How to perform Quality Control on your data

2020-11-05

Biodiversity data from the field to research

Yi-Ming Gan, Anton Van de Putte, Maxime Sweetlove

SCAR Antarctic Biodiversity Portal

Let us know if you prefer not to be recorded.

Code of conduct

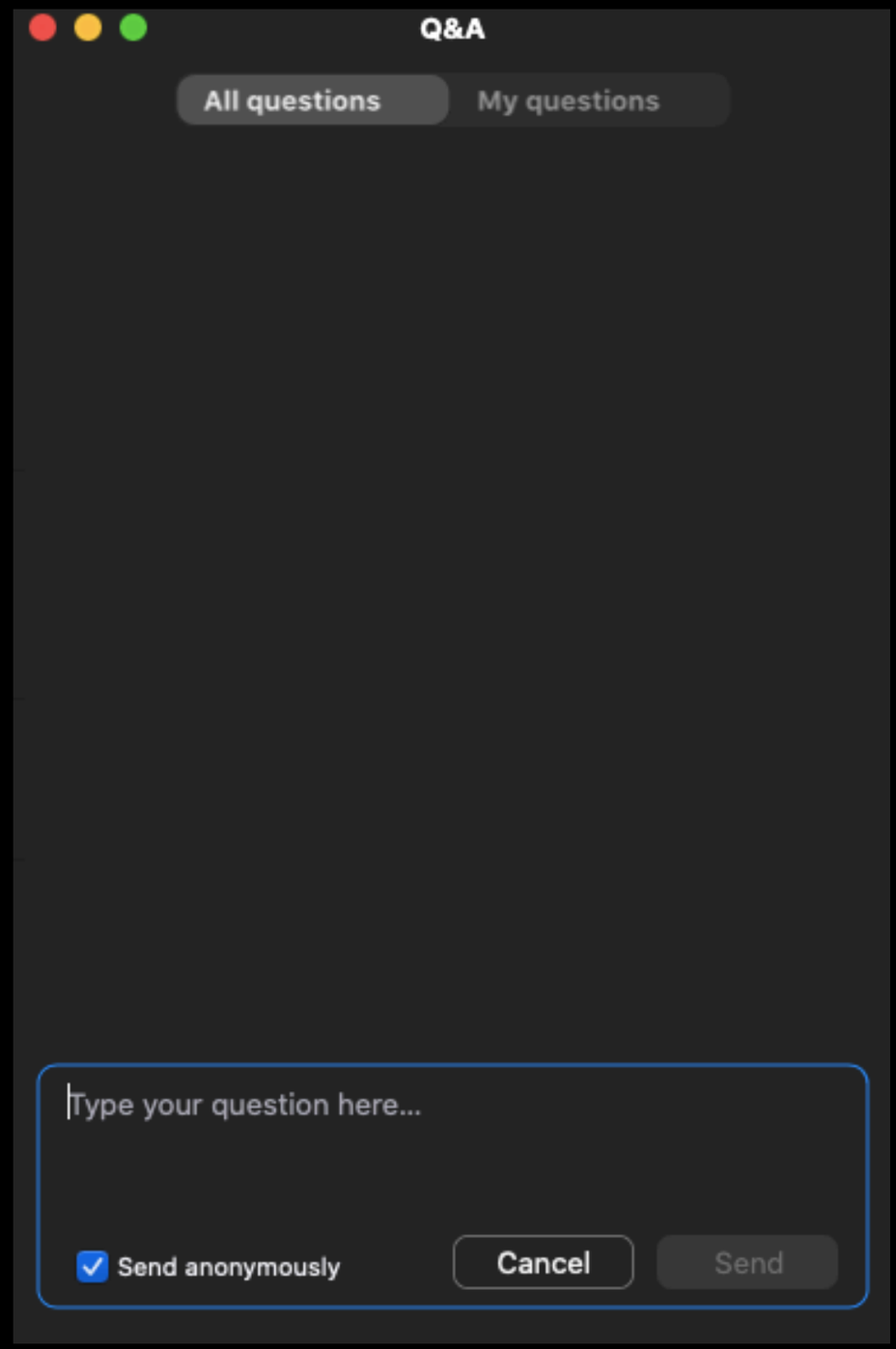
Be respectful

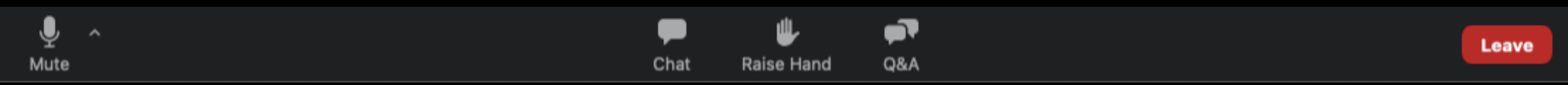
We will follow the principles of the rOpenSci Code of Conduct
https://ropensci.org/code-of-conduct/

Using Zoom in this webinar

Chat with panelists if you encounter technical problems

Ask questions using Q&A feature or raise your hand





Schedule for discussion

These webinars won’t cover everything

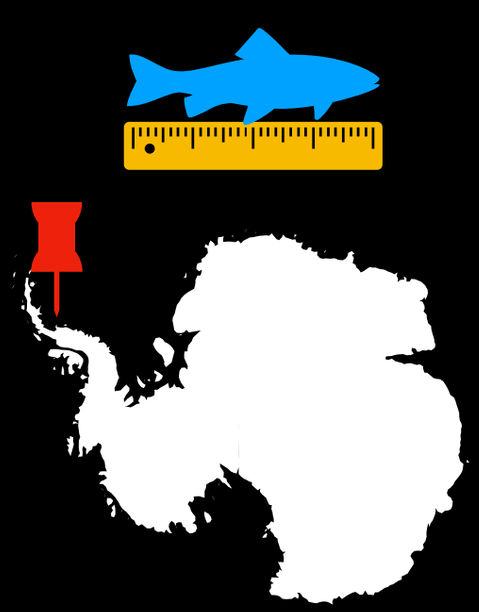
