SeaDataCloud

Specifications and development plan for online Biology Data QC service (D1.11)

Filip Waumans, Bart Vanhoorne, Paula Oset Garcia - Flanders Marine Institute

plenary meeting, Brest,18-10-2019
sdn-userdesk@seadatanet.org – www.seadatanet.org

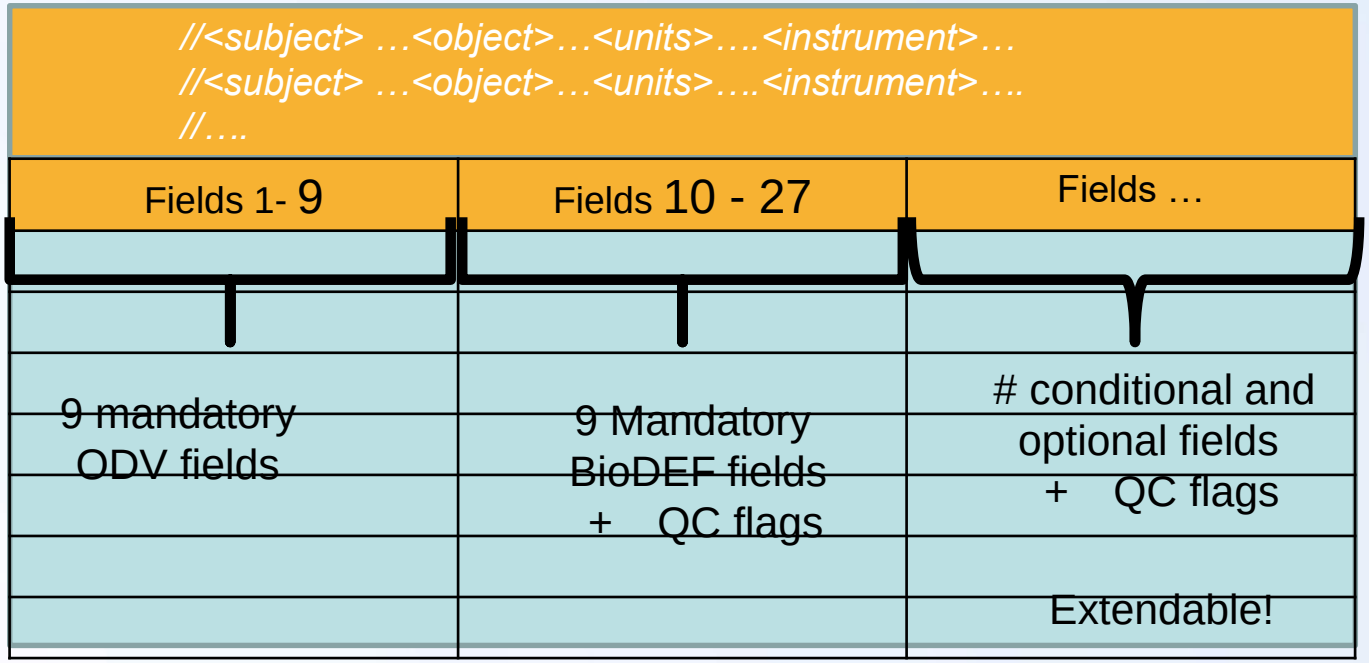
Introduction: bio – odv dataformat

- Seadatanet II: development of biological dataformat: BIO - DEF
  - To store and make accessible biological data using SeaDataNet infrastructure
  - To exchange biological data and to contribute to initiatives like (Eur)OBIS and GBIF.
- biological dataformat: structure
  - CDI XML files for biology metadata = general CDI files
  - specific variant of the ODV files for biology datasets based

BioODV Data Format

HEADER describing fields ←

DATA TABLE ←



Quality Control of Biological datasets in the Seadatacloud VRE

Description

- Development of a tool to process biological datasets and run some quality control checks on occurrence record level
- Analyze the quality and completeness of biology data

Aim

- select data that fit for certain analysis
- identify possible gaps and errors in datasets

Implemented QC checks in current version

Data format

- Are all required fields present and aren't any values missing.
- Is the date field valid (ISO format)

