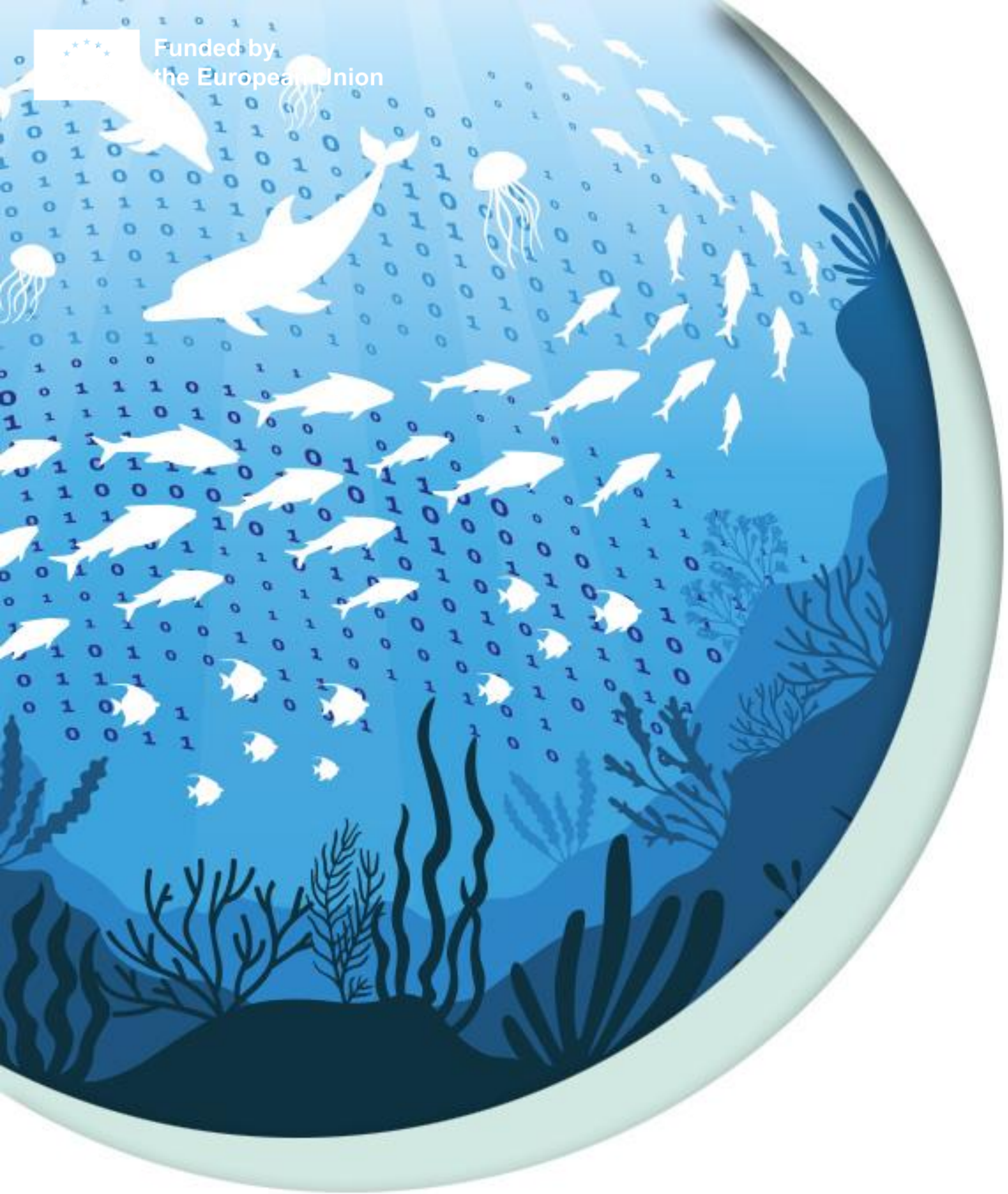


Funded by
the European Union

DTO-BioFlow

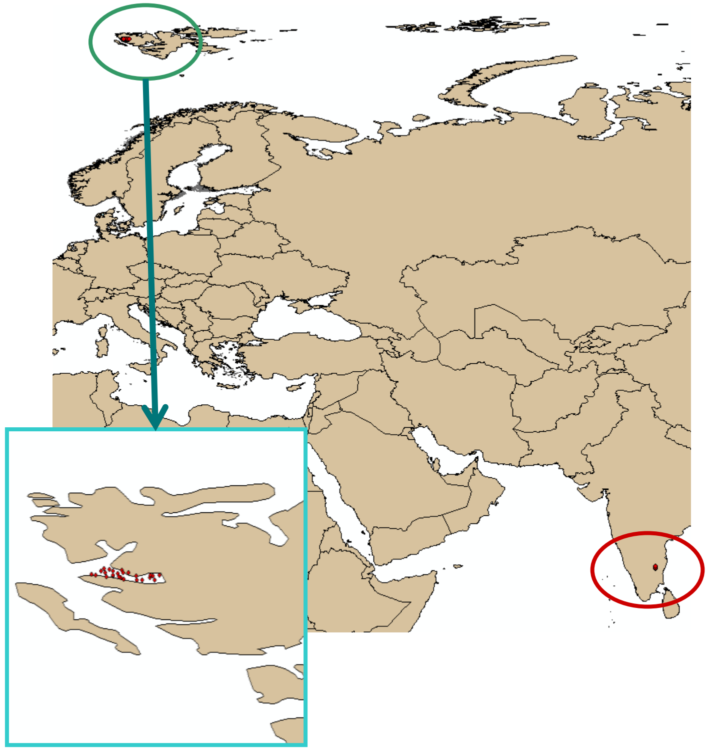
Integration of biodiversity monitoring
data into the Digital Twin Ocean