- Discover more based on the links that we provide
- Schedule a session to discuss directly
  - https://doodle.com/meetme/qc/XxYYpJwmbG

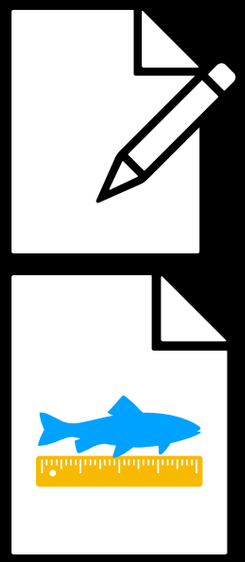
3 levels of quality control (QC)

Device QC & calibration

In situ measurement

Delayed mode QC





What, When, Where, Who, How

Document

Digitized

Distributed

3 levels of quality control (QC)

Device QC & calibration

In situ measurement QC

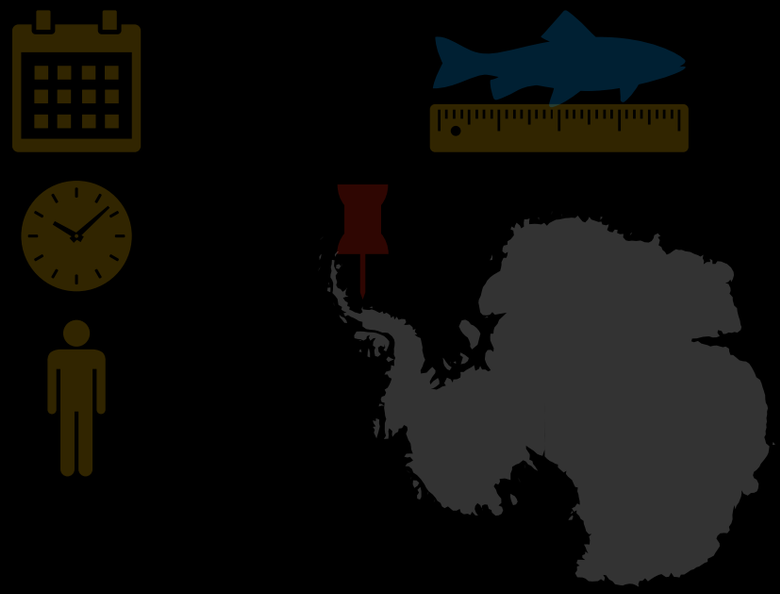
Delayed mode QC

Cruise reports

Aggregated data

Cleaned data

Metadata



What, When, Where, Who, How

Document

Digitized

Distributed

Legend

Done by scientists

Done by data manager

Use Case

- You want to combine data you receive from other participants in your expedition

