. Define relevant taxonomic qc functions:

- get the AphiaID for your taxon
- check the spelling of your taxa
- get the full classification for your taxa
- resolve your unaccepted names to accepted ones
- get all synonyms for a taxon
- fuzzy/near match your species list
- resolve a common name/vernacular to a scientific name...

### 2. Open webservice: platform-independent SOAP/WSDL standard.

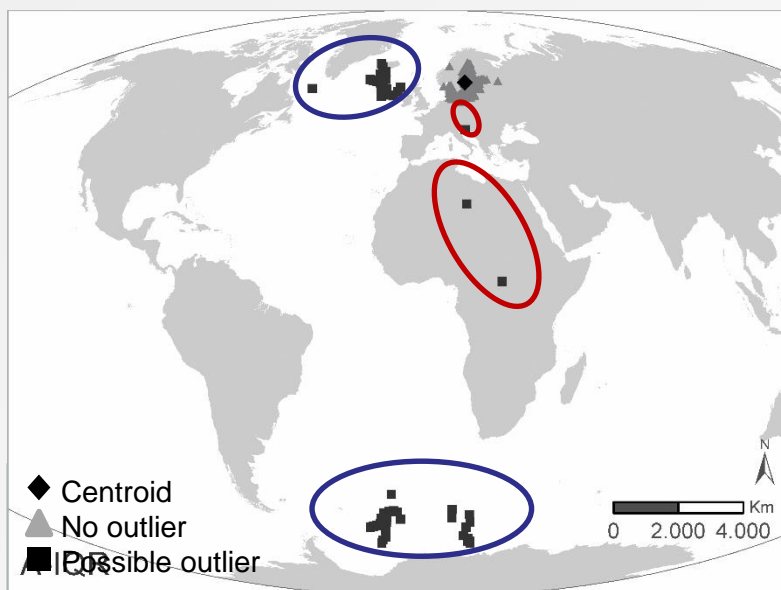
### 3. Implementations in SDC tools developed for submitting (MIKADO?) or using data





SeaDataNet

## Outlier analysis on dataset level

- Analysis on dataset level
- Possible location outlier(s)
- Methodology based on centroid calculations and assuming normal distribution => not applicable for strong asymmetric datasets...
- Communication with provider on results



Dataset: “ICES Biological Community” (DOME)

-  Also identified as incorrect in record-level check of lat-lon (=land)  
Not identified through record-level check of lat-lon (=sea), but seen as potential outlier through geographic outlier check
-  check of lat-lon (=sea), but seen as potential outlier through geographic outlier check

Provider communication:

- Antarctic locations are incorrect (data error)
- Northern locations are correct (sampling bias)

Vandepitte et al. (2015). *Fishing for data and sorting the catch [...]. Database*. DOI: 10.1093/database/bau125

