1st **SeaDataCloud** Training Workshop
Oostende, (20-27 June 2018)

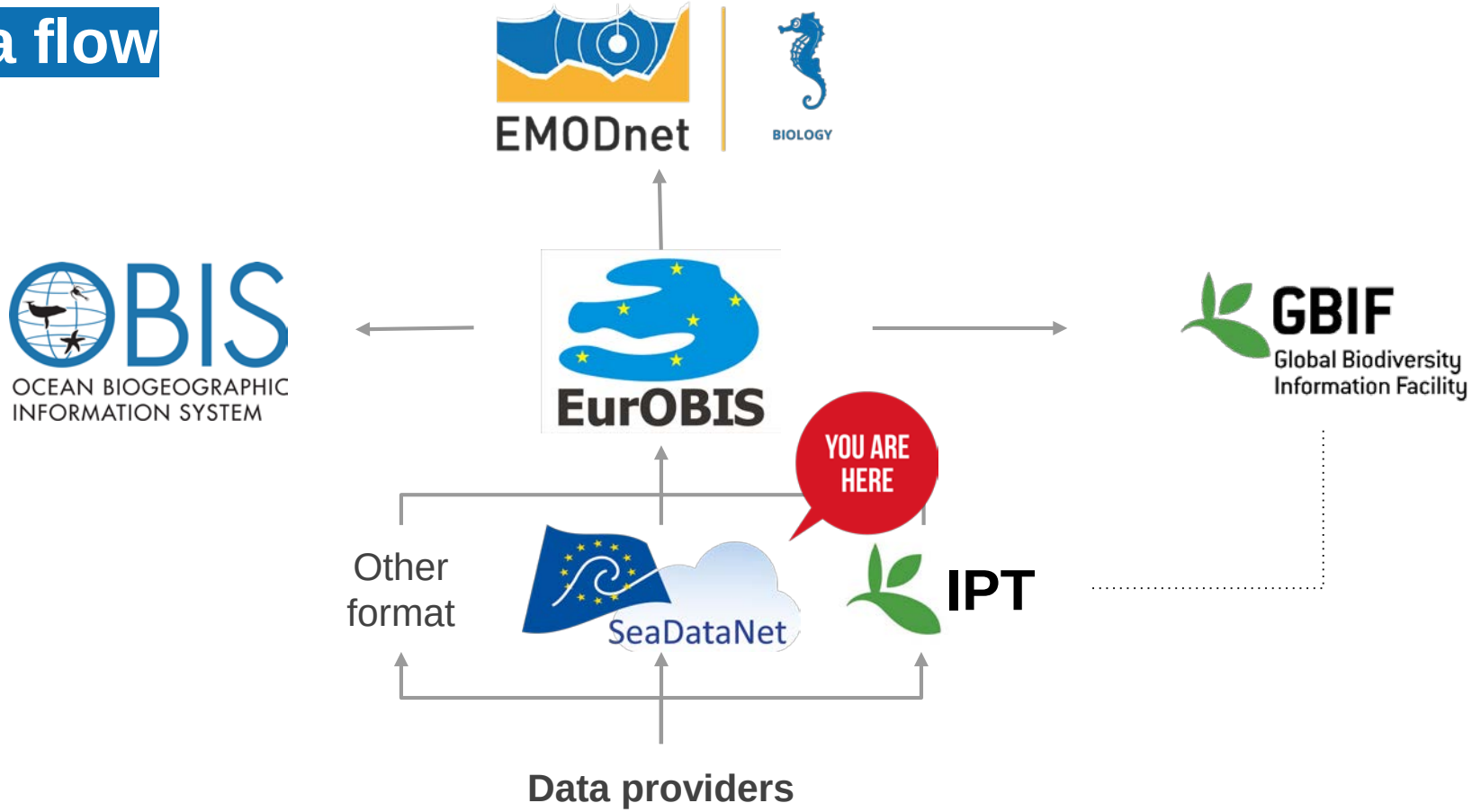
Biological data

Daphnis De Pooter & Paula Oset (VLIZ)

Outline

- EMODnet Biology data flow
- Introduction to biological data
- EMODnet Biology: OBIS Event Core format & Darwin Core
  - OBIS Event Core
  - Darwin Core terms
  - Standards: BODC vocabulary and taxonomy (WoRMS)
  - Exercise: Taxonomy and use of World Register of Marine Species
- SDC and biological data
  - Example 1: Data format: CDI + ODV biology variant
- Biological Data QC
  - Demo using R