- We assume you use a similar data template

QC workflow Tools

Cookiecutter

Excel

Openrefine

QGIS

R

COOKIECUTTER

- A well-defined, standard project structure
- Application/mindset
- Requires Python
- don’t need to know Python
- Good for any data project

```
├── LICENSE
├── Makefile           <- Makefile with commands like `make data` or `make train`
├── README.md          <- The top-level README for developers using this project.
├── data
│   ├── external       <- Data from third party sources.
│   ├── interim        <- Intermediate data that has been transformed.
│   ├── processed      <- The final, canonical data sets for modeling.
│   └── raw            <- The original, immutable data dump.
│
├── docs               <- A default Sphinx project; see sphinx-doc.org for details
│
├── models             <- Trained and serialized models, model predictions, or model summaries
│
├── notebooks          <- Jupyter notebooks. Naming convention is a number (for ordering),
│                         the creator's initials, and a short `-` delimited description, e.g.
│                         `1.0-jqp-initial-data-exploration`.
│
├── references         <- Data dictionaries, manuals, and all other explanatory materials.
│
├── reports            <- Generated analysis as HTML, PDF, LaTeX, etc.
│   └── figures        <- Generated graphics and figures to be used in reporting
│
├── requirements.txt   <- The requirements file for reproducing the analysis environment, e.g.
│                         generated with `pip freeze > requirements.txt`
│
├── setup.py           <- Make this project pip installable with `pip install -e`
├── src                <- Source code for use in this project.
│   ├── __init__.py    <- Makes src a Python module
│   │
│   ├── data           <- Scripts to download or generate data
│   │   └── make_dataset.py
│   │
│   ├── features       <- Scripts to turn raw data into features for modeling
│   │   └── build_features.py
│   │
│   ├── models         <- Scripts to train models and then use trained models to make
│   │   │                 predictions
│   │   ├── predict_model.py
│   │   └── train_model.py
│   │
│   └── visualization  <- Scripts to create exploratory and results oriented visualizations
│       └── visualize.py
│
└── tox.ini            <- tox file with settings for running tox; see tox.testrun.org
```

https://drivendata.github.io/cookiecutter-data-science/
https://github.com/cookiecutter/cookiecutter

Cookiecutter project structure

For R users

```
├── README.md          <- The top-level README.
├── data
│   ├── external       <- Data from third party sources.
│   ├── interim        <- Intermediate data that has been transformed.
│   ├── processed      <- The final, canonical data sets for modeling.
│   └── raw            <- The original, immutable data dump.
│
├── references         <- Data dictionaries, manuals, and all other explanatory
materials.
│
├── src                <- Source code for use in this project.
│   ├── data           <- Scripts to download or generate data.
│   │   └── make_dataset.R
│   │
│   ├── clean          <- Scripts to clean data.
│   │   └── clean_dataset.R
│   │
│   └── visualization  <- Scripts to create exploratory and results oriented
visualizations
│       └── visualize.R
│
└── reports            <- Generated QC reports.
    └── figures        <- Generated graphics and figures to be used in reporting.
```

Adapted from Python's version from Cookiecutter Data Science https://drivendata.github.io/cookiecutter-data-science/ accessed on 2020-10-29

Cookiecutter project structure

For Excel users

```
├── README.md          <- The top-level README.
├── data
│   ├── external       <- Data from third party sources.
│   ├── interim        <- Intermediate data that has been transformed.
│   ├── processed      <- The final, canonical data sets for modeling.
│   └── raw            <- The original, immutable data dump.
│
├── references         <- Data dictionaries, manuals, and all other explanatory
materials.
│
└── reports            <- Generated QC report.
    └── figures        <- Generated graphics and figures to be used in reporting
```