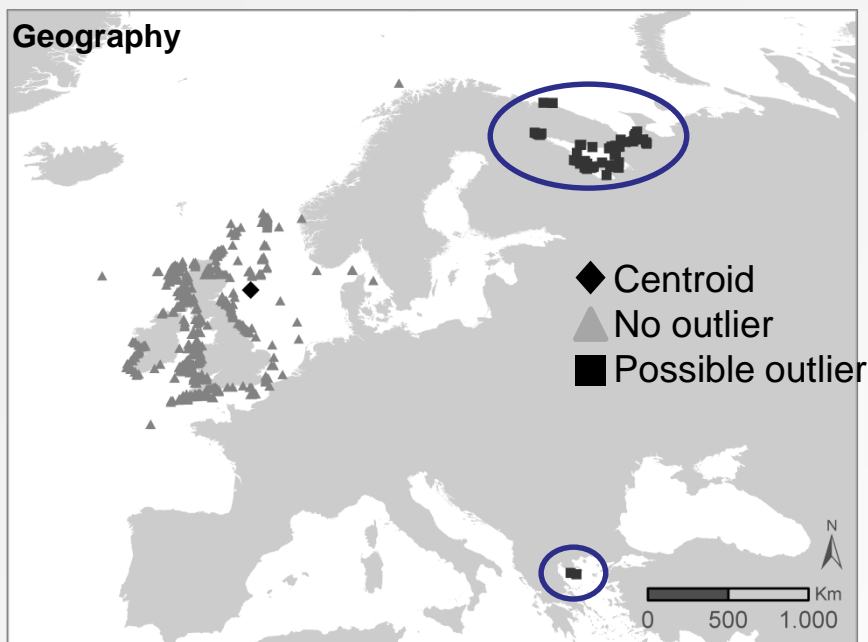




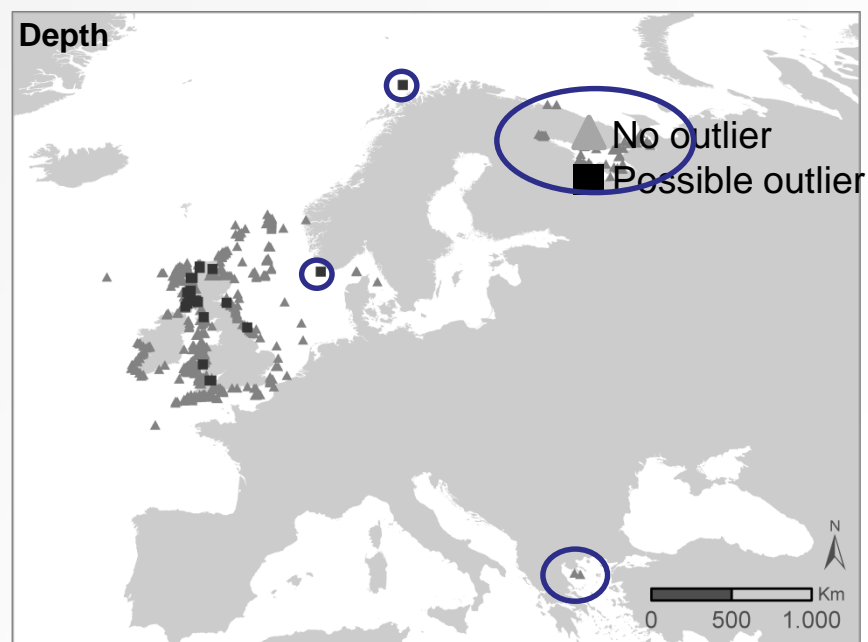
## SeaDataNet

# Outlier analysis based within the available distribution records of a species - Environmental outliers

- => Check for outliers within the available distribution records of a species
- => Geography, depth, sea surface salinity (SSS), sea surface temperature (SST)



*Verruca stroemia* (Crustacea: Cirripedia)



*Vandepitte et al. (2015)*

## WP10.2.4: development of online Biology Data quality checks (QC)

### D10.9: Specification of Biology Data QC online and development plan (M12)

1. Analyse and select relevant biological qc steps (OBIS, EurOBIS, WoRMS, ICES...)
2. Analyse how these services can be integrated in SDC and made available for SDC tools

D10.10: Phase 1 of Biology Data QC online operational (M24)

D10.11: Phase 2 of Biology Data QC online operational (M36)

D10.12: Phase 3 of Biology Data QC online operational (M42)

- *Technical:*

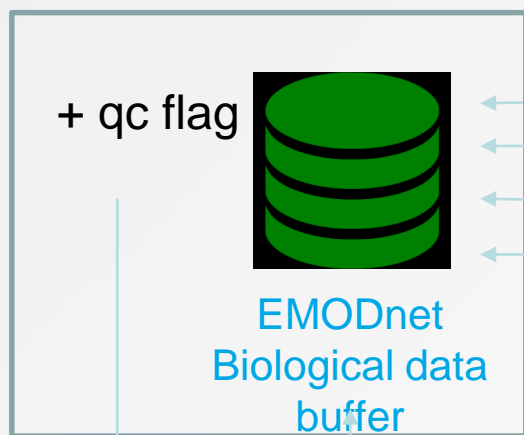
- 18 quality control steps, on individual record level
  - 10 outlier checks, on dataset or species level
  - Each QC step = yes (1)/no (0) question
  - Creation of a bit-sequence ( $2^{(x-1)}$ )
- => stored as an integer value for the QC
- => unique value for each possible combination

QC step	Value	Bit-seq.
1	1	$2^{(1-1)} = 1$
2	1	$2^{(2-1)} = 2$
3	0	$= 0$
4	1	$2^{(4-1)} = 8$
5	0	$= 0$
<b>TOTAL</b>		<b>= 11</b>

QC step	Value	Bit-seq.
1	1	$2^{(1-1)} = 1$
2	1	$2^{(2-1)} = 2$
3	1	$2^{(3-1)} = 4$
4	1	$2^{(4-1)} = 8$
5	1	$2^{(5-1)} = 16$
<b>TOTAL</b>		<b>= 31</b>



VRE



SOAP – Rest - WFS



?

A serene sunset scene over a beach. The sun is low on the horizon, partially obscured by clouds, casting a warm, golden glow across the sky and reflecting on the wet sand. In the foreground, a dark, vertical post stands in the shallow water, with a small bird perched on its top. The overall mood is peaceful and contemplative.

Thank you for your attention!