DTO-BioFlow data training workshop:

Quality Control in data management

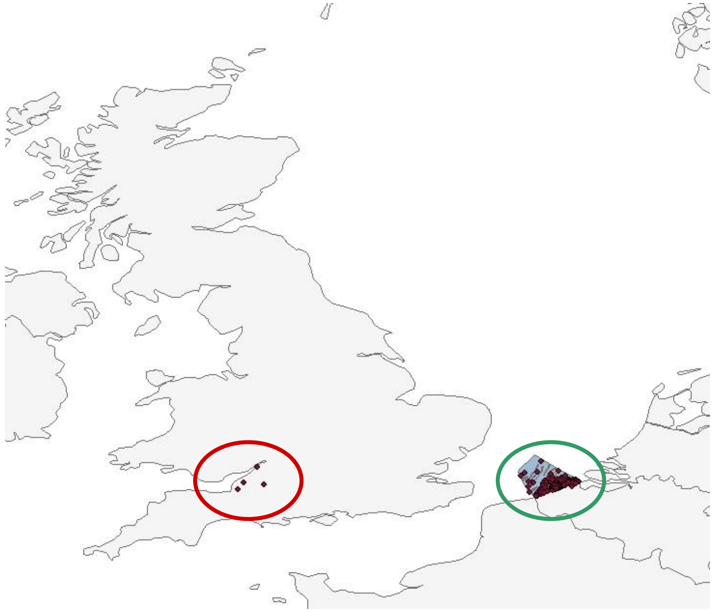
Why is quality control important?

“Monitoring in Kongsfjorden area”



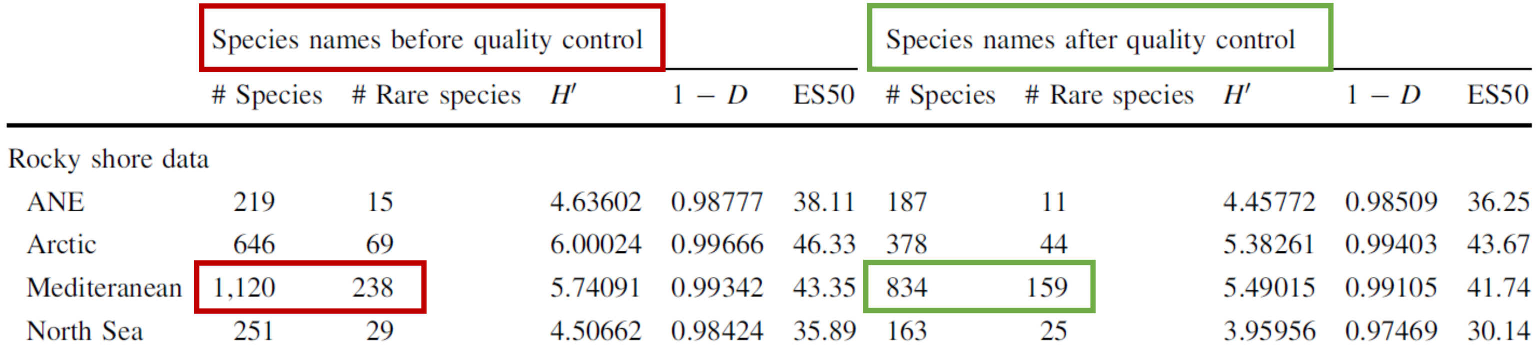
Latitude & longitude switched

“Monitoring in Belgian part of the North Sea”



“+” & “-” signs switched

Why is quality control important?



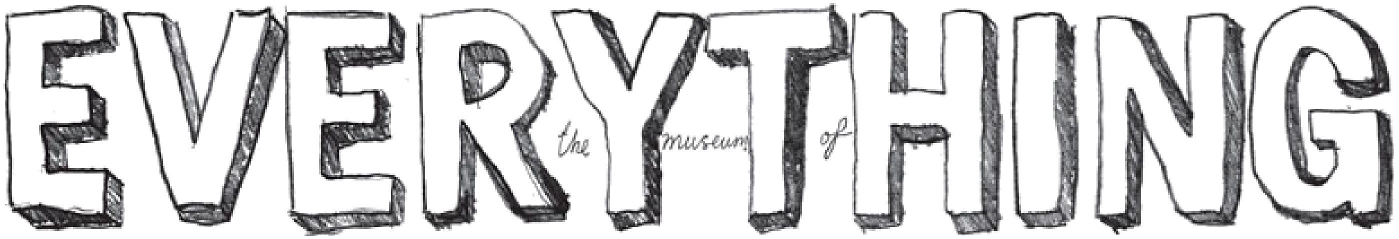
	Species names before quality control					Species names after quality control				
	# Species	# Rare species	H'	$1 - D$	ES50	# Species	# Rare species	H'	$1 - D$	ES50
Rocky shore data										
ANE	219	15	4.63602	0.98777	38.11	187	11	4.45772	0.98509	36.25
Arctic	646	69	6.00024	0.99666	46.33	378	44	5.38261	0.99403	43.67
Mediteranean	1,120	238	5.74091	0.99342	43.35	834	159	5.49015	0.99105	41.74
North Sea	251	29	4.50662	0.98424	35.89	163	25	3.95956	0.97469	30.14

“From 6,172 unique taxon names […] to 4,525, mostly due to spelling variations and synonymy.”