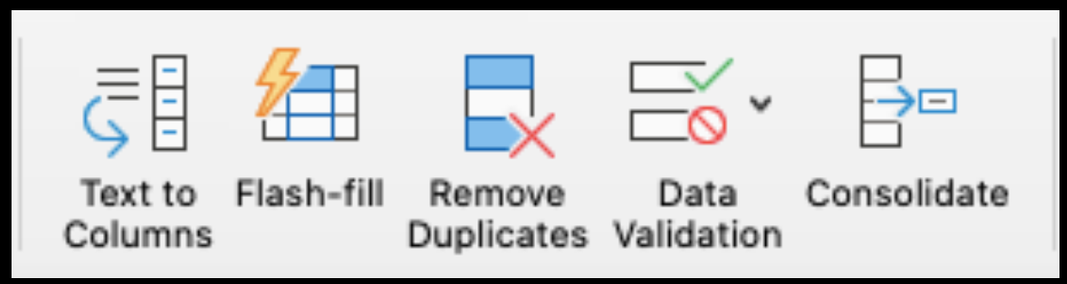
Adapted from Python's version from Cookiecutter Data Science https://drivendata.github.io/cookiecutter-data-science/ accessed on 2020-10-29

Cookiecutter project structure

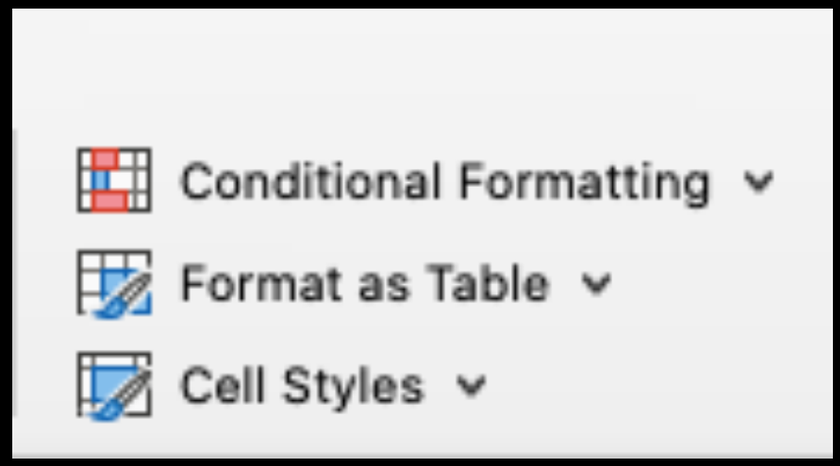
You can use it for anything

```
2020-11-05_data-qc
 |    ├── 2020-10-28_webinar_QC.key
 |    ├── data
 |    |    ├── 01_coordinate_invert-lat-lon.xlsx
 |    |    ├── 01_coordinate_invert-lat-lon.xlsx.Sheet1.vrt
 |    |    ├── 01_coordinate_negative-sign.xlsx
…
 |    ├── demo
 |    |    ├── 00_QGIS.mov
 |    |    ├── 01_excel_text-to-columns.mov
 |    |    ├── 02_WoRMS_taxon-match.mov
 |    |    ├── 03_excel_vlookup.mov
 |    |    └── 04_lifewatch_e-lab.mov
 |    ├── figures
 |    |    ├── 00_tips_cruise-events.png
 |    |    ├── 00_tips_lubridate.png
 |    |    ├── 00_tips_obistools-check-pts-on-land.png
 |    |    ├── 01_coordinate_invert-lat-lon.png
…
 |    |    ├── 04_taxon_worms-rest-api.png
 |    |    └── 04_taxon_worrms.png
```

Excel



- Some feature can be used for QC



- But dates and time….

OpenRefine

- is a powerful tool for working with messy data: cleaning it; transforming it from one format into another; and extending it with web services and external data.

- **1. Explore Data**
- **2. Clean and Transform Data**
- **3. Reconcile and Match Data**

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...

