SeaDataNet

PAN-EUROPEAN INFRASTRUCTURE FOR OCEAN & MARINE DATA MANAGEMENT

Status of the biological QC developments

Simon Claus, Daphnis De Pootere & Filip Waumans

VLIZ- Flanders Marine Institute

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) taxonomic quality
4) geographic quality
5) whether or not the record is a (possible) outlier.

D10.9: Specification of Biology Data QC online and development plan M12=>M18

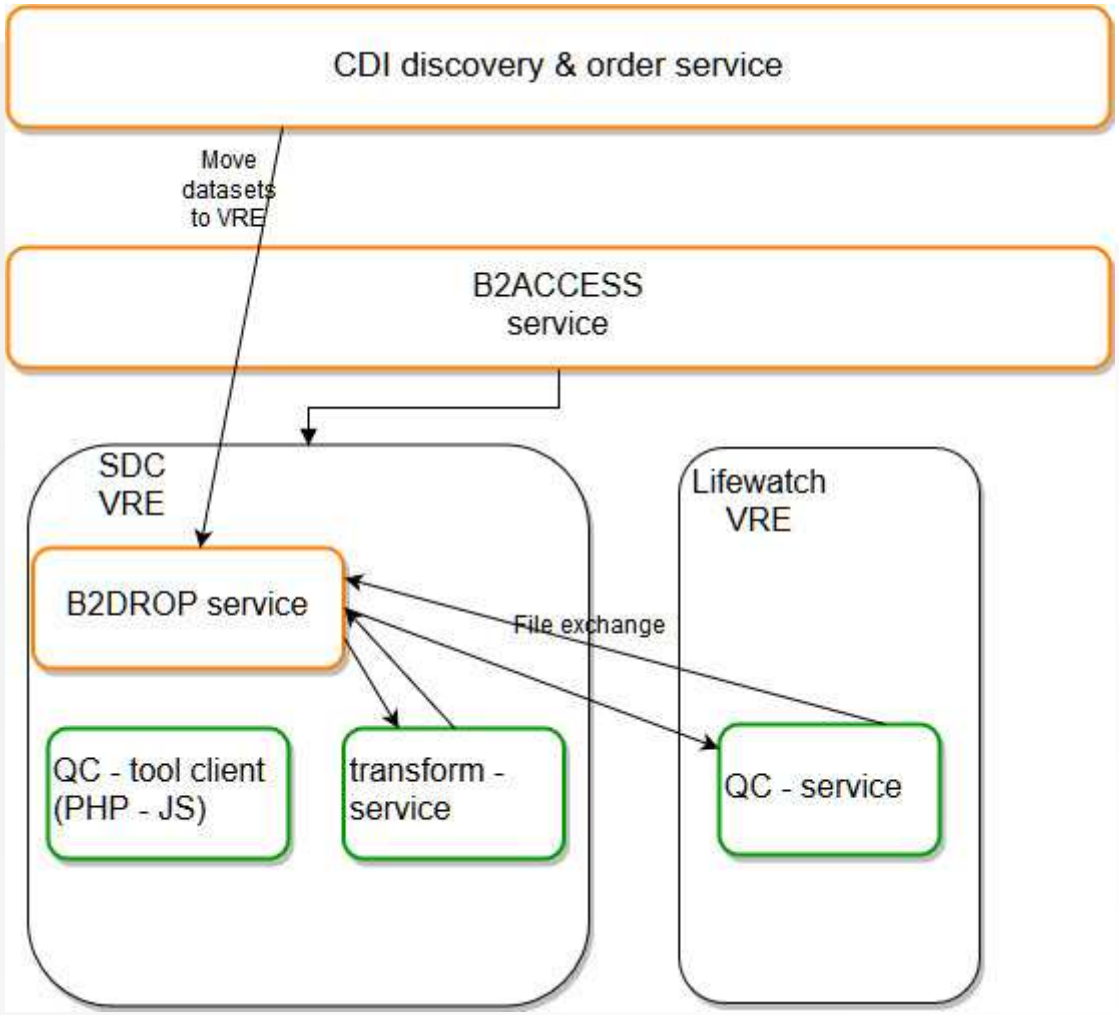
D10.10: Phase 1 of Biology Data QC online operational (M24)

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

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

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

- The development of the **transformation service** in a Docker container which will convert the BIODEF ODV format in DwC-A format and vice versa
- The implementation of the **QC steps** mentioned in point 3 of the use case description
- Tuning with the overall architecture and VRE development.