“ … Such [taxonomic] quality control is highly needed, since a misspelled or obsolete name could be compared to the introduction of a rare species, with adverse effects on further (biodiversity) calculations…”

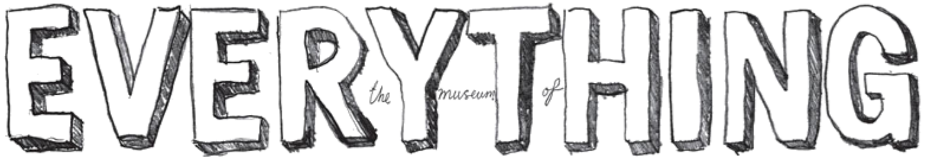
Source: Vandepitte *et al.* (2010)

What needs to be QCed



What needs to be QCed

- Adherence to chosen standard
- Content format (dates, coordinates, etc.)
- Unique value fields (IDs)
- Duplicated records and redundant information
- Impossible values and outliers
- Mandatory data
- Completeness (units, geodetic datums, etc.)
- ...

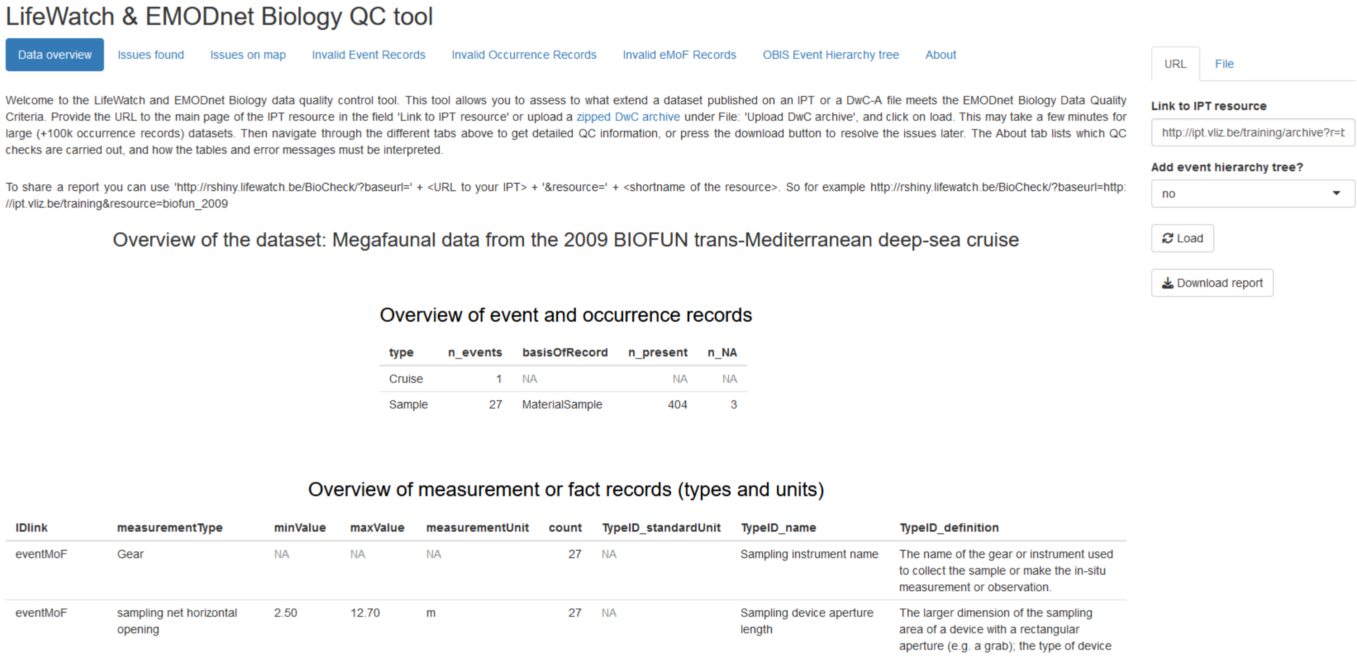


Quality control tools

- There are many different tools you can use to help you perform quality checks on your data. These include:
  - R package obistools
    - Checking required fields, coordinates, depth, outliers, dates, …
  - R package and function Hmisc:: describe
    - Summary statistics
  - LifeWatch & EMODnet Biocheck
  - Lifewatch data services
  - WoRMS taxon match tool
  - GBIF data validator
  - Excel Conditional Formatting tool - identify duplicated data
  - …

The LifeWatch & EMODnet Biocheck tool

- Shiny app: https://rshiny.lifewatch.be/BioCheck/
- R package: https://github.com/EMODnet/EMODnetBiocheck



LifeWatch & EMODnet Biology QC tool

Welcome to the LifeWatch and EMODnet Biology data quality control tool. This tool allows you to assess to what extend a dataset published on an IPT or a DwC-A file meets the EMODnet Biology Data Quality Criteria. Provide the URL to the main page of the IPT resource in the field 'Link to IPT resource' or upload a zipped DwC archive under File: 'Upload DwC archive', and click on load. This may take a few minutes for large (+100k occurrence records) datasets. Then navigate through the different tabs above to get detailed QC information, or press the download button to resolve the issues later. The About tab lists which QC checks are carried out, and how the tables and error messages must be interpreted.