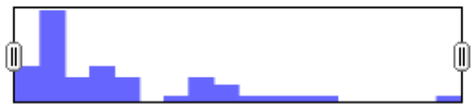
Method key collision ▾ Keying Function fingerprint ▾ **48** clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
3	63	• Pune Vidhyapeeth Gate (33 rows) • Pune Vidhyapeeth Gate (17 rows) • Pune Vidhyapeeth Gate (13 rows)	☐	Pune Vidhyapeeth Gate
3	101	• Hadapsar Gadital (83 rows) • Hadapsar Gadital (14 rows) • Hadapsar Gadital (4 rows)	☐	Hadapsar Gadital
3	12	• Devachi Uruli Phata (8 rows) • Devachi Uruli Phata (2 rows) • Uruli Devachi Phata (2 rows)	☐	Devachi Uruli Phata
2	2	• SRP Stadium (1 rows) • SRP Stadium (1 rows)	☐	SRP Stadium
2	8	• Khandoba Mandir Corner (6 rows) • Khandoba Mandir corner (2 rows)	☐	Khandoba Mandir Corner
2	7	• St Meera College (5 rows) • ST Meera College (2 rows)	☐	St Meera College
2	67	• Gunjan Corner (62 rows)	☐	Gunjan Corner

 # Choices in Cluster

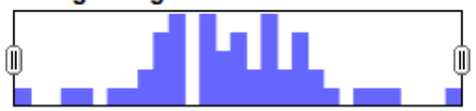
2 — 3

 # Rows in Cluster



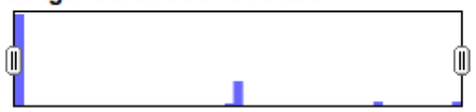
0 — 170

Average Length of Choices



3 — 31

Length Variance of Choices



0 — 1

- R is a free software environment for statistical computing and graphics

- R studio is an interface

- Packages can be found on CRAN
  - Obistools
  - Somap

R package for data QC

obistools: Tools for data enhancement and quality control.

build error | coverage 86% | DOI 10.5281/zenodo.3338213

Installation
Taxon matching
Check required fields
Plot points on a map
Identify points on a map
Check points on land
Check depth
Check outliers
Check eventID and parentEventID
Check eventID in an extension
Flatten event records
Flatten occurrence and event records
Calculate centroid and radius for WKT geometries
Map column names to Darwin Core terms
Check eventDate
Dataset structure
Data quality report
Lookup XY

https://github.com/iobis/obistools

GUI software to visualize geographical coordinates

Software Antarctic Layers

QGIS: https://www.qgis.org/en/site/, Quantarctica: https://www.npolar.no/quantarctica/

QGIS 2.18.3 - Quantarctica

Project Edit View Layer Settings Plugins Vector Raster Database Web MMQGIS Processing Help

Layers Panel

- Quantarctica v2
  - Miscellaneous ->
    - Overview Place Names
    - SCAR Place names
    - Map frame
    - COMNAP listed facilities
    - UTM zones
    - Antarctic Circle
    - Graticules, 30-dg. int.
    - Graticules, 20-dg. int.
    - Graticules, 15-dg. int.
    - Graticules, 10-dg. int.
    - Graticules, 5-dg. int.
    - Graticules, 1-dg. int. (Sc.dep.)
    - Graticules, 0.5-dg. lat. (Sc.dep.)
  - Elevation contours ->
  - Geophysics ->
    - ADMAP ->
    - GOCE ->
  - Glaciology ->
    - Subglacial lakes ->
    - ASAID Grounding/hydros...
    - ALBMAP compilation ->
    - Blue ice areas, Hui
    - MEaSUREs Flow speed ->
      - MEaSUREs Flow vector
      - MEaSUREs Flow speed error
      - MEaSUREs Flow speed
    - GSPC Drainage systems
  - Terrain rasters ->
    - BEDMAP2 ->
      - BEDMAP2 Data coverage
      - BEDMAP2 Exposed rock m...
      - BEDMAP2 Lake Vostok mask
      - BEDMAP2 Ice outline mask
      - BEDMAP2 Surface elevation
      - BEDMAP2 Bed elevation
      - BEDMAP2 Bed uncertainty
      - BEDMAP2 Ice thickness
      - BEDMAP2 Ice thickness un...
    - Hillshade 50x, Bamber
    - DEM, Bamber
    - RAMP2 Hillshade 50x
    - RAMP2 DEM
    - ETOPO1 Hillshade 50x
    - ETOPO1 DEM