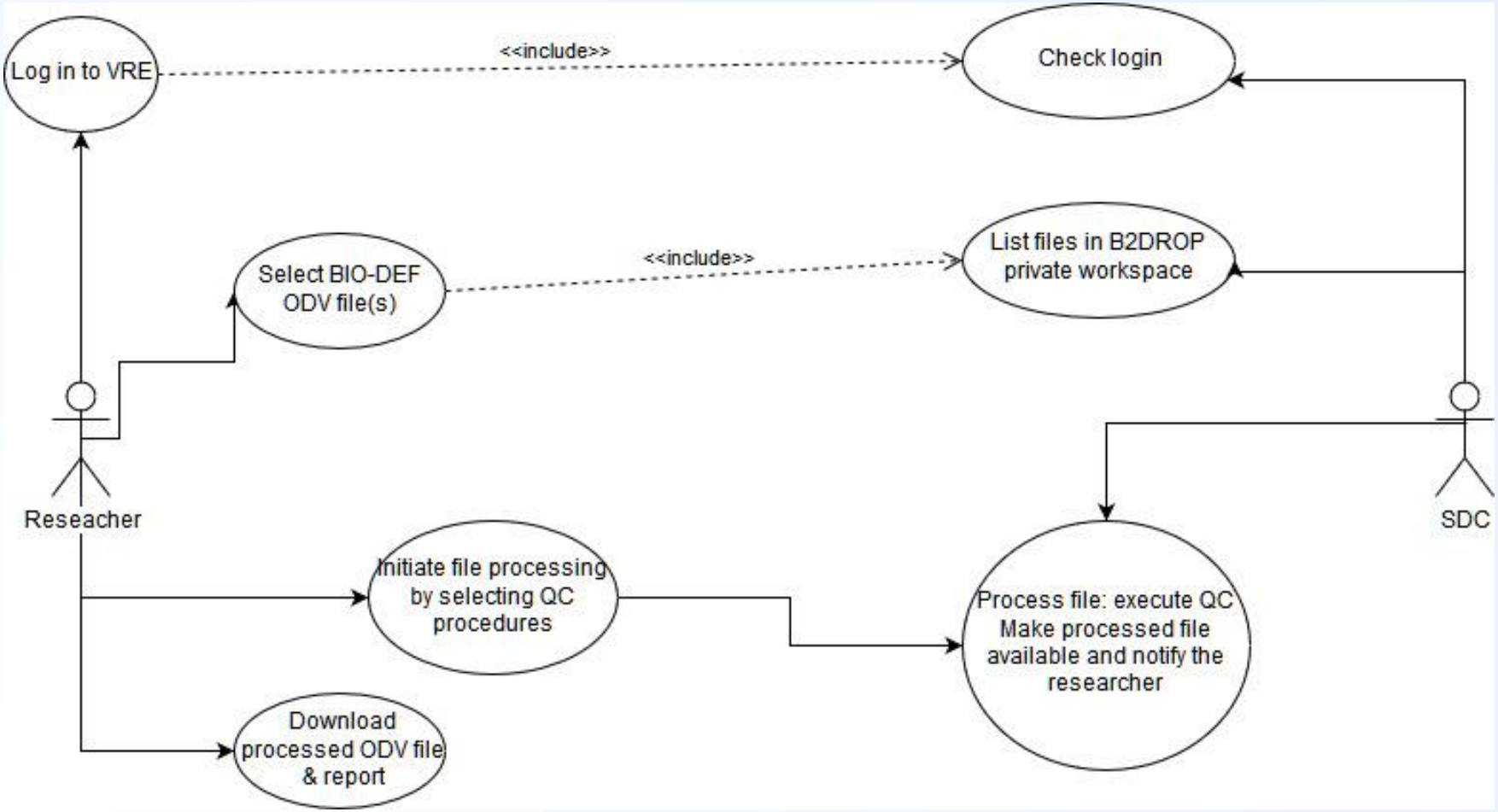
Depth and geograhic quality

- Are coordinates situated in sea (3000m buffer)?
- Have records potentially invalid depths indicated by bathymetry for that position (150m buffer)?

Geographic outlier analysis

- Dataset level: Spatial outliers are detected based on the distance to the geographical centroid of all unique coordinates in the file
- Species level: Outliers are detected based on the coordinates of existing occurrences (spatial outliers) in the OBIS database and based on environmental data

Current use case of the QC tool



Seadatacloud: Biological data Quality Control tool – use case

- **Input**:
  - one biological datafile (ODV format)

- **Processing:**
  - Transform the ODV file into Darwin Core Archive format
  - Execute Quality control procedures

