

Quality control in OBIS

Pieter Provoost

IOC-UNESCO



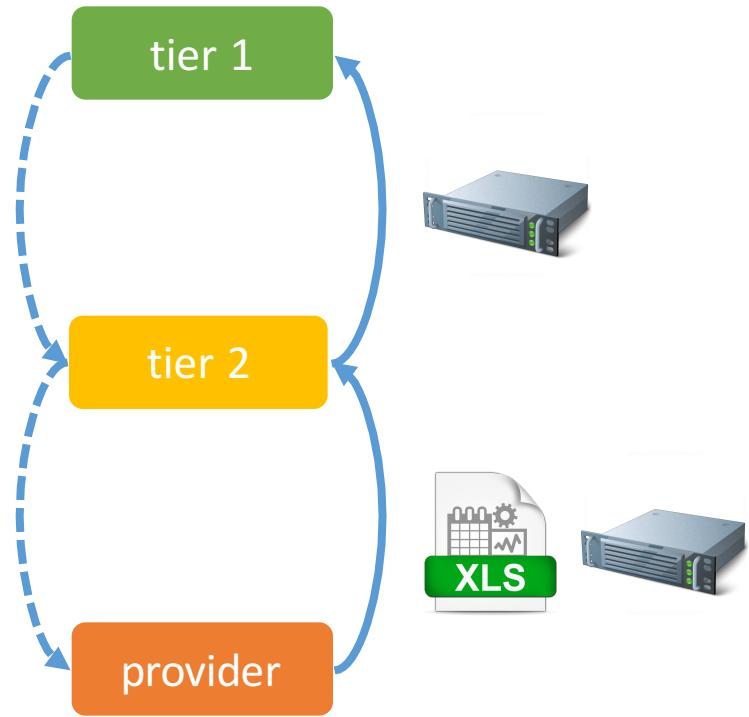
United Nations
Educational, Scientific and
Cultural Organization



Intergovernmental
Oceanographic
Commission

Quality control in OBIS

- Who does QC?
 - Data provider
 - Tier 2 OBIS nodes
 - Tier 1 OBIS node
 - Reject records or values
 - Tier 1 to provide guidelines, tools, and services
- What needs to be checked?
 1. Standards compliance
 2. OBIS required fields present?
 3. Values formatted correctly?
 4. Do the values make sense?



Quality control in OBIS

iOBIS Harvest Report 2015-09-04

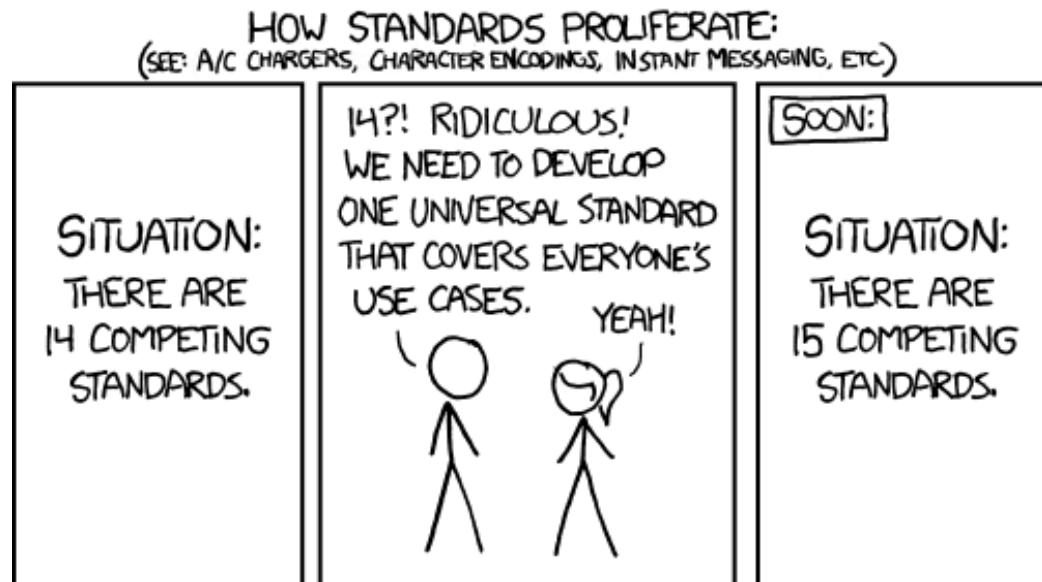
This is a summary report of iOBIS's last data crawling & harvesting activity. It may not reflect the final number of datasets and records added to / refreshed in the subsequently built database as there are more data quality checks performed as the database is built!

Use the links in this table to go to quickly go to the detailed report section...

OBIS Node	No. Resources Crawled	No. Resources Failing QC	No. New Resources	No. Resources Refreshed	No. Resources scheduled to delete	No. Records crawled in total	No. Absence Records	No. Occurrence Records	No. Records with QC Errors (Dropped Records)	No. Records with QC Warnings (Dropped Values)	No. Imported Occurrence Records
OBIS Argentina	1	0	1	0	0	517	0	517	0	458	517
AfrOBIS	14	0	0	14	0	148412	0	148412	11	18058	148401
Antarctic OBIS	86	5	0	81	0	289763	3394	286369	62166	18357	224203
ArCOD	1	0	0	1	0	1728	0	1728	0	0	1728
Caribbean OBIS	8	0	0	8	0	74910	0	74910	1	2	74909
EurOBIS	576	0	2	574	0	17059387	0	17059387	114	360384	17059273
IndiOBIS	48	2	44	2	0	90192	0	90192	17790	748	72402
OBIS Australia	23	0	0	23	0	437254	377	436877	0	3668	436877
OBIS Canada	107	1	4	102	0	3239875	449518	2790357	1245	404555	2789112
OBIS China	6	0	0	6	0	12179	0	12179	1	0	12178
OBIS Japan	5	0	0	5	0	309778	0	309778	0	0	309778
OBIS-SEAMAP	566	0	5	561	0	2503146	1417	2501729	273	18923	2501456
SEAOBIS	2	0	1	1	0	17918	0	17918	15730	0	2188
South Western Pacific OBIS	30	0	6	24	0	1130668	15	1130653	8	8552	1130645
OBIS-USA	31	0	1	30	0	19987712	17487526	2500186	36681	999752	2463505
Western South Atlantic OBIS	41	0	0	41	0	149568	0	149568	388	28786	149180
TOTAL	1545	8	64	1473	0	45453007	17942247	27510760	134408	1862243	27376352