Geology

Biology

Imagery

Ocean

Surface melt & firn

Facilities

Bed topography

Ice cores and climate

Atmosphere

DEMs & bathymetry

Ice flow

Detailed Basemap

2 000 km

Coordinate 62.3212,-38.6612 Scale :40,854,315 Magnifier 100% Rotation 0.0 Render EPSG:3031 (OTF)

Getting started

Naming files

Give logical names to your files

```
data.txt
taxon.txt
taxon_matched.txt
cleaned.txt
clean_data.R
```

README.md

```
data.txt - raw data received on 25/12/2018
taxon.txt - unique taxon from raw data
taxon_matched.txt - matched taxon from worms on
5/11/2020
cleaned.txt - cleaned data using clean_data.R script
clean_data.R - script to clean data
```

Give logical names to your files

```
data.txt
taxon.txt
taxon_matched.txt
cleaned.txt
clean_data.R
```

```
├──src/
│   └──data/
│       └──2020-11-05_clean-data.R
│
├──data/
│   ├──raw/
│   │   └─2018-12-25_data.txt
│   ├─-interim/
│   │   ├──2020-11-05_unique-taxon.txt
│   │   └──2020-11-05_taxon-matched.txt
│   └──processed/
│       └──2020-11-05_cleaned.txt
```

Give good name to your files

Machine readable, human readable, orderable

```
~/Desktop/projects/01_cruise-reports ▷ ls
2018-12-15_PS117_Cape-Town.pdf
2019-02-09_PS118_Punta-Arenas.pdf
2019-04-13_PS119_Punta-Arenas.pdf
2020-06-04_PS122-4_Arctic-Ocean.pdf
2020-08-12_PS122-5_Arctic-Ocean.pdf
```

Bryan, J. (2015) How to name files. Reproducible Science Workshop, Duke. https://speakerdeck.com/jennybc/how-to-name-files

Machine readable

Recover metadata from file names with delimiters

```
~/Desktop/projects/01_cruise-reports ▷ ls
2018-12-15_PS117_Cape-Town.pdf
2019-02-09_PS118_Punta-Arenas.pdf
2019-04-13_PS119_Punta-Arenas.pdf
2020-06-04_PS122-4_Arctic-Ocean.pdf
2020-08-12_PS122-5_Arctic-Ocean.pdf
```

Bryan, J. (2015) How to name files. Reproducible Science Workshop, Duke. https://speakerdeck.com/jennybc/how-to-name-files

Human readable

File names that give you an idea of its content

```
~/Desktop/projects/01_cruise-reports ▷ ls

2018-12-15.pdf          2018-12-15_PS117_Cape-Town.pdf
2019-02-09.pdf          2019-02-09_PS118_Punta-Arenas.pdf
2019-04-13.pdf          2019-04-13_PS119_Punta-Arenas.pdf
2020-06-04.pdf          2020-06-04_PS122-4_Arctic-Ocean.pdf
2020-08-12.pdf          2020-08-12_PS122-5_Arctic-Ocean.pdf
```

Date + Expedition code + Location

Bryan, J. (2015) How to name files. Reproducible Science Workshop, Duke. https://speakerdeck.com/jennybc/how-to-name-files

Orderable

Order your files with ISO 8601 date format

YYYY-MM-DD

```
2018-12-15_PS117_Cape-Town.pdf
2019-02-09_PS118_Punta-Arenas.pdf
2019-04-13_PS119_Punta-Arenas.pdf
2020-06-04_PS122-4_Arctic-Ocean.pdf
2020-08-12_PS122-5_Arctic-Ocean.pdf
```

DD-MM-YYYY

```
04-06-2020_PS122-4_Arctic-Ocean.pdf
09-02-2019_PS118_Punta-Arenas.pdf
12-08-2020_PS122-5_Arctic-Ocean.pdf
13-04-2019_PS119_Punta-Arenas.pdf
15-12-2018_PS117_Cape-Town.pdf
```

Bryan, J. (2015) How to name files. Reproducible Science Workshop, Duke. https://speakerdeck.com/jennybc/how-to-name-files