Data flow

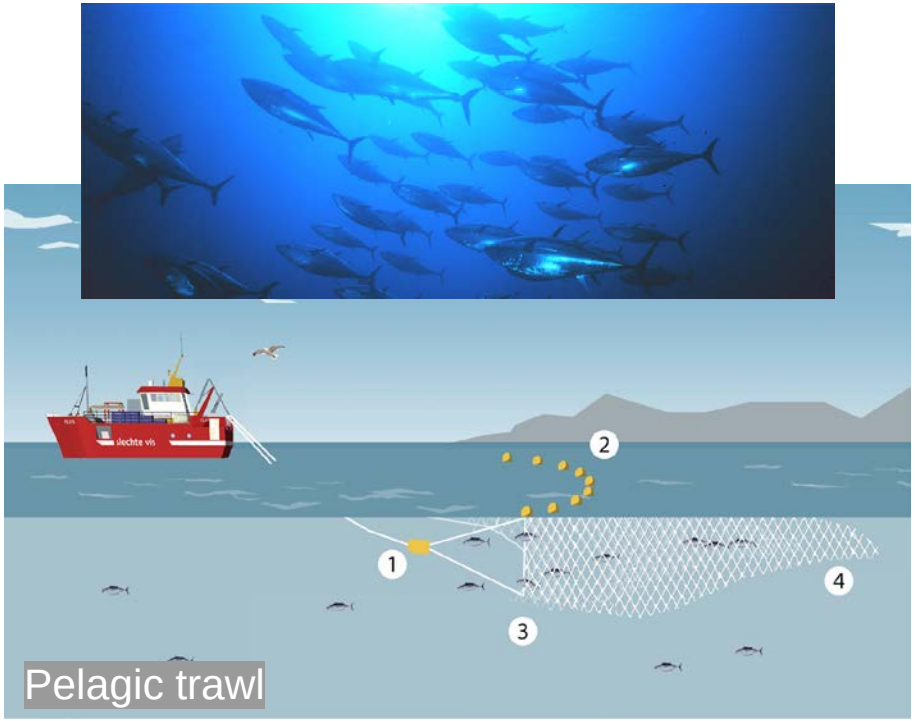

EMODnet
BIOLOGY
OBIS
OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM
EurOBIS
GBIF
Global Biodiversity Information Facility
YOU ARE HERE
Other format
SeaDataNet
IPT
Data providers

Introduction to biological data

Biodiversity

- Biodiversity, the variety of life found in a place on Earth. Common measure: the count of species in an area (species richness).
- Basic information:
  - What
  - Where
  - When
- Additional information:
  - How many/much
  - Under which environmental conditions

Biological sampling:



Biological sampling:

Bottom trawl

Van Veen grab

Biological sampling:

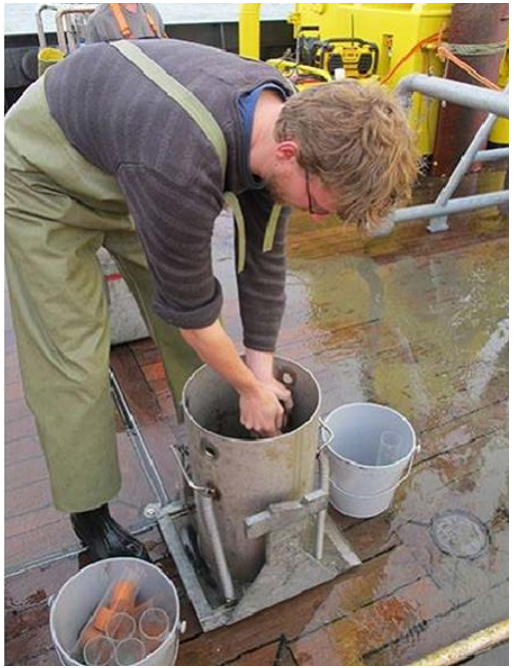
Biological sampling:


Temperature, salinity


Pigments, nutrients

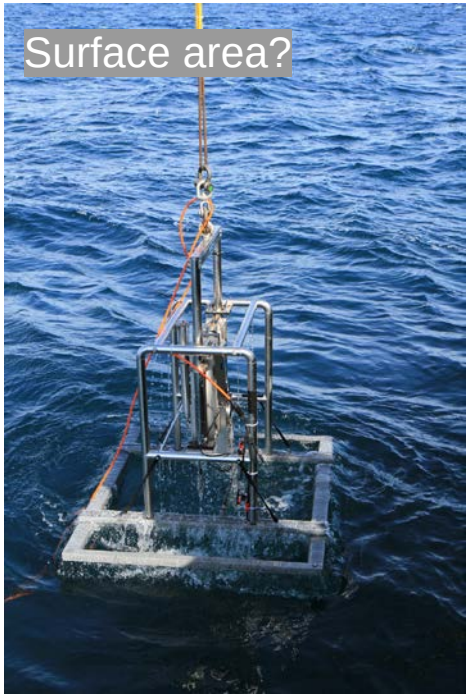

Sediment characteristics

Biological sampling:



Subsamples, slices. replicas...

Biological sampling:



- How to capture these complexities and different types of information?

Overview of biological data format

01 Data schema / structure

- 3 tables (OBIS-Event core data format):
  - Event core table.
  - Occurrence table.
  - Extended Measurements or Facts (eMoF).

02 Field nomenclature

- Darwin Core (DwC) standard terms.
- Minimum of fields required per table.

03 Content - Controlled vocabulary and standards

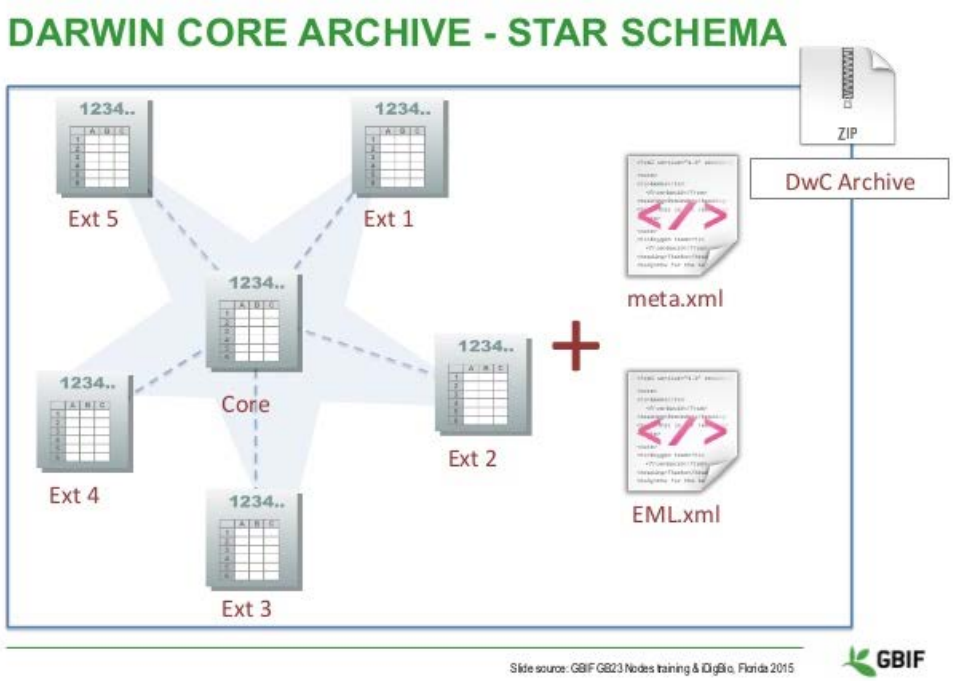
- Date/time & Lat/lon.
- EventID and OccurrenceID.
- Taxonomic information: LSID*
- Other parameters: BODC-NERC vocabulary.

OBIS Event core format

Darwin Core schema

The conceptual data model of the Darwin Core Archive is a “**star schema**” (Robertson et al. 2014):

- **Core record**, such as an occurrence or an event, as the center of the star.
- **Extension records**, radiating out of the star, can optionally be associated with the core, linked by database keys such as an ID column.