Quality control in OBIS

- Standards: Darwin Core (DwC)
 - Managed by TDWG (Biodiversity Information Standards)
 - Terms: <http://rs.tdwg.org/dwc/terms/>
 - Guidelines on how to use these terms in XML, CSV, etc.
 - <http://beta.iobis.org/manual/darwincore/>
 - "OBIS schema" no longer used



Quality control in OBIS

The screenshot shows a web browser window with the following details:

- Title Bar:** Darwin Core Terms: A quick reference guide
- User:** Pieter
- Address Bar:** rs.tdwg.org/dwc/terms/
- Left Sidebar (TDWG Navigation):**
 - Biodiversity Information Standards TDWG
 - Introduction
 - References
 - Quick Reference Guide
 - Term Index
 - Record-level Terms
 - Occurrence
 - Organism
 - MaterialSample
 - LivingSpecimen
 - PreservedSpecimen
 - FossilSpecimen
 - Event
 - HumanObservation
 - MachineObservation
 - Location
 - GeologicalContext
 - Identification
 - Taxon
 - MeasurementOrFact
 - ResourceRelationship
 - Term Definitions
 - Simple Darwin Core
- Content Area:**

Darwin Core Terms: A quick reference guide

Title: Darwin Core Terms: A quick reference guide

Date Issued: 2009-02-12

Date Modified: 2015-06-02

Abstract: This document is a quick reference for all recommended Darwin Core terms. For complete historical term information, including version changes and pre-standard terms, see [HISTORY]. For a comparative table of elements from pre-standard versions of Darwin Core to the current terms in the standard, see [VERSIONS].

Contributors: John Wieczorek (MVZ), Markus Döring (GBIF), Renato De Giovanni (CRIA), Tim Robertson (GBIF), Dave Vieglais (KUNHM)

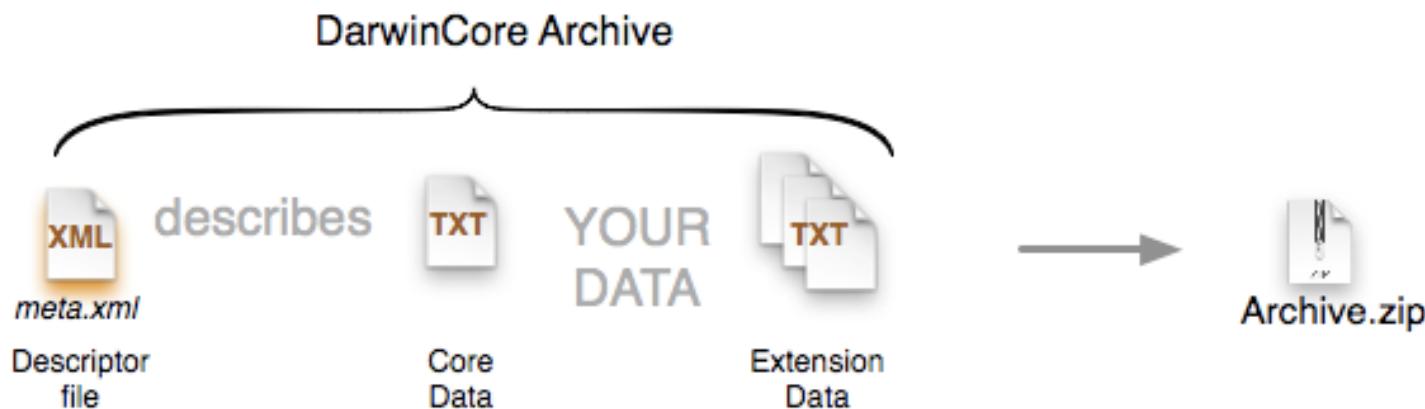
Legal: This document is governed by the standard legal, copyright, licensing provisions and disclaimers issued by the Taxonomic Databases Working Group.

Part of TDWG Standard:

 - Creator:** Darwin Core Task Group
 - Identifier:** <http://rs.tdwg.org/dwc/2015-03-19/terms/>
 - Latest Version:** <http://rs.tdwg.org/dwc/terms/>
 - Replaces:** <http://rs.tdwg.org/dwc/2014-11-08/terms/>
 - Document Status:** Current Standard

Quality control in OBIS

- Standards: Darwin Core Archive (DwC-A)
 - <http://rs.gbif.org/>
 - Managed by GBIF (Global Biodiversity Information Facility)
 - Single self-contained dataset
 - Archive descriptor, metadata and data files



Quality control in OBIS

- OBIS required terms
 - **eventDate** (when?)
 - **decimalLatitude, decimalLongitude** (where?)
 - **scientificName, scientificNameID** (what?)
 - Keep original name!
 - **occurrenceStatus** (present or absent?)
 - **basisOfRecord** (DwC-A required field)

Quality control in OBIS

- OBIS required terms
 - **eventDate** (when?)
 - **decimalLatitude, decimalLongitude** (where?)
 - **scientificName, scientificNameID** (what?)
 - Keep original name!
 - **occurrenceStatus** (present or absent?)
 - **basisOfRecord** (DwC-A required field)

Quality control in OBIS

- Date and time
 - ISO 8601
 - 1973-02-28T15:25:00
 - 1973-02-28T15:25:00Z (UTC)
 - 1993-01-26T04:39+12/1993-01-26T05:48+12
 - 2015-023
 - 2014-W26

PUBLIC SERVICE ANNOUNCEMENT:

OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.