To share a report you can use 'http://rshiny.lifewatch.be/BioCheck/?baseurl=' + <URL to your IPT> + '&resource=' + <shortname of the resource>. So for example http://rshiny.lifewatch.be/BioCheck/?baseurl=http://ipt.vliz.be/training&resource=biofun_2009

Link to IPT resource: http://ipt.vliz.be/training/archive?r=b

Add event hierarchy tree?: no

Overview of the dataset: Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise

Overview of event and occurrence records

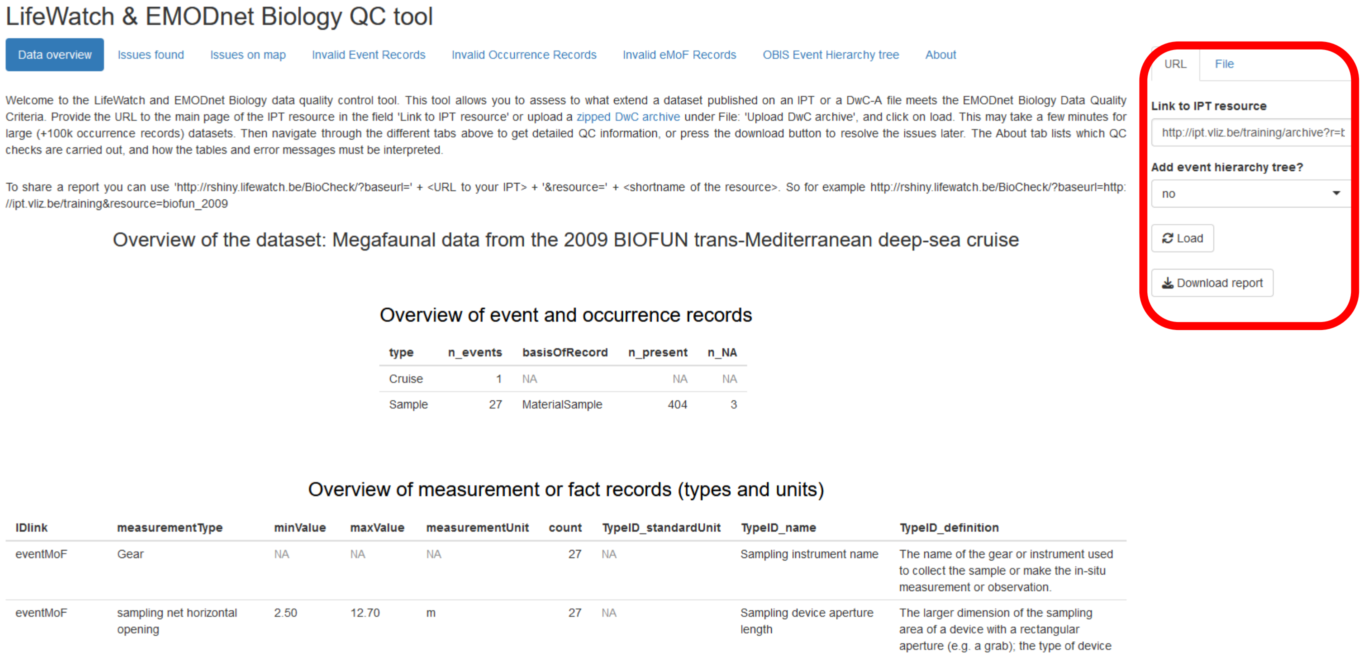
type	n_events	basisOfRecord	n_present	n_NA
Cruise	1	NA	NA	NA
Sample	27	MaterialSample	404	3

Overview of measurement or fact records (types and units)

IDlink	measurementType	minValue	maxValue	measurementUnit	count	TypeID_standardUnit	TypeID_name	TypeID_definition
eventMoF	Gear	NA	NA	NA	27	NA	Sampling instrument name	The name of the gear or instrument used to collect the sample or make the in-situ measurement or observation.
eventMoF	sampling net horizontal opening	2.50	12.70	m	27	NA	Sampling device aperture length	The larger dimension of the sampling area of a device with a rectangular aperture (e.g. a grab); the type of device

The LifeWatch & EMODnet Biocheck tool

- IPT resource URL or Darwin Core Archive file



LifeWatch & EMODnet Biology QC tool

Welcome to the LifeWatch and EMODnet Biology data quality control tool. This tool allows you to assess to what extend a dataset published on an IPT or a DwC-A file meets the EMODnet Biology Data Quality Criteria. Provide the URL to the main page of the IPT resource in the field 'Link to IPT resource' or upload a zipped DwC archive under File: 'Upload DwC archive', and click on load. This may take a few minutes for large (+100k occurrence records) datasets. Then navigate through the different tabs above to get detailed QC information, or press the download button to resolve the issues later. The About tab lists which QC checks are carried out, and how the tables and error messages must be interpreted.