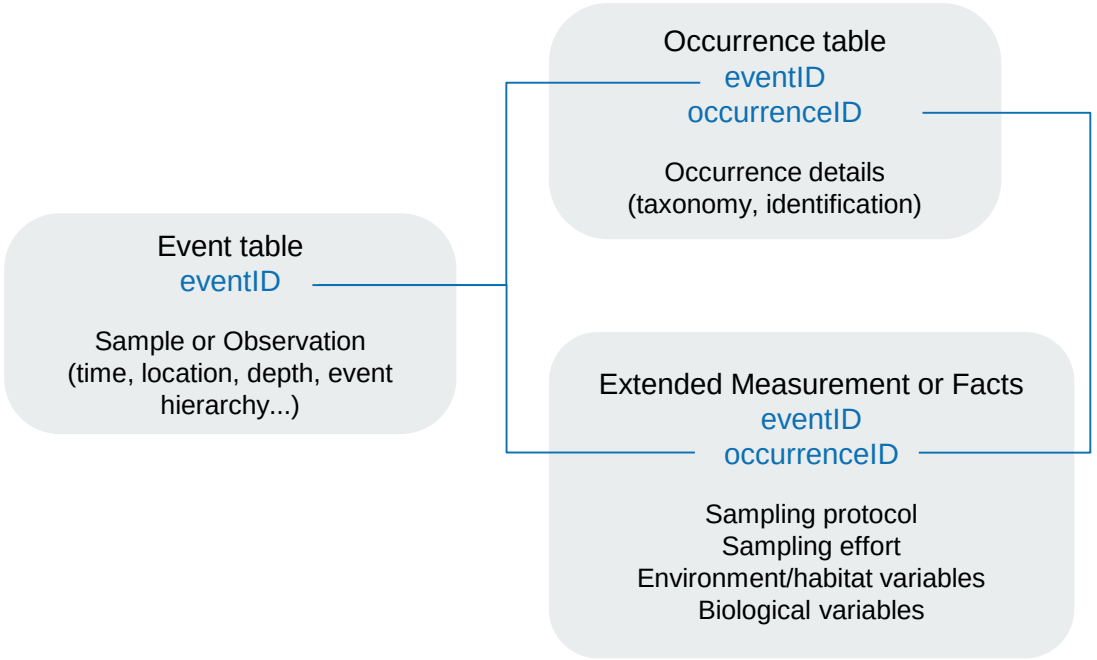


01 Data schema / structure

- 3 tables (OBIS-Event core data format):
  - Event core table.
  - Occurrence table.
  - Extended Measurements or Facts (eMoF).

Data is structured in 3 tables related to each other via the eventID and the occurrenceID.

This structure allows to store not only occurrences but also sampling information and additional biological and/or abiotic measurements.



Event
Core

Cruise

Station

sample

subsample

Field names: Darwin Core terms

02 Field nomenclature

- Darwin Core (DwC) standard terms.
- Minimum of fields required per table.

The DwC terms that are most relevant to EMODnet Biology format are the following (those in **bold** are mandatory):

Event table

datasetName, ***eventID***, parentEventID, ***eventDate***, institutionCode, habitat, type, minimumDepthInMeters, maximumDepthInMeters, ***decimalLatitude***, ***decimalLongitude***, coordinateUncertaintyInMeters, footprintWKT, modified

Occurrence table

eventID, ***occurrenceID***, ***scientificName***, scientificNameAuthorship, ***scientificNameID***, kingdom, taxonRank, identificationQualifier, ***occurrenceStatus***, ***basisOfRecord***, modified

Extended MeasurementorFact table

measurementID, ***eventID***, occurrenceID, measurementType, measurementTypeID, measurementValue, measurementValueID, measurementUnit, measurementUnitID, measurementAccuracy, measurementRemarks

Content: controlled vocabulary and standards

03 Content - Controlled vocabulary and standards

- Date/time & Lat/lon.
- EventID and OccurrenceID.
- Taxonomic information: LSID
- Other parameters: BODC-NERC vocabulary.

Besides the field names, the content or the data itself has to follow certain standards. For example, the date-related fields have to be ISO 8601 compliant, the latitude and longitude have to be in decimal degrees and referenced to the WGS84 projection (EPSG:4326 datum).

An overview of the required format for the content of the different fields is available here.

- EventID and OccurrenceID, coordinates and date (Common terms)
- Taxonomic information: LSID
- Other parameters (eMoF table): BODC-NERC controlled vocabulary.