Orderable

Left pad your file names with 0

```
01-01_argo-float.txt
01-02_bathythermograph.txt
02-01_argo-float.txt
02-02_bathythermograph.txt
10-01_argo-float.txt
10-02_bucket-water-sampling.txt
10-11_ctd-rosette.txt
11-01_plankton-net.txt
11-02_bucket-water-sampling.txt
```

```
1-1_argo-float.txt
1-2_bathythermograph.txt
10-1_argo-float.txt
10-11_ctd-rosette.txt
10-2_bucket-water-sampling.txt
11-1_plankton-net.txt
11-2_bucket-water-sampling.txt
2-1_argo-float.txt
2-2_bathythermograph.txt
```

Bryan, J. (2015) How to name files. Reproducible Science Workshop, Duke. https://speakerdeck.com/jennybc/how-to-name-files

Don’t overwrite data

```
├── README.md          <- The top-level README.
├── data
│   ├── external       <- Data from third party sources.
│   ├── interim        <- Intermediate data that has been transformed.
│   ├── processed      <- The final, canonical data sets for modeling.
│   └── raw            <- The original, immutable data dump.
│
```

General inspection

General inspection

Insufficient information, unclear headers

I **guess** these are
decimalLatitude &
decimalLongitude in EPSG:4326
coordinate system

What does this mean?

code	LAT	LONG	Depth
??N dia dia	-74.15	-29.68333333	2012
A mur	-59.9333333	-65.3	3687
A mur bil	-54.8333333	-129.8	1035
G and	-71.25	-13.06666667	186

I **guess** this is recorded in meters

General inspection

Unit not specified

code	LAT	LONG	Depth
??N dia dia	-74.15	-29.68333333	2012
A mur	-59.9333333	-65.3	3687
A mur bil	-54.8333333	-129.8	1035
G and	-71.25	-13.06666667	186

I **guess** this is recorded in meters?

Confusing data without unit:

- Temperature: °C or °F
- Depth: m, km, ft, fathoms …

Use Openrefine to

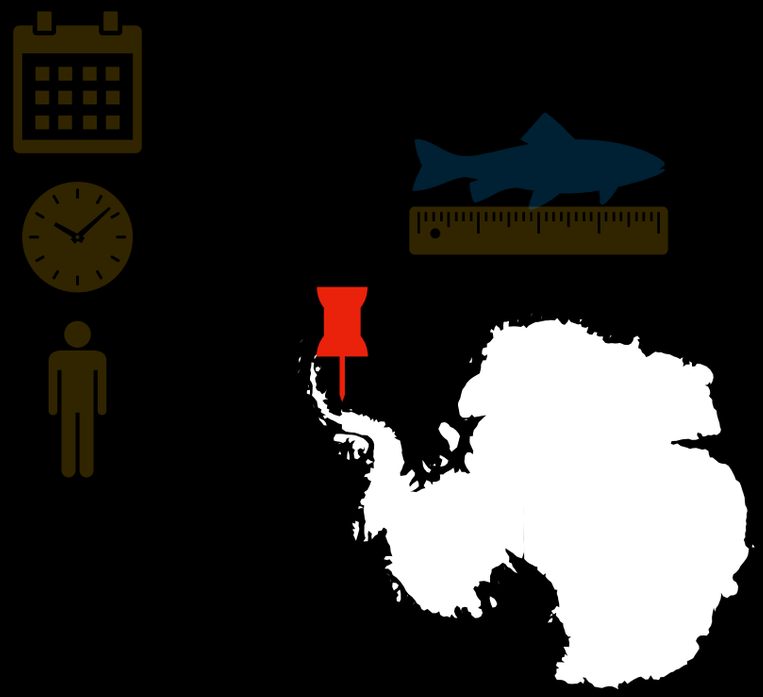
do a general cleaning
Remove typos
Standardise terms within the file

occurrenceID

- Ensure that occurrenceID is unique for each record

basisOfRecord

- Ensure standard labels are being used
- Remove spaces
- Correct lower case and upper case