CDI discovery & order service
Move datasets to VRE
B2ACCESS service
SDC VRE
B2DROP service
QC - tool client (PHP - JS)
transform - service
Lifewatch VRE
File exchange
QC - service

http://marinespecies.org/rest/AphiaIDByName/abra%20alba?marine_only=true
http://marinespecies.org/rest/AphiaIDByName/abra%20alba?marine_only=true

4. Transformation service

SPECIFIC DOCUMENTATION FOR **BIOLOGICAL DATA**

DESCRIPTION OF THE SEADATANET DATA TRANSPORT FORMAT FOR BIOLOGICAL DATA

DOWNLOAD (991.96 KB)

TEMPLATE AND EXAMPLES OF BIOLOGICAL DATA FILES (ZIP FILE)

DOWNLOAD (550.84 KB)

Darwin Core Archive

OBIS-ENV-DATA

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

anche.R | EMODnet_allservices.R | bd | bod | prepdata.Rmd | EMODnet_biology.R | dome | dome1 | BaltizZPprepData.Rmd

Source on Save

```
36 close(con)
37
38 ##read datafile in dataframe variable skipping commentlines
39 data<-read.delim("data_from_collection_MinimumObservationDepth.txt",
40                  skip = commentlines, colClasses = "character", na.strings ='' )
41
42 #-------------------------------------------------------------------------#
43 ####                    map header with DwC terms                      ####
44 #-------------------------------------------------------------------------#
45 # map sdc --> dwc
46 map_sdc_terms <- c(LocationID= "Station name",
47                    eventDate= "yyyy.mm.ddThh.mm.ss.sss",
48                    Locality= "Station",
49                    decimalLongitude= "Longitude..degrees_east.",
50                    decimalLatitude= "Latitude..degrees_north.",
51                    minimumdepthinmeters= "minimumdepth",
52                    maximumdepthinmeters= "maximumdept",
53                    Samplingprotocol= "Samplingprotocol..#.")
54
55 dwc_terms <- c("eventDate",
56                "decimalLongitude",
57                "decimalLatitude",
58                "scientificName",
59                "scientificNameID",
60                "occurrenceStatus",
61                "basisOfRecord",
62                "minimumDepthInMeters",
63                "maximumDepthInMeters")
64
```

http://marinespecies.org/rest/AphiaIDByName/abra%20alba?marine_only=true
http://marinespecies.org/rest/AphiaIDByName/abra%20alba?marine_only=true

- Is the dataset integrety ok?
  - Do all eventID's in the occurrence extention refer to an Event Record?
  - Do all eventID's in the eMoF extention refer to an Event Record?
  - Do all occurrenceID's in the eMoF extention refer to an occurrence Record?
  - In case of a biometrical parameter, is the eventID in the eMoF link the same one as in the occurrence extension?
  - Are there 'duplicate occurrences' meaning is the same taxon listed twice at the same EventID without any diference in any of the biometric paramters?
  - Are there 'duplicate measurements' meaning does the same measurement occur twice for the same occurrenceID or the same EventID?
- Are all mandatory fields present in the dataset?
- Are all mandatory field filled out?
- Does eventDate follow the required format?
- Does the scientificNameID follow the required format?
- Are there coordinates located on land? (buffer of 3km is taken under consideration)
- Are there depths at the location deeper than the depths stored by GEBCO? (a margin of 150m is taken under consideration)
- Do all measurementtypes have a MeasurementTypeID?
- Does the measurementTypeID / measurementValueID refer to an existing term in the BODC vocabulary?
- Do all measurementValues that refer to facts have a measurementValueID?
- Are there records where measurementValue is NULL?
- Are there records that refer to biological measurement where measurementValue = 0 (and where occurrenceStatus is not absent)?

http://marinespecies.org/rest/AphiaIDByName/abra%20alba?marine_only=true
http://marinespecies.org/rest/AphiaIDByName/abra%20alba?marine_only=true

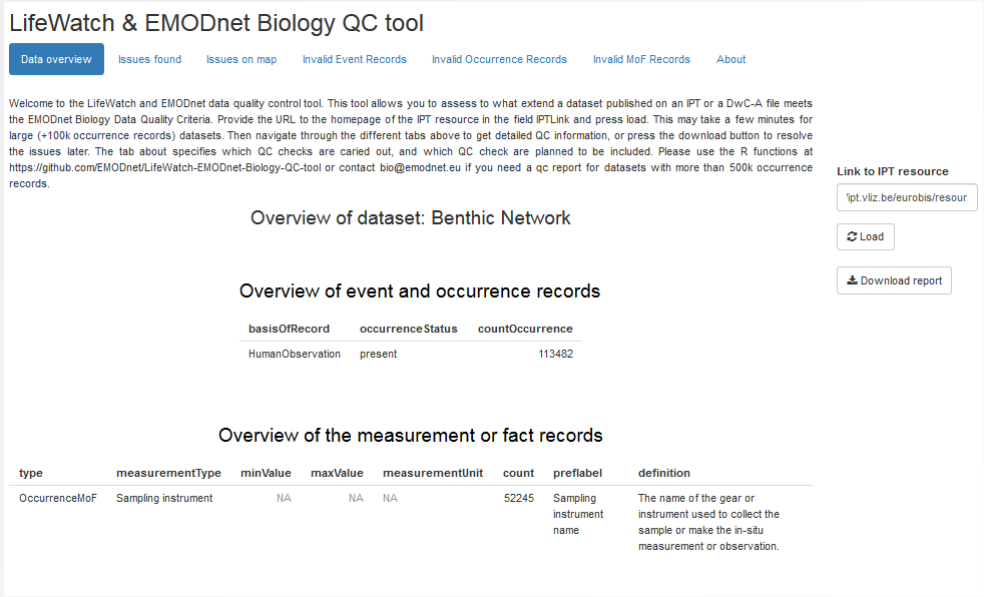
Additional checks which are planned to be implemented are:

- Is there a sampling instrument present?
- Are there other sampling descriptors present?
- Provide an overview of the number of taxa per kingdom and class.
- Check for non-marine and non-brackish taxa.
- Has the unit provided the same base unit as specified by the measurementTypeID?
- add the name of the dataset + URL at the top + number of records per file
- List the non-matched taxa
- Add geographical cover and temporal cover from the eml to the overvieuw page.

http://marinespecies.org/rest/AphiaIDByName/abra%20alba?marine_only=true
http://marinespecies.org/rest/AphiaIDByName/abra%20alba?marine_only=true

QC service

- Tool developed in R, all scripts are available from the EMODnet github repository at https://github.com/EMODnet/LifeWatch-EMODnet-Biology-QC-tool.
- Is using the OBIS tools package https://github.com/iobis/obistools
- Available from the LifeWatch services at http://rshiny.lifewatch.be/BioCheck/



http://marinespecies.org/rest/AphiaIDByName/abra%20alba?marine_only=true
http://marinespecies.org/rest/AphiaIDByName/abra%20alba?marine_only=true

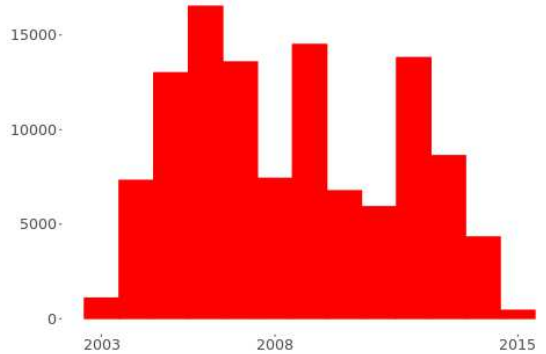
QC service

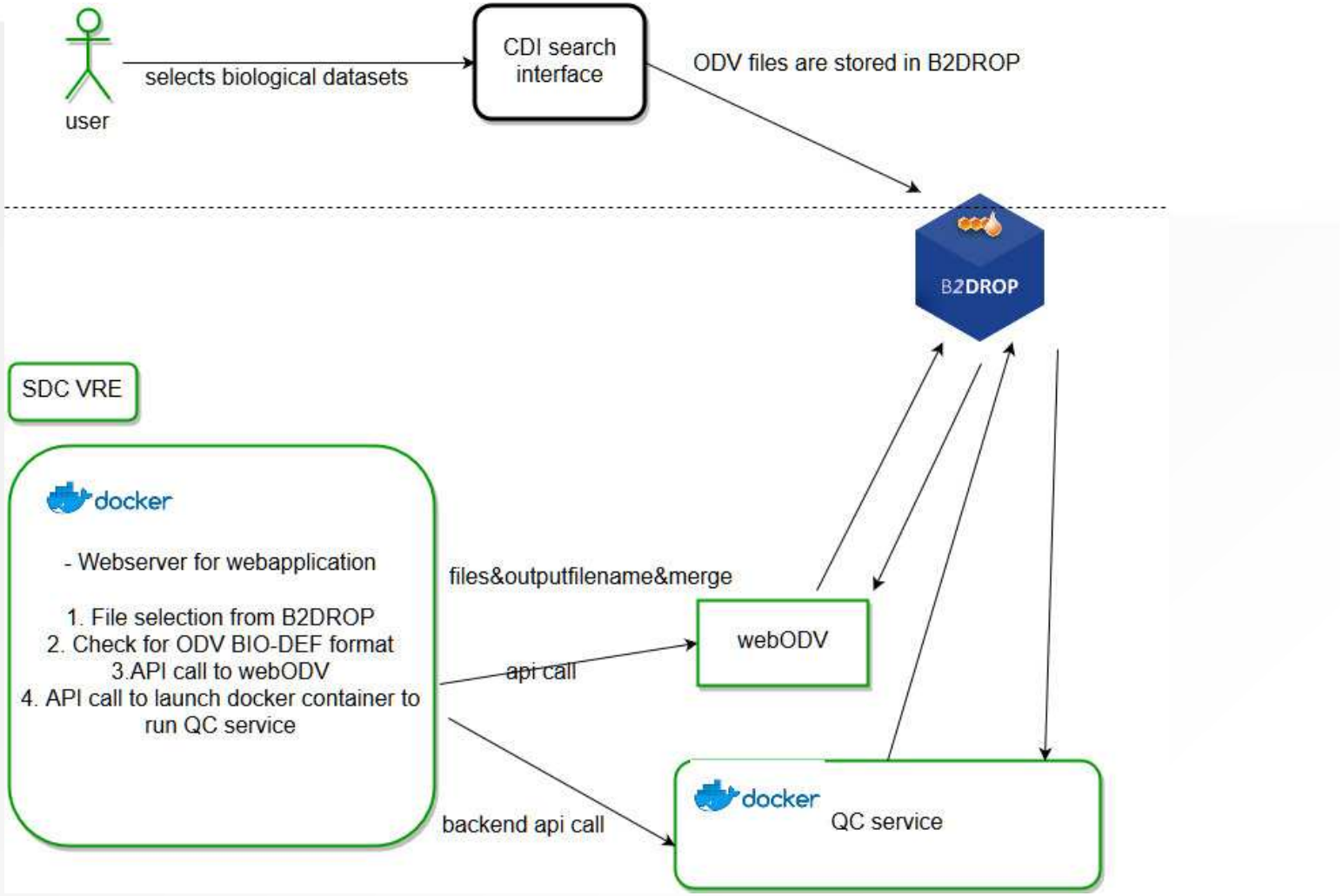
	row	maximumDepthInMeters_error	minimumDepthInMeters_error
1	1	Empty value for recommended field maximumDepthInMeters	Empty value for recommended field minimumDepthInMeters
2	2	Empty value for recommended field maximumDepthInMeters	Empty value for recommended field minimumDepthInMeters
3	3	Empty value for recommended field maximumDepthInMeters	Empty value for recommended field minimumDepthInMeters
4	4	Empty value for recommended field maximumDepthInMeters	Empty value for recommended field minimumDepthInMeters
5	5	Empty value for recommended field maximumDepthInMeters	Empty value for recommended field minimumDepthInMeters
6	6	Empty value for recommended field maximumDepthInMeters	Empty value for recommended field minimumDepthInMeters
7	7	Empty value for recommended field maximumDepthInMeters	Empty value for recommended field minimumDepthInMeters
8	8	Empty value for recommended field maximumDepthInMeters	Empty value for recommended field minimumDepthInMeters
9	9	Empty value for recommended field maximumDepthInMeters	Empty value for recommended field minimumDepthInMeters
