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

OBIS

EurOBIS

YOU ARE HERE

Other format

SeaDataNet

IPT

GBIF

Global Biodiversity Information Facility

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:

1. Pelagic trawl
2. Water column
3. Pelagic organisms
4. Zooplankton net
Biological sampling:

Bottom trawl

Van Veen grab
Biological sampling:

Lifestage

Biometrics

Biometrics

Biomass

Abundance
Biological sampling:

Temperature, salinity
Pigments, nutrients
Sediment characteristics
Biological sampling:

Subsamples, slices. replicas...
Biological sampling:

Mesh size?

Trawling duration, speed?

Surface area?
● 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
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.
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 table**
  - eventID
  - Sample or Observation (time, location, depth, event hierarchy...)

- **Occurrence table**
  - eventID
  - occurrenceID
  - Occurrence details (taxonomy, identification)

- **Extended Measurement or Facts (eMoF)**
  - eventID
  - occurrenceID
  - Sampling protocol
  - Sampling effort
  - Environment/habitat variables
  - Biological variables

- **3 tables** (OBIS-Event core data format):
  - Event core table.
  - Occurrence table.
  - Extended Measurements or Facts (eMoF).
Field names: Darwin Core terms
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
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
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/](https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/)

<table>
<thead>
<tr>
<th>MeasurementType</th>
<th>MeasurementTypeID</th>
</tr>
</thead>
<tbody>
<tr>
<td>(free text)</td>
<td>(controlled vocabulary)</td>
</tr>
<tr>
<td>Body length</td>
<td><a href="http://vocab.nerc.ac.uk/collection/P01/current/OBSINDLX">http://vocab.nerc.ac.uk/collection/P01/current/OBSINDLX</a></td>
</tr>
<tr>
<td>Length</td>
<td><a href="http://vocab.nerc.ac.uk/collection/P01/current/OBSINDLX">http://vocab.nerc.ac.uk/collection/P01/current/OBSINDLX</a></td>
</tr>
<tr>
<td>Length (mm)</td>
<td><a href="http://vocab.nerc.ac.uk/collection/P01/current/OBSINDLX">http://vocab.nerc.ac.uk/collection/P01/current/OBSINDLX</a></td>
</tr>
<tr>
<td>length_in_mm</td>
<td><a href="http://vocab.nerc.ac.uk/collection/P01/current/OBSINDLX">http://vocab.nerc.ac.uk/collection/P01/current/OBSINDLX</a></td>
</tr>
<tr>
<td>Length of specimen</td>
<td><a href="http://vocab.nerc.ac.uk/collection/P01/current/OBSINDLX">http://vocab.nerc.ac.uk/collection/P01/current/OBSINDLX</a></td>
</tr>
</tbody>
</table>

Example of parameter standardization using controlled vocabulary.
## Controlled vocabulary (eMoF table)

<table>
<thead>
<tr>
<th>id</th>
<th>Occurrence ID</th>
<th>Measurement Type</th>
<th>Measurement TypeID</th>
<th>Measurement Value</th>
<th>Measurement Unit</th>
<th>Measurement UnitID</th>
</tr>
</thead>
<tbody>
<tr>
<td>BF1M1</td>
<td>BIOFUN1_BF1M1_1</td>
<td>abundance</td>
<td><a href="http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01">http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01</a></td>
<td>26</td>
<td>Individuals</td>
<td><a href="http://vocab.nerc.ac.uk/collection/P06/current/UUUU">http://vocab.nerc.ac.uk/collection/P06/current/UUUU</a></td>
</tr>
<tr>
<td>BF1M1</td>
<td>BIOFUN1_BF1M1_1</td>
<td>density</td>
<td><a href="http://vocab.nerc.ac.uk/collection/P01/current/SDBIOL02">http://vocab.nerc.ac.uk/collection/P01/current/SDBIOL02</a></td>
<td>0.000329114</td>
<td>N/km²</td>
<td><a href="http://vocab.nerc.ac.uk/collection/P06/current/NPKM">http://vocab.nerc.ac.uk/collection/P06/current/NPKM</a></td>
</tr>
<tr>
<td>BF1M1</td>
<td>BIOFUN1_BF1M1_1</td>
<td>Wet Weight Biomass</td>
<td><a href="http://vocab.nerc.ac.uk/collection/P01/current/SDBIOL05">http://vocab.nerc.ac.uk/collection/P01/current/SDBIOL05</a></td>
<td>0.091139241</td>
<td>kg/km²</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>id</th>
<th>measurementType</th>
<th>measurementTypeID</th>
<th>measurementValue</th>
<th>measurementValueID</th>
<th>measurementAccuracy</th>
<th>measurementUnit</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIOFUN_BF1M1_1</td>
<td>Trawling speed</td>
<td><a href="http://vocab.nerc.ac.uk/collection/P01/current/APSAZZ01">http://vocab.nerc.ac.uk/collection/P01/current/APSAZZ01</a></td>
<td>2.7</td>
<td>2.7</td>
<td>0.1</td>
<td>knots</td>
</tr>
<tr>
<td>BIOFUN_BF1A0_1</td>
<td>Gear</td>
<td><a href="http://vocab.nerc.ac.uk/collection/Q01/current/Q01000002">http://vocab.nerc.ac.uk/collection/Q01/current/Q01000002</a></td>
<td>Agassiz dredge</td>
<td><a href="http://vocab.nerc.ac.uk/collection/L22/current/TOOL0991/">http://vocab.nerc.ac.uk/collection/L22/current/TOOL0991/</a></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 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 actimomycetemcomitans
Actinobacillus actinmucetemcomitans
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

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 AphialIDs, LSIDs, valid names, authorities, classification and any other output you have selected. The WoRMS LSID is used for DwC field scientificNameID.

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

<table>
<thead>
<tr>
<th>Fields 1-9</th>
<th>Fields 10 - 27</th>
<th>Fields ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>9 mandatory ODV fields</td>
<td>9 Mandatory BioDEF fields</td>
<td># conditional and optional fields</td>
</tr>
<tr>
<td>+ QC flags</td>
<td>+ QC flags</td>
<td>+ QC flags</td>
</tr>
</tbody>
</table>

Extendable!
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
- EventStartLongitude
- EventEndLongitude
- EventStartLatitude
- EventEndLatitude
- SubsampleID
- SubSamplingCoefficient
- Samplingprotocol
- Occurrencestatus

Time range for sampling
Sampling along track (e.g. trawling)
In case of subsampling
BioDEF Data Format

Optional BioDEF fields

- Abundance per unit area
- Wet weight biomass per unit area
- Coverage
- .....  
- Length of biological entity
- Size class of biological entity
- ...
- Sediment type category
- JNCC habitat type (version 04.03)
- ...

Biota quantification

Biometrics

Abiotic data
## BioDEF Data Format

### Template and examples

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RV Mechelen/197606</td>
<td>330</td>
<td>*</td>
<td>1976-06-05T06:02:13</td>
<td>2.92361</td>
<td>51.24411</td>
<td>MACROBEL_1351</td>
<td>422</td>
<td>25</td>
<td>21.5</td>
</tr>
<tr>
<td>RV Mechelen/197606</td>
<td>2602</td>
<td>*</td>
<td>1976-07-05T07:04:10</td>
<td>3.21113</td>
<td>51.35106</td>
<td>MACROBEL_1206</td>
<td>422</td>
<td>28</td>
<td>23.1</td>
</tr>
</tbody>
</table>
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?  QC flag → fitness for purpose
Depths possible?
QC Procedures Biological data - demo

Download R script

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