BODC controlled vocabulary in the eMoF table

Controlled vocabulary (eMoF table)

- The eMoF extension is used to store:
  - information related to sampling method and sampling effort (via eventID).
  - measurements linked to a biological occurrence (via occurrenceID).
  - environmental measurements (via eventID).
- The MoF terms: measurementType, measurementValue and measurementUnit are completely unconstrained and can be populated with free text annotation.
  - Free text: to capture complex and as yet unclassified information
  - But heterogeneity (e.g. of spelling or wording) becomes a major challenge for effective data integration and analysis.

Controlled vocabulary (eMoF table)

- Three fields to standardise the measurement types, values and units: measurementTypeID, measurementValueID and measurementUnitID.
- These terms are populated using controlled vocabularies from the NERC Vocabulary Server, developed by the British Oceanographic Data Centre (BODC)
  - https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/

MeasurementType	MeasurementTypeID
(free text)	(controlled vocabulary)
Body length	http://vocab.nerc.ac.uk/collection/P01/current/OBSINDLX
Length	http://vocab.nerc.ac.uk/collection/P01/current/OBSINDLX
Length (mm)	http://vocab.nerc.ac.uk/collection/P01/current/OBSINDLX
length_in_mm	http://vocab.nerc.ac.uk/collection/P01/current/OBSINDLX
Length of specimen	http://vocab.nerc.ac.uk/collection/P01/current/OBSINDLX

Example of parameter standardization using controlled vocabulary.

Controlled vocabulary (eMoF table)

id	Occurrence ID	Measurement Type	Measurement TypeID	Measurement Value	Measurement Unit	Measurement UnitID
BF1M1	BIOFUN1_BF1M1_1	abundance	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01	26	Individuals	http://vocab.nerc.ac.uk/collection/P06/current/UUUU
BF1M1	BIOFUN1_BF1M1_1	density	http://vocab.nerc.ac.uk/collection/P01/current/SDBIOL02	0.000329114	N/km2	http://vocab.nerc.ac.uk/collection/P06/current/NPKM
BF1M1	BIOFUN1_BF1M1_1	Wet Weight Biomass	http://vocab.nerc.ac.uk/collection/P01/current/SDBIOL05	0.091139241	kg/km2	

Example of measurements linked to a biological occurrence (via occurrenceID)

id	measurementType	measurementTypeID	measurementValue	measurementValueID	measurementAccuracy	measurementUnit
BIOFUN_BF1M15	Trawling speed	http://vocab.nerc.ac.uk/collection/P01/current/APSAZZ01	2.7		0.1	knots
BIOFUN_BF1A01	Gear	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002	Agassiz dredge	http://vocab.nerc.ac.uk/collection/L22/current/TOOL0991/		

Example of information related to sampling method and sampling effort (via eventID)

Useful links

http://iobis.org/manual/emof-bodc/
http://iobis.org/vocab/
https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/
http://seadatanet.maris2.nl/v_bodc_vocab_v2/search.asp?lib=P01
(abundance%entity)

Example dataset fully processed

Taxonomic information:
World Register of Marine Species (WoRMS)

Why a taxonomic standard?

Taxonomy = the academic discipline of defining groups of biological organisms on the basis of shared characteristics and giving names to those groups.

Spelling errors

Actinobacillus actimomycetemcomitans
Actinobacillus actimycetemcomitans
Actinobacillus actinmycetemcomitans
Actinobacillus actinomicetemcomitans
...

Spelling variations (correct)

Agalinus paupercula borealis
Agalinus paupercula var. borealis
Agalinus paupercula var. borealis Pennell
Agalinus paupercula (Gray) Britt. var. borealis Pennell
...

Synonymy

Halichondria panicea Pallas, 1766
Hymeniacidon parfitti Parfitt, 1868
Halichondria paciscens Schmidt, 1875
Menanetia minchini Topsent, 1896
>60 synonyms -> check