To share a report you can use 'http://rshiny.lifewatch.be/BioCheck/?baseurl=' + <URL to your IPT> + '&resource=' + <shortname of the resource>. So for example http://rshiny.lifewatch.be/BioCheck/?baseurl=http://ipt.vliz.be/training&resource=biofun_2009

Link to IPT resource: http://ipt.vliz.be/training/archive?r=t

Add event hierarchy tree?: no

Overview of the dataset: Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise

Overview of event and occurrence records

type	n_events	basisOfRecord	n_present	n_NA
Cruise	1	NA	NA	NA
Sample	27	MaterialSample	404	3

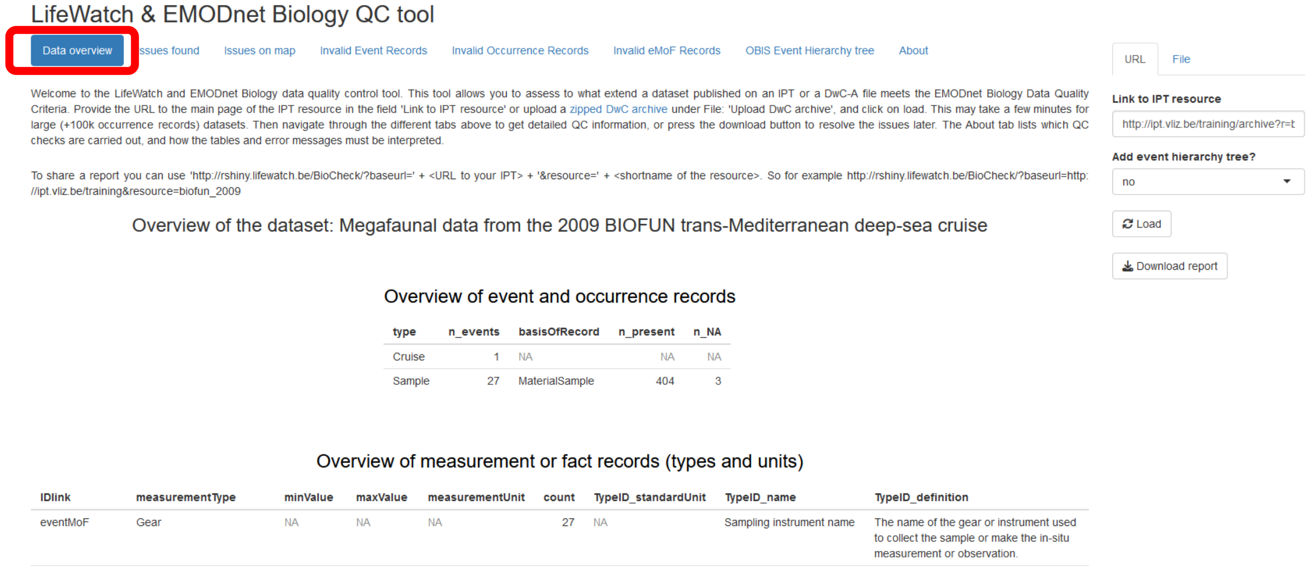
Overview of measurement or fact records (types and units)

IDlink	measurementType	minValue	maxValue	measurementUnit	count	TypeID_standardUnit	TypeID_name	TypeID_definition
eventMoF	Gear	NA	NA	NA	27	NA	Sampling instrument name	The name of the gear or instrument used to collect the sample or make the in-situ measurement or observation.
eventMoF	sampling net horizontal opening	2.50	12.70	m	27	NA	Sampling device aperture length	The larger dimension of the sampling area of a device with a rectangular aperture (e.g. a grab); the type of device

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

The LifeWatch & EMODnet Biocheck tool

≋Explore dataset



LifeWatch & EMODnet Biology QC tool

Welcome to the LifeWatch and EMODnet Biology data quality control tool. This tool allows you to assess to what extend a dataset published on an IPT or a DwC-A file meets the EMODnet Biology Data Quality Criteria. Provide the URL to the main page of the IPT resource in the field 'Link to IPT resource' or upload a zipped DwC archive under File: 'Upload DwC archive', and click on load. This may take a few minutes for large (+100k occurrence records) datasets. Then navigate through the different tabs above to get detailed QC information, or press the download button to resolve the issues later. The About tab lists which QC checks are carried out, and how the tables and error messages must be interpreted.

To share a report you can use 'http://rshiny.lifewatch.be/BioCheck/?baseurl=' + <URL to your IPT> + '&resource=' + <shortname of the resource>. So for example http://rshiny.lifewatch.be/BioCheck/?baseurl=http://ipt.vliz.be/training&resource=biofun_2009

Link to IPT resource: http://ipt.vliz.be/training/archive?r=t

Add event hierarchy tree?: no

Overview of the dataset: Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise

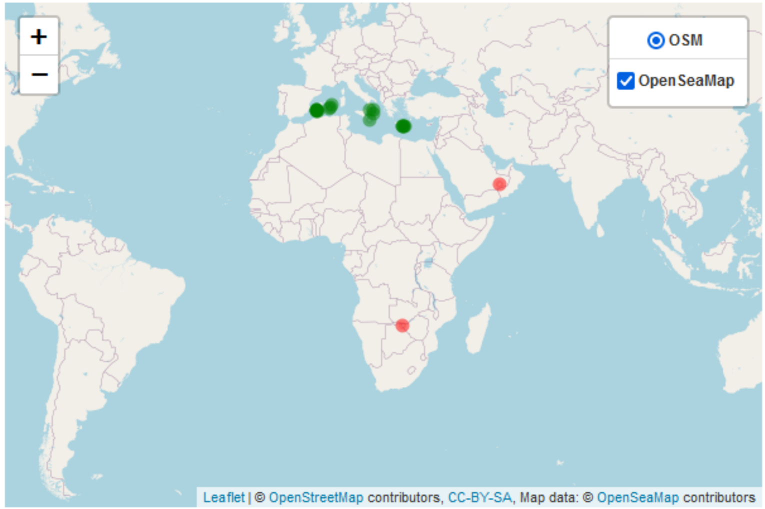
Overview of event and occurrence records

type	n_events	basisOfRecord	n_present	n_NA
Cruise	1	NA	NA	NA
Sample	27	MaterialSample	404	3