- **Outputfile:**
  - DWA-a occurrence file + report on original ODV file

	A	B	D	F	H	I
1	eventID	CollectionCc	occurrenceID	ScientificName	scientificNameID	occurre
2	BIOFUN1_BF1M1	BIOFUN1	CSIC_BIOFUN1_1	Alepocephalus rostratus	urn:lsid:marinespecies.org:taxname:12668	present
3	BIOFUN1_BF1M1	BIOFUN1	CSIC_BIOFUN1_2	Bathypterois mediterraneus	urn:lsid:marinespecies.org:taxname:29994	present
4	BIOFUN1_BF1M1	BIOFUN1	CSIC_BIOFUN1_3	Coelorinchus mediterraneus	urn:lsid:marinespecies.org:taxname:28031	present
5	BIOFUN1_BF1M1	BIOFUN1	CSIC_BIOFUN1_4	Galeus melastomus	urn:lsid:marinespecies.org:taxname:10581	present
6	BIOFUN1_BF1M1	BIOFUN1	CSIC_BIOFUN1_5	Lepidion lepidion	urn:lsid:marinespecies.org:taxname:12649	present

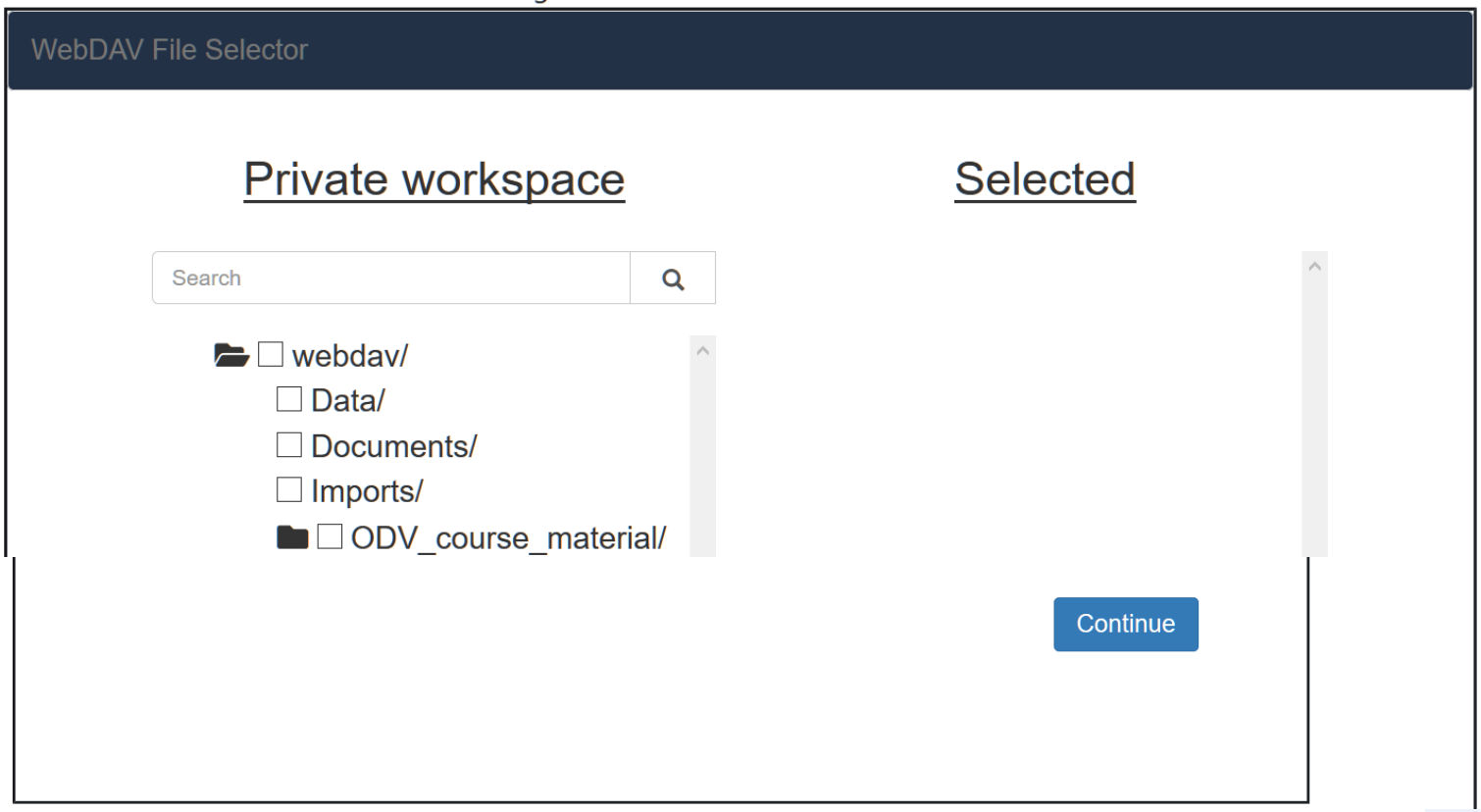
Occurrence table

SeaDataCloud

QC tool prototype

1. Select a biological ODV file | 2. Select QC procedures | 3. Run quality control & confirmation

Select a biological ODV file



Input file name

example_file_bio_odv.txt

SeaDataCloud

QC tool prototype

1. Select a biological ODV file | 2. Select QC procedures | 3. Run quality control & confirmation

Select QC procedures

QC checks

- [x] Check required fields
- [x] Check event data