Homonymy

Alebion

World Register of Marine Species

taxonomy

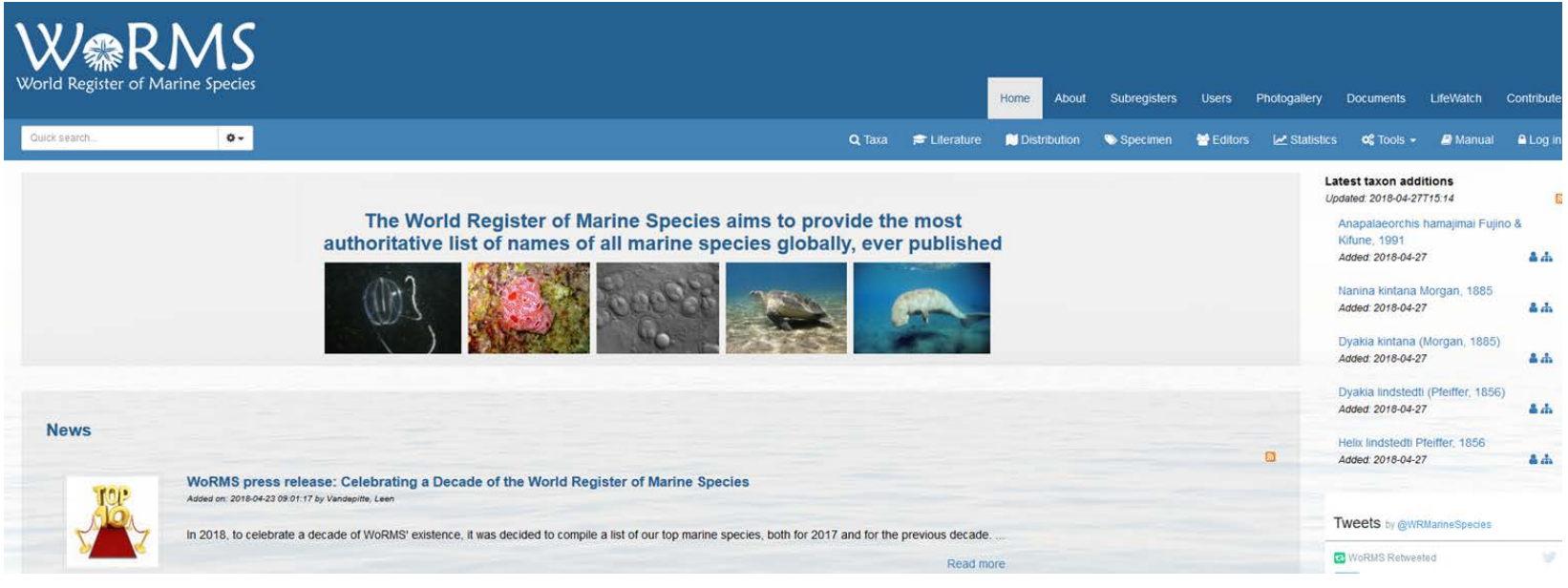
Expert editors

traits

distribution

The World Register of Marine Species (WoRMS) is the taxonomic backbone of EMODnet-Biology and OBIS.

→ All taxa in your dataset need a scientificNameID from WoRMS!



LSID: taxonomic information

All the occurrences are given a unique scientificNameID. This is done by matching the Scientific Names of your occurrence table with the World Register of Marine Species, using the taxon match tool. Information on how to use the taxon match tool here.

After matching, the tool will return you a file with the AphiaIDs, LSIDs, valid names, authorities, classification and any other output you have selected.The WoRMS LSID is used for DwC field scientificNameID.

ScientificName	scientificNameID
Alepocephalus rostratus	urn:lsid:marinespecies.org:taxname:126684
Bathypterois mediterraneus	urn:lsid:marinespecies.org:taxname:299942
Coelorinchus mediterraneus	urn:lsid:marinespecies.org:taxname:280313
Galeus melastomus	urn:lsid:marinespecies.org:taxname:105812
Lepidion lepidion	urn:lsid:marinespecies.org:taxname:126495
Mora moro	urn:lsid:marinespecies.org:taxname:126497

Important: DO NOT provide the "accepted" scientificName and scientificNameID,

Please DO provide the original taxon name and the scientificNameID of the match

Taxon match guided exercise

Download exercise

Go to http://marinespecies.org/aphia.php?p=match

Taxon match guided exercise

Download solution

SDC and Biological Data

- SDC Data to EurOBIS

- Automated transformation should be possible
  - Same terms
  - Similar structure

→ Biological Data Exchange Format (BioDEF)