THIS IS **THE** CORRECT WAY TO WRITE NUMERIC DATES:

2013-02-27

THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:

02/27/2013 02/27/13 27/02/2013 27/02/13

20130227 2013.02.27 27.02.13 27-02-13

27.2.13 2013. II. 27. 2^{1/2}-13 2013.158904109

MMXIII-II-XXVII MMXIII ^{LVII}_{CCCLXV} 1330300800

((3+3)×(111+1)-1)×3/3-1/3³ 2013 ²⁰¹³_{misses}

10/11011/1101 02/27/20/13 ^{2 3 1 4}_{5 6 7 8}



Quality control in OBIS

- Not ok:
 - 2015/07/11
 - 09-Dec-2009
 - 10/01/2013
 - 1:25
 - Jan
 - 1915-6-9 0:00:00
 - 1995-7-0
 - 00:18:00+0:00

PUBLIC SERVICE ANNOUNCEMENT:

OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.

THIS IS **THE** CORRECT WAY TO WRITE NUMERIC DATES:

2013-02-27

THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:

02/27/2013 02/27/13 27/02/2013 27/02/13

20130227 2013.02.27 27.02.13 27-02-13

27.2.13 2013. II. 27. $\frac{2}{3}$ -13 2013.158904109

MMXIII-II-XXVII MMXIII $\frac{LVII}{CCCLXV}$ 1330300800

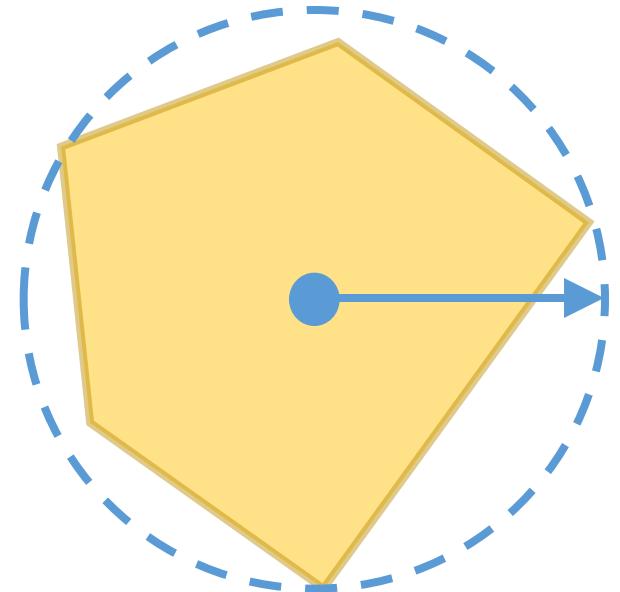
$((3+3)\times(111+1)-1)\times3/3-1/3^3$ 2013 2013 Mississ

10/11011/1101 02/27/20/13 $\frac{2}{5} \frac{3}{1} \frac{4}{7} \frac{2}{8} \frac{7}{3} \frac{4}{1} \frac{2}{7} \frac{3}{8}$



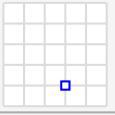
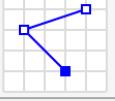
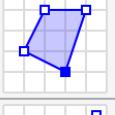
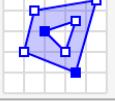
Quality control in OBIS

- Location related terms:
 - **decimalLatitude**
 - **decimalLongitude**
 - **coordinateUncertaintyInMeters**
 - geodeticDatum (EPSG:4326)
 - **footprintWKT**
 - minimumDepthInMeters
 - maximumDepthInMeters
 - locality

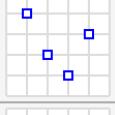
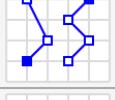
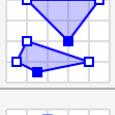
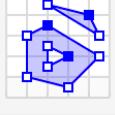


Quality control in OBIS

Geometry primitives (2D)

Type	Examples	
Point		POINT (30 10)
LineString		LINESTRING (30 10, 10 30, 40 40)
Polygon		POLYGON ((30 10, 40 40, 20 40, 10 20, 30 10))
		POLYGON ((35 10, 45 45, 15 40, 10 20, 35 10), (20 30, 35 35, 30 20, 20 30))

Multipart geometries (2D)

Type	Examples	
MultiPoint		MULTIPOINT ((10 40), (40 30), (20 20), (30 10))
		MULTIPOINT (10 40, 40 30, 20 20, 30 10)
MultiLineString		MULTILINESTRING ((10 10, 20 20, 10 40), (40 40, 30 30, 40 20, 30 10))
MultiPolygon		MULTIPOLYGON (((30 20, 45 40, 10 40, 30 20)), ((15 5, 40 10, 10 20, 5 10, 15 5)))
		MULTIPOLYGON (((40 40, 20 45, 45 30, 40 40)), ((20 35, 10 30, 10 10, 30 5, 45 20, 20 35)), (30 20, 20 15, 20 25, 30 20)))

Quality control in OBIS

Screenshot of the OBIS Quality Control interface (obis.github.io/map) showing a map of the Nice area and a list of locations.