Overview of measurement or fact records (types and units)

IDlink	measurementType	minValue	maxValue	measurementUnit	count	TypeID_standardUnit	TypeID_name	TypeID_definition
eventMoF	Gear	NA	NA	NA	27	NA	Sampling instrument name	The name of the gear or instrument used to collect the sample or make the in-situ measurement or observation.
		?.50	12.70	m	27	NA	Sampling device aperture length	The larger dimension of the sampling area of a device with a rectangular aperture (e.g. a grab); the type of device

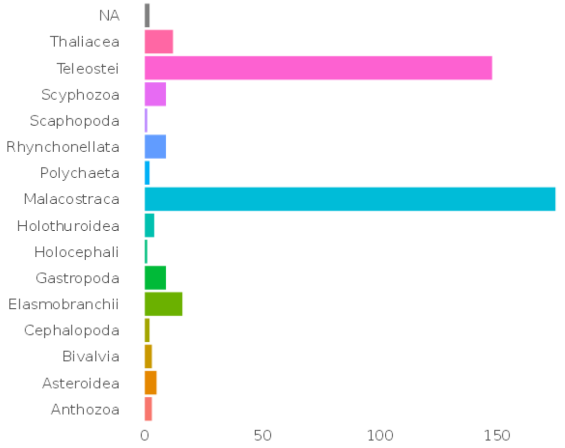
Geographical cover of the dataset



+
−
OSM
OpenSeaMap
Leaflet | © OpenStreetMap contributors, CC-BY-SA, Map data: © OpenSeaMap contributors

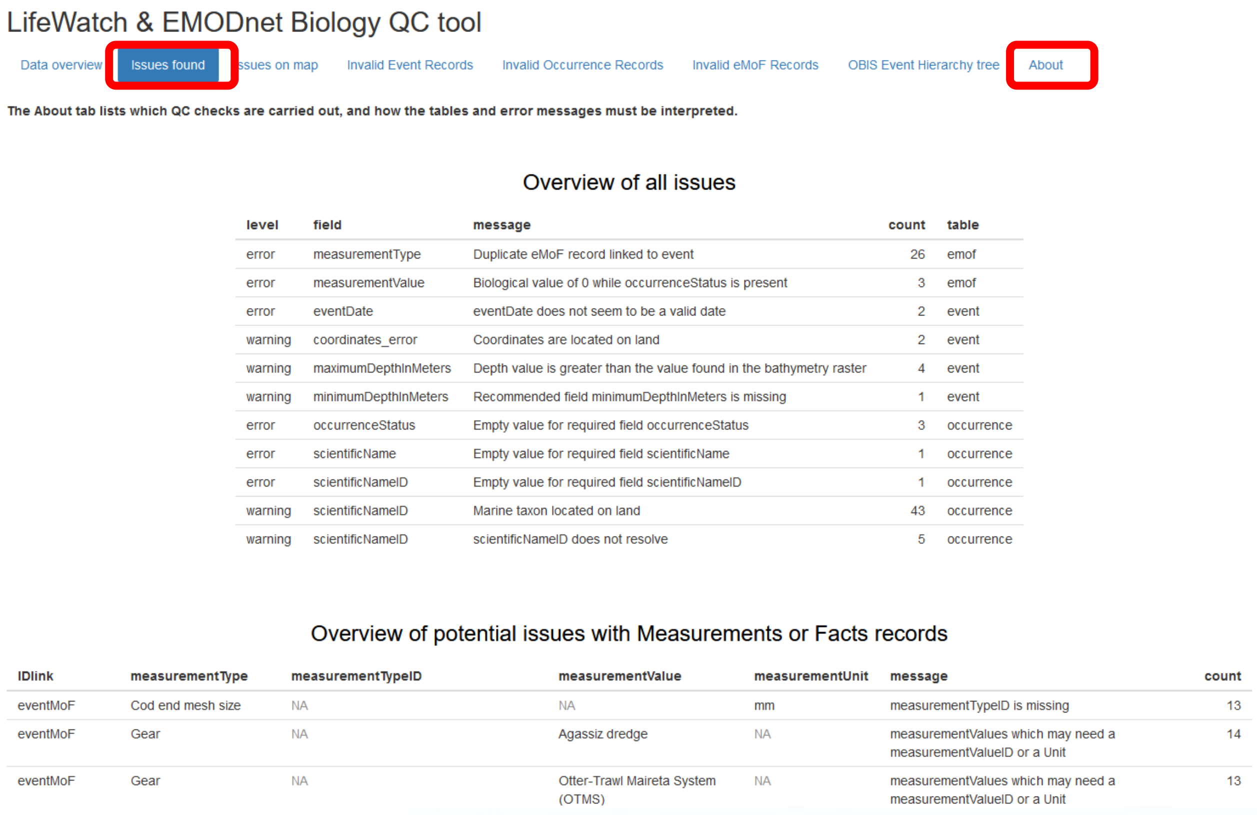
Taxonomic cover of the dataset

Select kingdom: Animalia



The LifeWatch & EMODnet Biocheck tool

- Identify issues
  - List of checks under "About"
- Check and deal with all error and warning messages!



LifeWatch & EMODnet Biology QC tool

The About tab lists which QC checks are carried out, and how the tables and error messages must be interpreted.