- [ ] Check if coördinate is situated on land
- [ ] Check the depth value
- [x] Check for geographic outliers in the dataset
- [x] Check of geographic outliers in the observations of the species

Back Next

SeaDataCloud

QC tool prototype

1. Select a biological ODV file | 2. Select QC procedures | 3. Run quality control & confirmation

Run quality control & confirmation

Back | Submit this job

DEMO

QC tool prototype

Job 104

Status	Running
Input file	ODV file
Output report - ODV	
Darwin Core Occurence file	
Submit date	2019-10-15 08:07:45
Start date	2019-10-15 08:07:47
End date	
Error message	

SeaDataCloud

QC tool prototype

Job 104

Status	Success
Input file	ODV file
Output report - ODV	QC report
Darwin Core Occurence file	DWC-a occurence file
Submit date	2019-10-15 08:07:45
Start date	2019-10-15 08:07:47
End date	2019-10-15 08:08:21
Error message	

QC report on bio ODV file

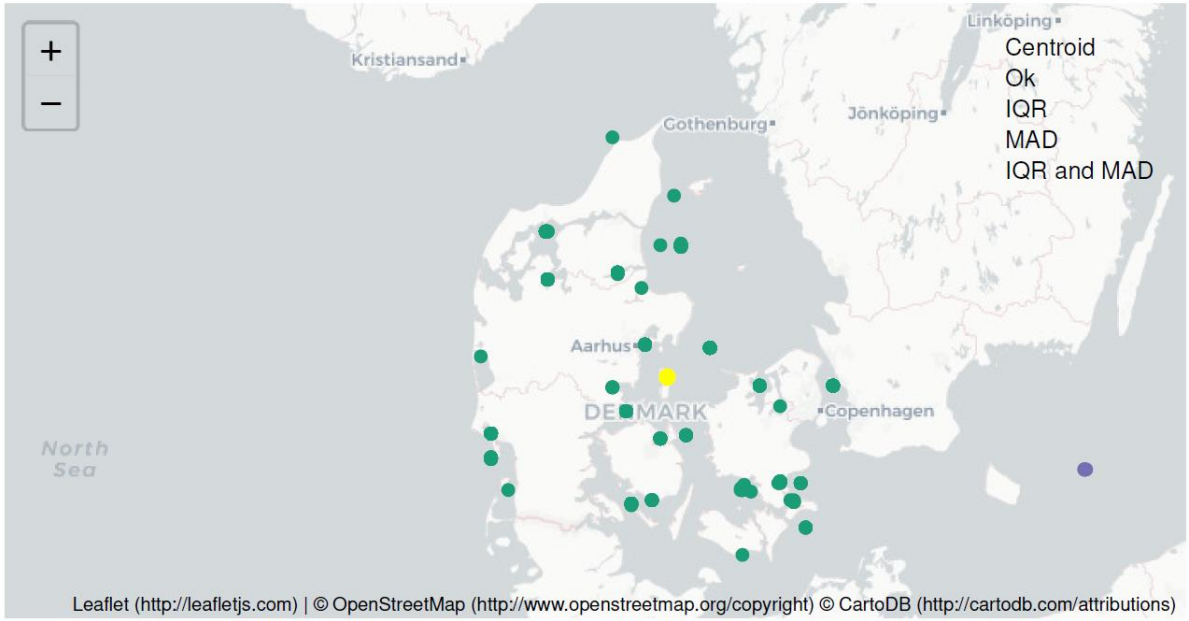
Issues

Outliers

Out

More f

map w



and shown on a

e

field				
Outlie				3838.1]
Outlie				3838.1]
Outlie				3838.1]
Outlie				3838.1]
Outlie				3838.1]
Outliers Dataset	3046	warning	spatial [346251.8] is not within MAD limits [-84769.78, 293838.1]	