Layers
Switch layers on or off.
EEZ **IHO**

WKT
Generate WKT (points not yet included). **WKT**

Coordinates
Add a location using decimal longitude and latitude.
Enter coordinates **Add**

Geocoding
Find locations by name and add them to the locations list.
Enter location name **Submit** **Clear**

Type	Name	Longitude	Latitude
No results			

Locations

Longitude	Latitude	Radius	Name	EEZ	IHO
7.3163	43.6964	709			

Map: A map of the Nice area showing several coastal regions labeled: NICE EST, CORNE D'OR NORD, VILLEFRANCHE-SUR-MER, POVETTE, LES DEUX RUPS, SAINT-ESTÈVE, LA DARSE, LE LAZARET, and SAINT-JEAN-CAP-FERRAT. An orange polygon highlights a specific area near Villefranche-sur-Mer.

Quality control in OBIS

Marine Regions · Oosterschelde Pieter
marineregions.org/gazetteer.php?p=details&id=5332



Marineregions.org

towards a standard for georeferenced marine names

About Marine Gazetteer EEZ boundaries Sources Statistics Downloads

Marine Gazetteer Placedetails

Search

Browse

About

Tutorial

Webservices

Login

MRGID 5332

Status Proposed standard 

Names	Language	Name	Name source
Dutch	Oosterschelde	BIOMARE	
English	Eastern Scheldt	The SeaVoX Salt and Fresh Water Body Gazetteer	

PlaceType Inlet

Latitude 51° 33' 38.9" N (51.56081°)

Longitude 3° 57' 10.3" E (3.95285°)

Precision 23616 meter

Min. Lat 51° 25' 51.4" N (51.4309°)

Min. Long 3° 40' 57.6" E (3.6827°)

Max. Lat 51° 41' 26.4" N (51.6907°)

Max. Long 4° 13' 22.9" E (4.223°)

Source BIOMARE, available online at <http://www.biomareweb.org>

Image 

Quality control in OBIS

Switch layers on or off.

EEZ IHO

Generate WKT (points not yet included).

Add

Add a location using decimal longitude and latitude.