10	10	Empty value for recommended field maximumDepthInMeters	Empty value for recommended field minimumDepthInMeters

Previous 1 2 3 4 5 ... 10687 Next

Geographical cover of the dataset

Temporal cover of the dataset




user
selects biological datasets
CDI search interface
ODV files are stored in B2DROP
B2DROP
SDC VRE
docker
- Webserver for webapplication
1. File selection from B2DROP
2. Check for ODV BIO-DEF format
3.API call to webODV
4. API call to launch docker container to run QC service
files&outputfilename&merge
webODV
api call
backend api call
docker
QC service


Thank you for your attention!

http://marinespecies.org/rest/AphiaIDByName/abra%20alba?marine_only=true

The different components are:

1. CDI discovery service

This service allows users to select BIO-DEF ODV files based via the CDI system.

2. B2ACCESS/ Marine ID service

These service will be used for authorization and authentication to the VRE. The user logs in via his/her Marine ID account, that maps to B2ACCESS by which the SDC VRE components will be secured.

3. File repository

The B2DROP environment will be used as a shared file repository. In the SDC VRE the user collects the biological datasets in the personal space. Results of transformations and QC are collected again in the same space.

http://marinespecies.org/rest/AphiaIDByName/abra%20alba?marine_only=true
http://marinespecies.org/rest/AphiaIDByName/abra%20alba?marine_only=true

The different components are:

4. Transformation service

Biodiversity services such as the QC service work preferable with Darwin Core standards. This transformation service running within the SDC VRE platform (in a Docker container like the other transformation and processing services on the EUDAT platform) will convert the BIODEF ODV files into Darwin Core, ready to serve the files later to the QC service.

5. Client application

Within the SDC VRE, the GUI of the SeaDataNet Biology quality assessment tool will be developed as a HTML, JavaScript, PHP web application running in a container on the EUDAT B2HOST platform. The front-end application will communicate with several services on the LifeWatch platform.

6. QC service

This service will execute the QC procedures and add the information to the ODV-file(s). This service will be developed as a REST-service calls by the user interface (component 5).