Overview of all issues

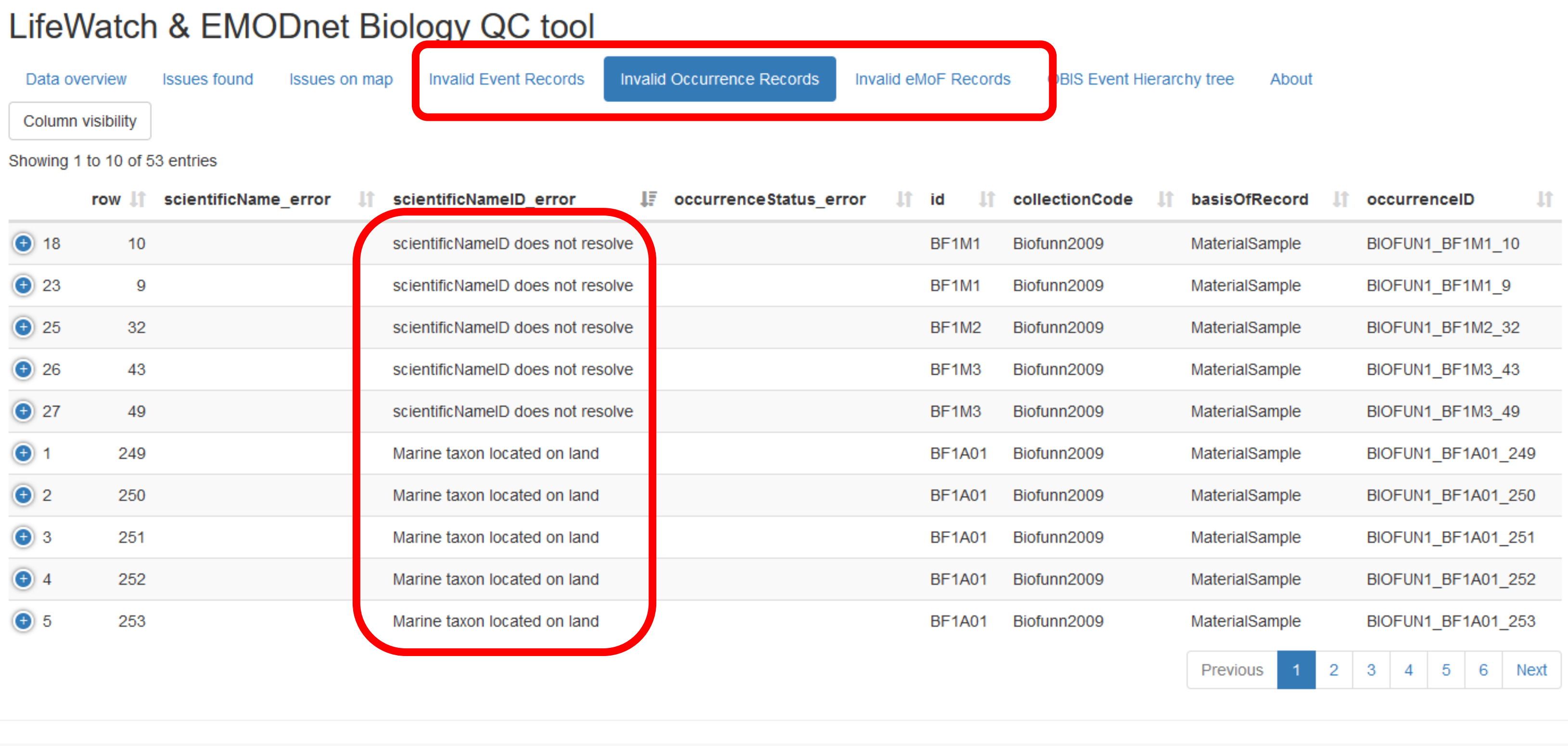
level	field	message	count	table
error	measurementType	Duplicate eMoF record linked to event	26	emof
error	measurementValue	Biological value of 0 while occurrenceStatus is present	3	emof
error	eventDate	eventDate does not seem to be a valid date	2	event
warning	coordinates_error	Coordinates are located on land	2	event
warning	maximumDepthInMeters	Depth value is greater than the value found in the bathymetry raster	4	event
warning	minimumDepthInMeters	Recommended field minimumDepthInMeters is missing	1	event
error	occurrenceStatus	Empty value for required field occurrenceStatus	3	occurrence
error	scientificName	Empty value for required field scientificName	1	occurrence
error	scientificNameID	Empty value for required field scientificNameID	1	occurrence
warning	scientificNameID	Marine taxon located on land	43	occurrence
warning	scientificNameID	scientificNameID does not resolve	5	occurrence

Overview of potential issues with Measurements or Facts records

IDlink	measurementType	measurementTypeID	measurementValue	measurementUnit	message	count
eventMoF	Cod end mesh size	NA	NA	mm	measurementTypeID is missing	13
eventMoF	Gear	NA	Agassiz dredge	NA	measurementValues which may need a measurementValueID or a Unit	14
eventMoF	Gear	NA	Otter-Trawl Maireta System (OTMS)	NA	measurementValues which may need a measurementValueID or a Unit	13