Enter coordinates

AMERICA

EUROPE

AFRICA

SOUTH AMERICA

ASIA

AUSTRALIA

North Atlantic Ocean

South Atlantic Ocean

Indian Ocean

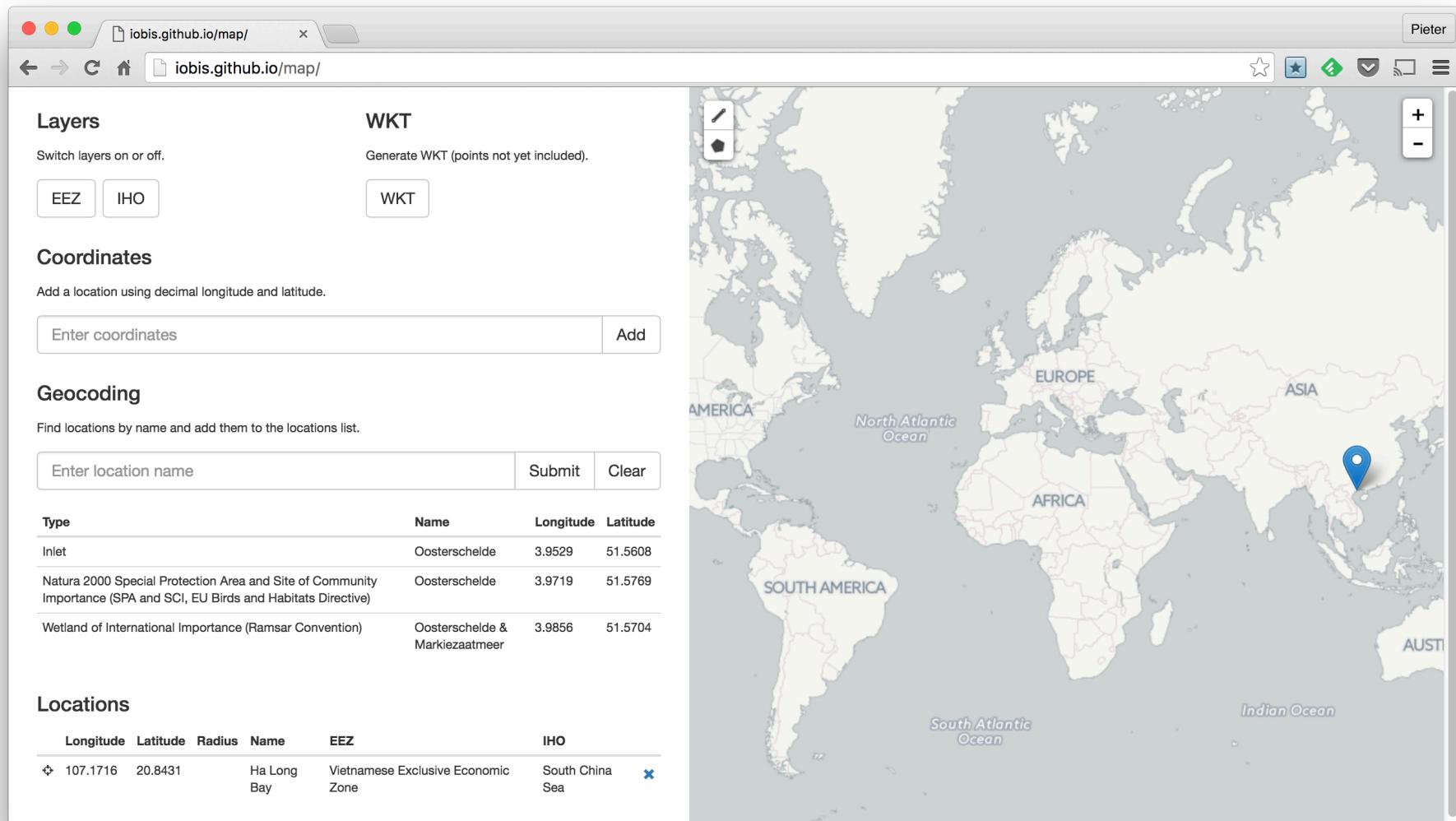
Enter location name

Type Name Longitude Latitude

Type	Name	Longitude	Latitude
Inlet	Oosterschelde	3.9529	51.5608
Natura 2000 Special Protection Area and Site of Community Importance (SPA and SCI, EU Birds and Habitats Directive)	Oosterschelde	3.9719	51.5769
Wetland of International Importance (Ramsar Convention)	Oosterschelde & Markiezaatmeer	3.9856	51.5704

Locations

Longitude	Latitude	Radius	Name	EEZ	IHO
107.1716	20.8431		Ha Long Bay	Vietnamese Exclusive Economic Zone	South China Sea



Quality control in OBIS

- Not ok:
 - -54,00000
 - -550.244.444.444.444
 - -551.675
 - 34.5673°
 - 14°35'26''N

Quality control in OBIS

pieterprovoost/parsnip: Par x Pieter

GitHub, Inc. [US] https://github.com/pieterprovoost/parsnip

parsnip

build passing coverage 97%

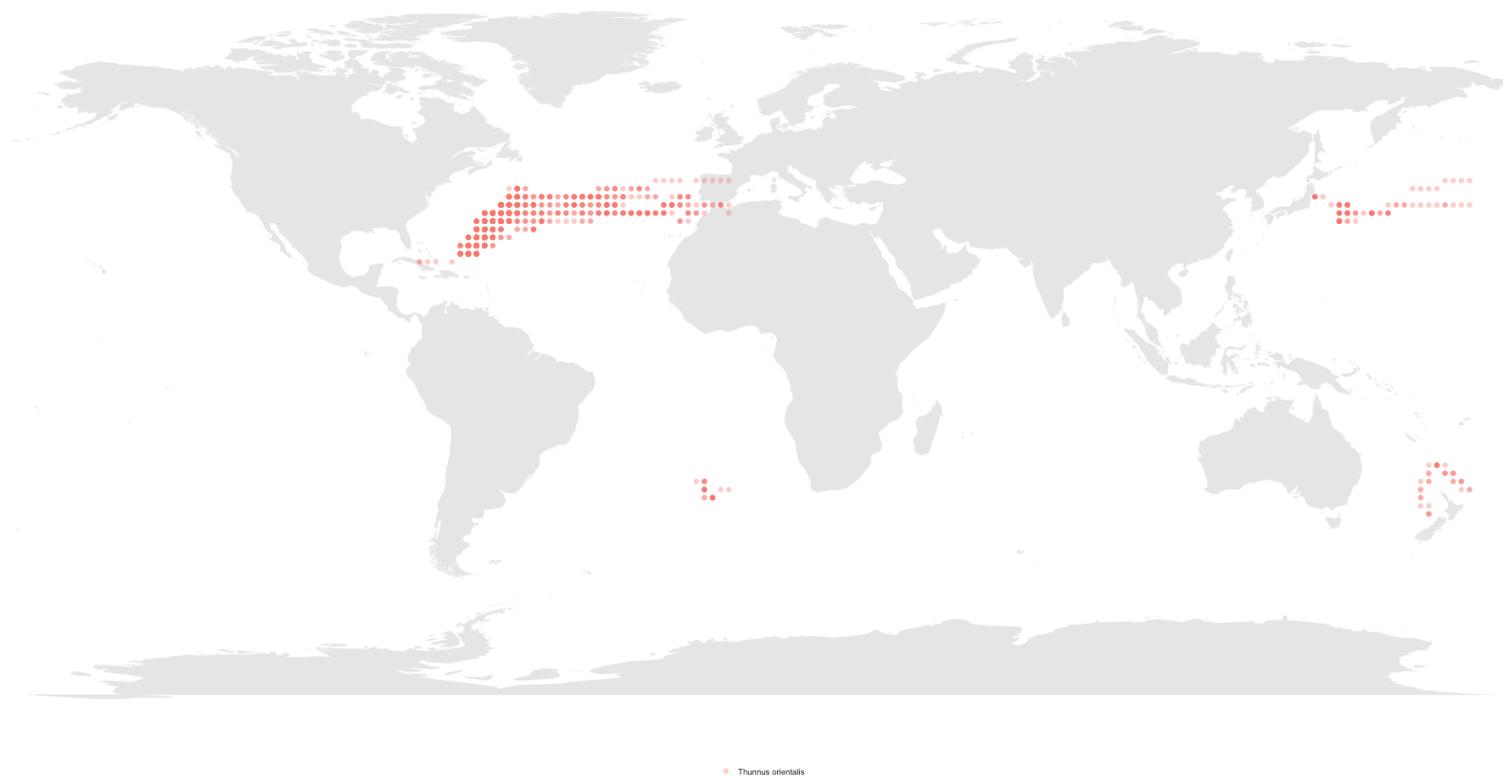
Parsing geographical coordinate strings

Parsing degrees/minutes/seconds

```
parsedms("51°28'38''N 101°16'56''W")
parsedms("51°28'38\"N 101°16'56\"W")
parsedms("51°28`38''N 101°16'56"W")
parsedms("51o28`38''N 101°16'56"W")
parsedms("51°28`38''n 101°16'56"w")
parsedms("51° 28' 38'' N 101° 16' 56'' W")
parsedms("51 ° 28 ' 38 '' N 101 ° 16 ' 56 '' W")
parsedms("51°28'38''N -101°16'56''E")
parsedms("51° N 101° W")
parsedms("51° N")
parsedms("12° N 109° 58' 37" W")
```

Quality control in OBIS

- Make maps!