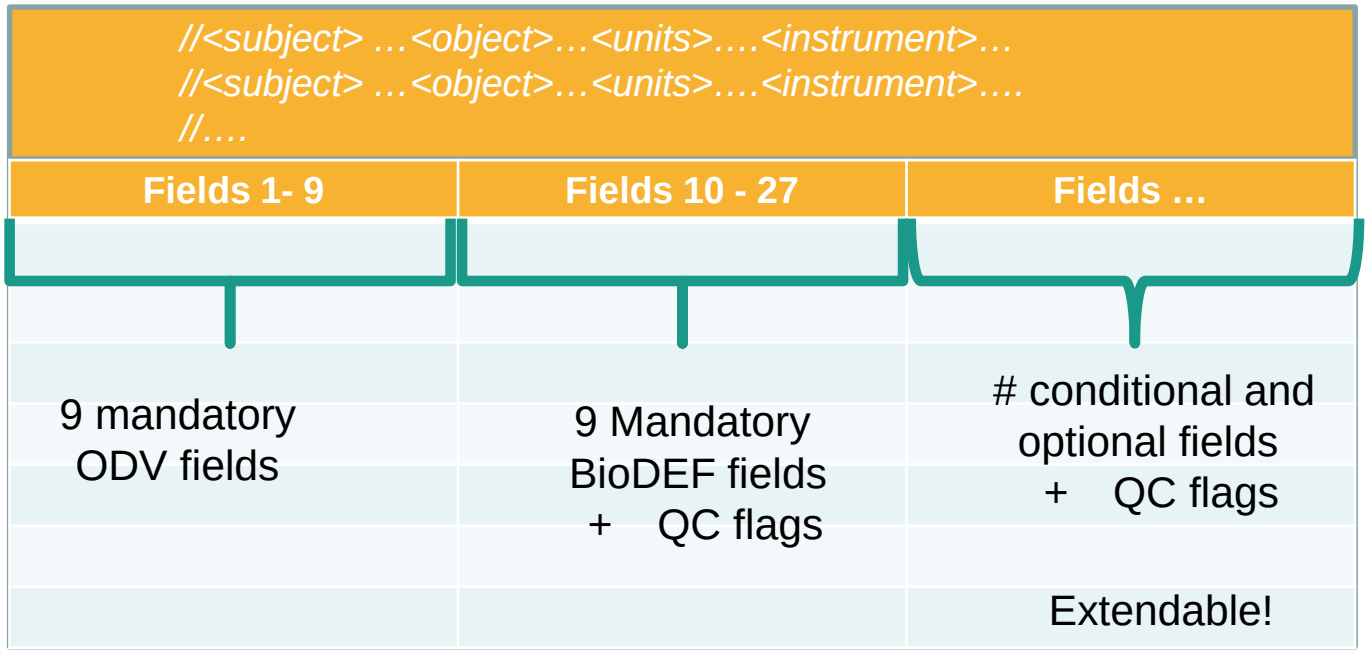
Biological Data Exchange Format (BioDEF)

- CDI metadata
- ODV biology variant version 2.0
  - Semantic header
  - 9 Mandatory ODV fields
  - 9 Mandatory BioDEF fields
  - Conditional BioDEF fields
  - Optional BioDEF fields
  - Quality flags

BioDEF Data Format

HEADER describing fields ←

DATA TABLE ←



BioDEF Data Format

9 Mandatory ODV fields

- Cruise
- Station
- Type
- yyyy-mm-ddThh:mm:ss.sss
- Longitude [degrees_east]
- Latitude [degrees_north]
- LOCAL_CDI_ID
- EDMO_code
- Bot. Depth [m]

BioDEF Data Format

9 Mandatory BioDEF fields

- MinimumDepthOfObservation
- MaximumDepthOfObservation
- SampleID
- SamplingEffort
- ScientificName
- ScientificNameID
- Sex
- LifeStage
- ObservedIndividualCount

4 possible P01 codes

1. Area sampled of the bed
2. Volume sampled of the water body
3. Length of sampling track
4. Sample duration

BioDEF Data Format

Conditional BioDEF fields

- EventStartDateTime
- EventEndDateTime

Time range for sampling

- EventStartLongitude
- EventEndLongitude
- EventStartLatitude
- EventEndLatitude

Sampling along track (e.g. trawling)

- SubsampleID
- SubSamplingCoefficient

In case of subsampling

- Samplingprotocol
- Occurrencestatus

BioDEF Data Format

Optional BioDEF fields

- Abundance per unit area
- Wet weight biomass per unit area
- Coverage
- …..