The LifeWatch & EMODnet Biocheck tool

Explore records with issues



LifeWatch & EMODnet Biology QC tool

Showing 1 to 10 of 53 entries

	row	scientificName_error	scientificNameID_error	occurrenceStatus_error	id	collectionCode	basisOfRecord	occurrenceID
18	10		scientificNameID does not resolve		BF1M1	Biofunn2009	MaterialSample	BIOFUN1_BF1M1_10
23	9		scientificNameID does not resolve		BF1M1	Biofunn2009	MaterialSample	BIOFUN1_BF1M1_9
25	32		scientificNameID does not resolve		BF1M2	Biofunn2009	MaterialSample	BIOFUN1_BF1M2_32
26	43		scientificNameID does not resolve		BF1M3	Biofunn2009	MaterialSample	BIOFUN1_BF1M3_43
27	49		scientificNameID does not resolve		BF1M3	Biofunn2009	MaterialSample	BIOFUN1_BF1M3_49
1	249		Marine taxon located on land		BF1A01	Biofunn2009	MaterialSample	BIOFUN1_BF1A01_249
2	250		Marine taxon located on land		BF1A01	Biofunn2009	MaterialSample	BIOFUN1_BF1A01_250
3	251		Marine taxon located on land		BF1A01	Biofunn2009	MaterialSample	BIOFUN1_BF1A01_251
4	252		Marine taxon located on land		BF1A01	Biofunn2009	MaterialSample	BIOFUN1_BF1A01_252
5	253		Marine taxon located on land		BF1A01	Biofunn2009	MaterialSample	BIOFUN1_BF1A01_253

Demo

Biocheck QC tool

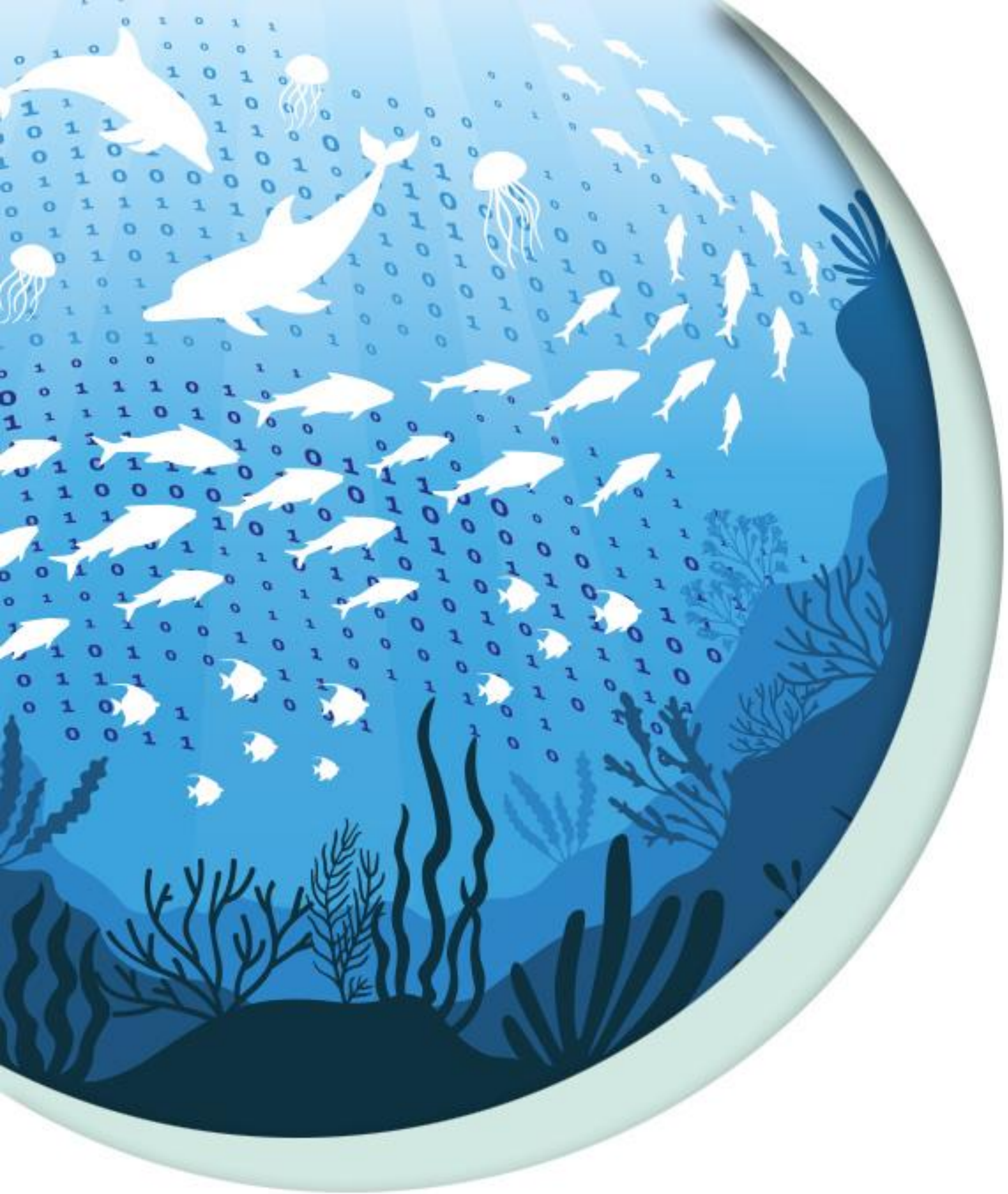
https://rshiny.lifewatch.be/BioCheck/

Relevant sources

- Vandepitte *et al.* (2010). Data integration for European marine biodiversity research: creating a database on benthos and plankton to study large-scale patterns and long-term changes. *Hydrobiologia 644: 1-13*

- The LifeWatch & EMODnet Biocheck tool

- The GBIF data validator



DTO-BioFlow

Integration of biodiversity monitoring data into the Digital Twin Ocean

THANKS!