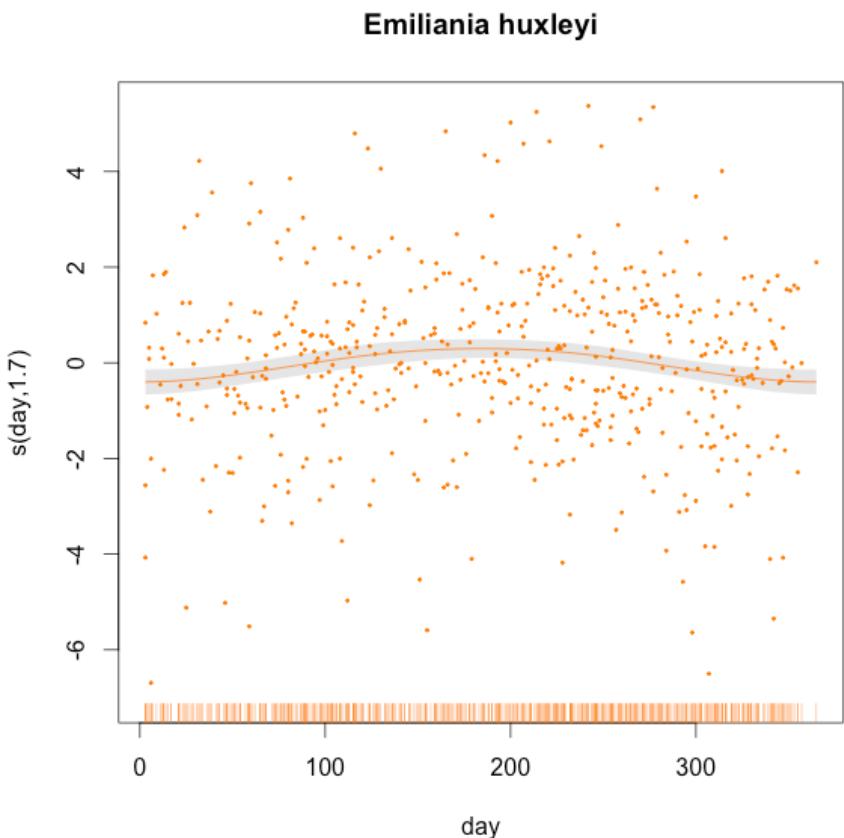
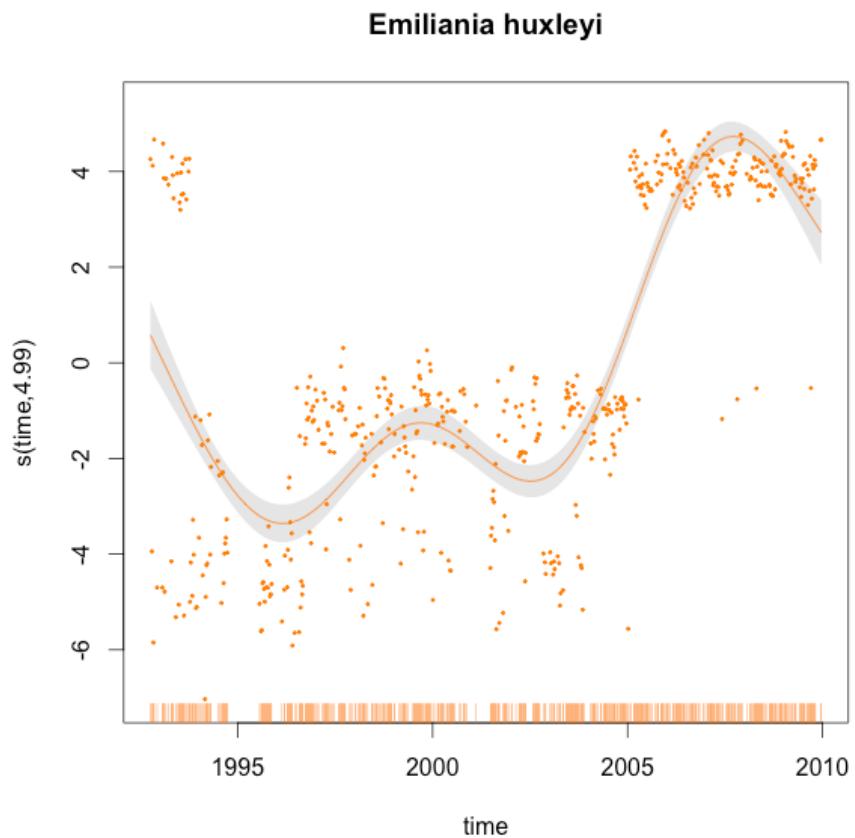
Quality control in OBIS

- Make maps!



Quality control in OBIS

- Make graphs!



Quality control in OBIS

- Make graphs!