Biota quantification

- Length of biological entity
- Size class of biological entity
- ...

Biometrics

- Sediment type category
- JNCC habitat type (version 04.03)
- ...

Abiotic data

BioDEF Data Format

Template and examples

A	B	C	D	E	F	G	H	I	J
//<sdn_reference xlink:href="http://seadatanet.maris2.nl/v_cdi_v3/print_xml.asp?edmo=422&identifier=MACROBEL_1351" xlink:role="isDescribedBy" xlink:type="SDN:L23::CDI" sdn:scope="422:MACROBEL_1351"/>									
//<sdn_reference xlink:href="http://seadatanet.maris2.nl/v_cdi_v3/print_xml.asp?edmo=422&identifier=MACROBEL_1206" xlink:role="isDescribedBy" xlink:type="SDN:L23::CDI" sdn:scope="422:MACROBEL_1206"/>									
//SDN_parameter_mapping									
//<subject>SDN:LOCAL:MinimumObservationDepth</subject><object>SDN:P01::MINWDIST</object><units>SDN:P06::ULAA</units>									
//<subject>SDN:LOCAL:MaximumObservationDepth</subject><object>SDN:P01::MAXWDIST</object><units>SDN:P06::ULAA</units>									
//<subject>SDN:LOCAL:SampleID</subject><object>SDN:P01::SAMPID01</object><units>SDN:P06::UUUU</units>									
//<subject>SDN:LOCAL:SamplingEffort</subject><object>SDN:P01::AREABEDS</object><units>SDN:P06::UUUU</units>									
//<subject>SDN:LOCAL:SubsampleID</subject><object>SDN:P01::SSAMID01</object><units>SDN:P06::UUUU</units>									
//<subject>SDN:LOCAL:SubSamplingCoefficient</subject><object>SDN:P01::SSAMPC01</object><units>SDN:P06::UUUU</units>									
//<subject>SDN:LOCAL:ScientificName</subject><object>SDN:P01::SCNAME01</object><units>SDN:P06::UUUU</units>									
//<subject>SDN:LOCAL:ScientificNameID</subject><object>SDN:P01::SNANID01</object><units>SDN:P06::UUUU</units>									
//<subject>SDN:LOCAL:Sex</subject><object>SDN:P01::ENTSEX01</object><units>SDN:P06::UUUU</units>									
//<subject>SDN:LOCAL:LifeStage</subject><object>SDN:P01::LSTAGE01</object><units>SDN:P06::UUUU</units>									
//<subject>SDN:LOCAL:ObservedIndividualCount</subject><object>SDN:P01::OCOUNT01</object><units>SDN:P06::UUUU</units><instrument>SDN:L22::TOOL0653</instrument>									
//<subject>SDN:LOCAL:DensityPerUnitEffort</subject><object>SDN:P01::SDBIOL02</object><units>SDN:P06::UPMS</units>									
//<subject>SDN:LOCAL:WWBiomass</subject><object>SDN:P01::OWETBM01</object><units>SDN:P06::UGRM</units>									
//<subject>SDN:LOCAL:AFDWBiomassPerUnitEffort</subject><object>SDN:P01::SDBIOL03</object><units>SDN:P06::UGMS</units>									
//									
Cruise	**Station**	**Type**	**YYYY-MM-DDThh:mm:ss**	**Longitude [degrees_east]**	**Latitude [degrees_north]**	**LOCAL_CDI_ID**	**EDMO_code**	**Bot. Depth [m]**	**MinimumObservationDepth [m]**
RV Mechelen/197606	330	*	1976-06-05T06:02:13	2.92361	51.24411	MACROBEL_1351	422	25	21.5
									21.5
									21.5
									21.5
									21.5
									21.5
RV Mechelen/197606	ZG02	*	1976-07-05T07:04:10	3.21113	51.35106	MACROBEL_1206	422	28	23.1
									23.1

QC Procedures Biological data

To Verify

- Are all mandatory fields **present**?
- Are all values of mandatory fields **filled out**?
- Are all values in correct **format**?
- Are all values **possible**?

On land?

Depths possible?

...

QC flag → fitness for purpose

QC Procedures Biological data - demo

Download R script

R-package documentation:
https://github.com/iobis/obistools