The screenshot shows a Microsoft Excel spreadsheet titled "L4_zooplankton_20091009". The table has 904 rows and 9 columns. The columns are labeled: CF, CG, CH, CI, CJ, CK, CL, CM, and CN. The data consists primarily of zeros, with some non-zero values appearing in the first few columns. For example, row 806 has values 0,00, 0,00, 171,60, 0,00, 0,00, 163,80, 7,80, 0,00, and 0,00. Row 822 has values 0,00, 0,00, 248129234,00, 1168917162,00, 0,00, 9829530682,00, 1381447555,00, 0,00, and 0,00. The last row (row 904) has values 0,00, 0,00, 241081231,00, 0,00, 0,00, 2452078431,00, 2174606208,00, 0,00, and 0,00.

	CF	CG	CH	CI	CJ	CK	CL	CM	CN
806	0,00	0,00	171,60	0,00	0,00	163,80	7,80	0,00	0,00
807	0,00	0,00	110,24	0,00	0,00	62,99	47,25	0,00	0,00
808	0,00	0,00	111,00	0,00	0,00	104,92	6,08	0,00	0,00
809	0,00	0,00	197,29	0,00	0,00	197,29	0,00	0,00	0,00
810	0,00	0,00	400,97	0,00	0,00	74,95	326,02	0,00	0,00
811	0,00	0,00	339,11	0,00	0,00	318,63	20,48	0,00	0,00
812	0,00	0,00	360,23	0,00	0,00	132,94	227,29	0,00	0,00
813	31,58	0,00	126,32	0,00	0,00	99,25	27,07	0,00	0,00
814	0,00	0,00	476,11	6,26	0,00	156,62	313,23	0,00	0,00
815	0,00	0,00	150,61	0,00	0,00	37,65	112,96	0,00	0,00
816	0,00	0,00	709,31	0,00	0,00	680,75	28,56	0,00	0,00
817	0,00	0,00	1368,10	0,22	0,00	1148,76	219,13	0,00	119
818	0,00	0,00	865,50	0,00	0,00	727,02	138,48	0,00	0,00
819	0,24	0,00	476,11	0,00	0,00	431,68	44,44	0,00	0,00
820	0,00	0,00	332,16	0,00	0,00	249,12	83,04	0,00	0,00
821	0,00	0,00	330,45	0,19	0,00	196,99	133,26	0,00	0,00
822	0,00	0,00	248129234,00	1168917162,00	0,00	9829530682,00	1381447555,00	0,00	0,00
823	0,00	0,00	9263157895,00	0,21	0,00	197309417,00	7269294312,00	0,00	0,00
824	0,00	0,00	1297465887,00	0,00	0,00	8583235867,00	4391423002,00	0,00	0,00
825	0,00	0,00	1710427011,00	0,38	0,00	9599723981,00	7466451985,00	0,00	0,00
826	0,00	0,00	284307098,00	0,00	0,00	9082032298,00	193486775,00	0,00	0,00
827	0,00	0,00	5611336032,00	0,00	0,00	1726564933,00	3884771099,00	0,00	0,00
828	0,00	0,00	313492823,00	0,00	0,00	1245933014,00	1888995215,00	0,00	0,00
829	0,00	0,00	849668616,00	3404510768,00	0,00	3064059691,00	5398581361,00	0,00	0,00
830	0,00	0,00	3333333333,00	0,00	0,00	1166666667,00	2166666667,00	0,00	0,00
831	0,00	0,00	2537799043,00	0,00	0,00	1359535202,00	1178263841,00	0,00	0,00
832	0,00	0,00	7384615385,00	0,00	0,00	320,00	4184615385,00	0,00	0,00
833	0,00	0,00	930994152,00	0,00	0,00	8583634026,00	8451578118,00	0,00	0,00
834	0,00	0,00	1494892788,00	1026712079,00	0,00	3388149862,00	1155051089,00	0,00	0,00
835	0,00	0,00	241081231,00	0,00	0,00	2452078431,00	2174606208,00	0,00	0,00

Original article

Fishing for data and sorting the catch: assessing the data quality, completeness and fitness for use of data in marine biogeographic databases

**Leen Vandepitte^{1,*}, Samuel Bosch^{1,2}, Lennert Tyberghein¹,
Filip Waumans¹, Bart Vanhoorne¹, Francisco Hernandez¹,
Olivier De Clerck² and Jan Mees¹**

¹Flanders Marine Institute (VLIZ), InnovOcean Site, Wandelaarkaai 7, 8400 Oostende, Belgium and

²Research Group Phycology, University of Ghent, Krijgslaan 281/S8, 9000 Ghent, Belgium

*Corresponding author: Tel: +32-(0)59-34 01 55; Fax: +32-(0)59-34 21 31; Email: leen.vandepitte@vliz.be

Citation details: Vandepitte,L., Bosch,S., Tyberghein,L., et al. Fishing for data and sorting the catch: assessing the data quality, completeness and fitness for use of data in marine biogeographic databases. *Database* (2014) Vol. 2014: article ID bau125; doi:10.1093/database/bau125.

Received 5 September 2014; Revised 19 November 2014; Accepted 8 December 2014

Quality control in OBIS

QC-number	Category	Question	Bit-sequence, if answer is yes	Available as online data service	Implemented in
2	Taxonomy	Is the taxon name matched to WoRMS?	2	Yes (taxon match)	EurOBIS + OBIS
3	Taxonomy	Is the taxon level lower than family?	4	Yes (taxon match)	EurOBIS + OBIS
4	Geography: lat/lon	Are the latitude/longitude values different from zero?	8	Yes (check OBIS format)	EurOBIS + OBIS
5	Geography: lat/lon	Are the latitude/longitude values within their possible boundaries?	16	Yes (check OBIS format)	EurOBIS + OBIS
6	Geography: lat/lon	Are the coordinates situated in sea or along the coastline (20 km buffer)?	32	Yes (check OBIS format)	EurOBIS + OBIS
9	Geography: lat/lon	Are the coordinates situated in the expected geographic area (compare metadata)?	256	No, but visual check possible through separate data validation service	EurOBIS
18	Geography: depth	Is minimum depth \leq maximum depth?	131 072	Not yet available	EurOBIS + OBIS
19	Geography: depth	Is the sampling depth possible when compared with GEBCO depth map (incl. margin)?	262 144	No, but depths per lat-lon can be requested through geographic web services	EurOBIS + OBIS

Quality control in OBIS

QC ID	Category	Description	Count	Status	Source
25	Outliers:environment	Is the observation within six MADs from the median SST of this taxon?	16 777 216	Not yet available	OBIS
26	Outliers:environment	Is the observation within three IQRs from the first & third quartile SST of this taxon?	33 554 432	Not yet available	OBIS
27	Outliers:geography	Is the observation within six MADs from the distance to the centroid of this taxon?	67 108 864	Not yet available	OBIS
28	Outliers:geography	Is the observation within three IQRs from the first & third quartile distance to the centroid of this taxon?	134 217 728	Not yet available	OBIS
29	Outliers:geography	Is the observation within six MADs from the distance to the centroid of this dataset?	268 435 456	Not yet available	OBIS
30	Outliers:geography	Is the observation within three IQRs from the first & third quartile distance to the centroid of this dataset?	536 870 912	Not yet available	OBIS

Quality control in OBIS

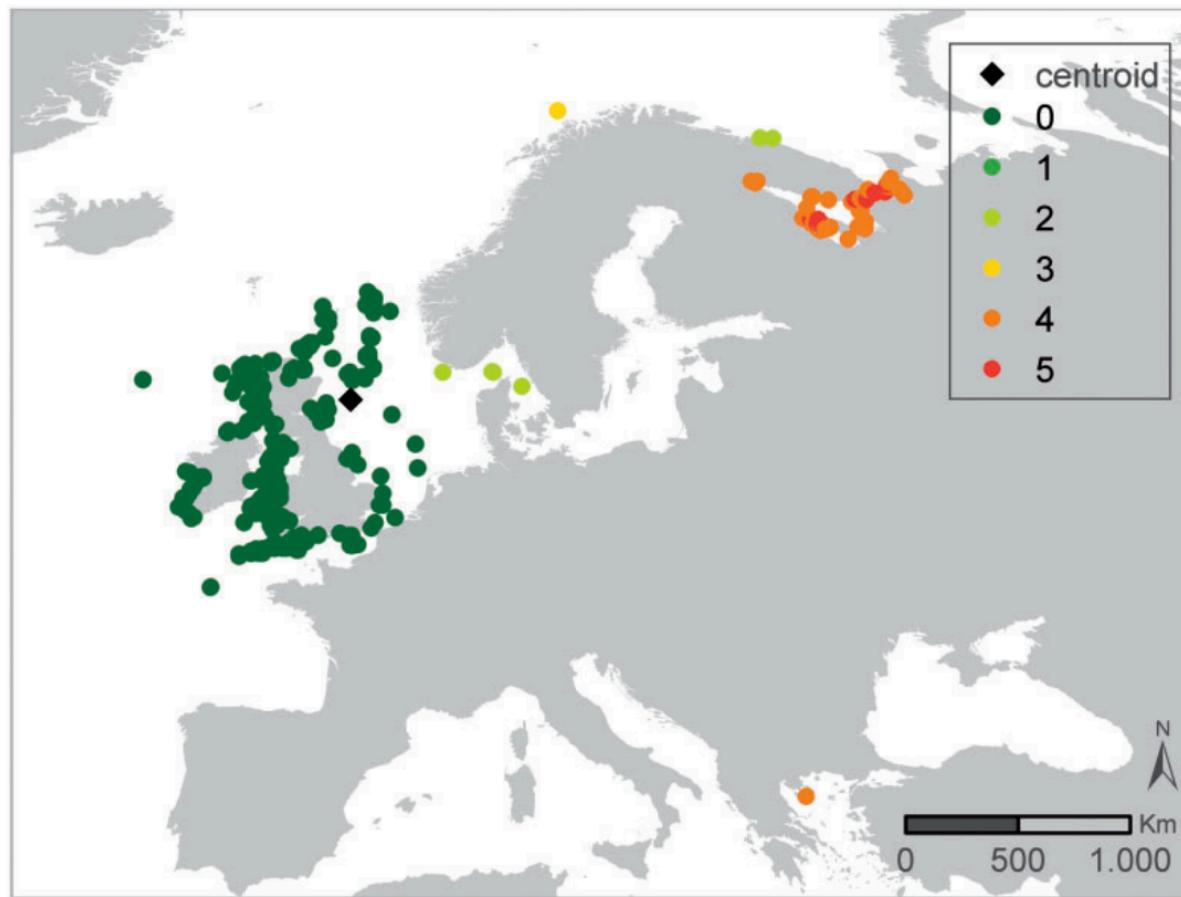


Figure 5. Synthesis map representing the combined results of the outlier analyses of *Verruca stroemia* from [Figure 4](#). The scale represents the number of times a species distribution is seen as an outlier, when combining the eight outlier analyses—geography, bathymetry, Sea Surface Salinity (SSS) and Sea Surface Temperature (SST) SSS and SST according to the IQR and MAD approach—from [Figure 4](#). The black diamond indicates the centroid of the investigated data.