Where, When, What, Who, How

decimalLongitude, decimalLatitude

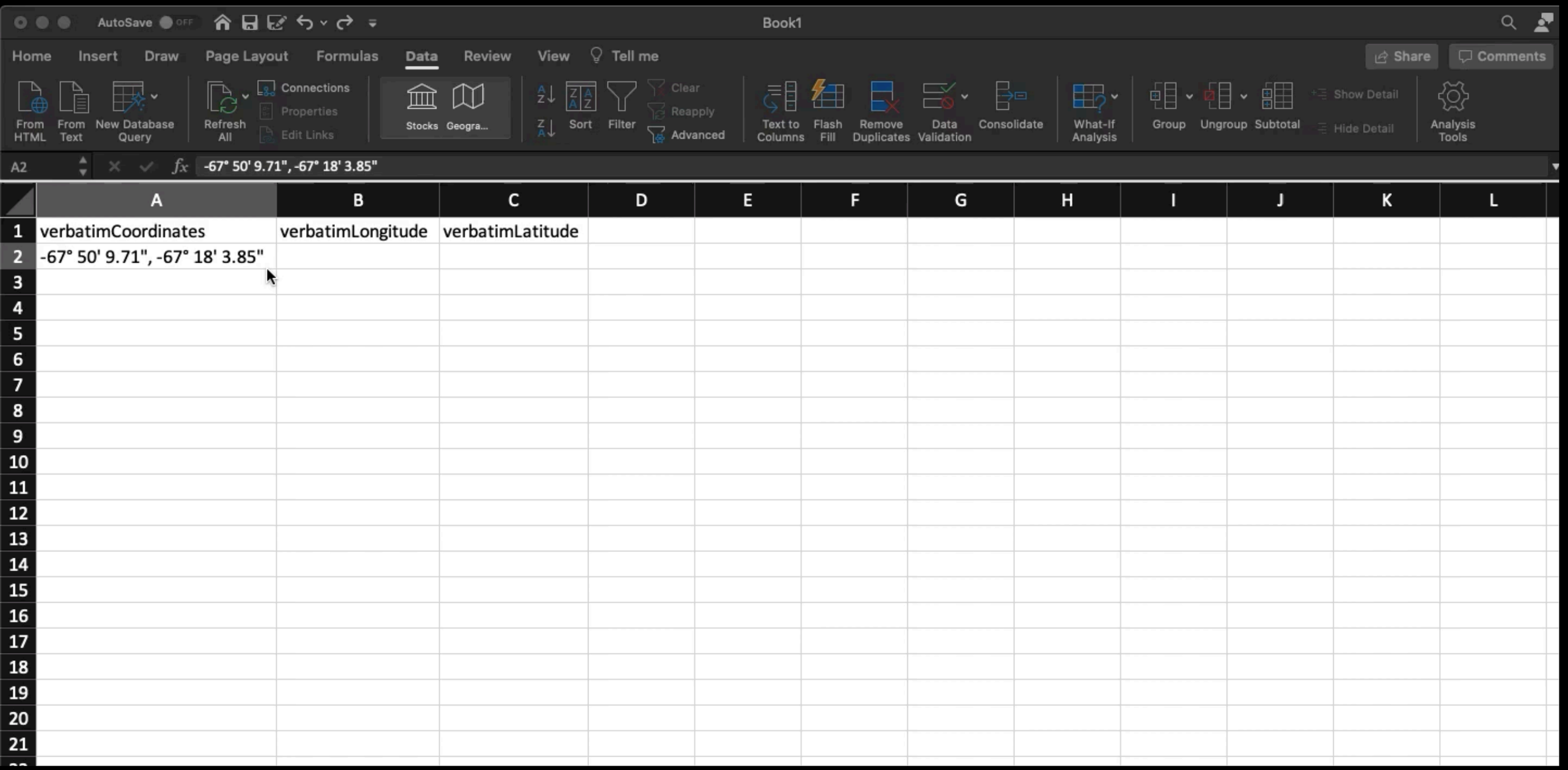
Atomize columns

To ease downstream batch operations

verbatimCoordinates
58°28'30''W, 62°7'0,10''S
58°28'42,00''W, 62°7'15,00''S
62°08.0958'S, 058°24,1625'W
62°10.4680'S, 058°25.2137'W
62°09.7174'S, 058°21,5886'W
62°11.2075'S, 058°18,9951W
62°11.9764'S, 058°22.5800W
62°12.1942'S, 058°23.4483'W
62°53.6407'S, 58°26.8232'W

Atomize columns

Excel: Split text to columns



Always keep original data

So that you can trace it back

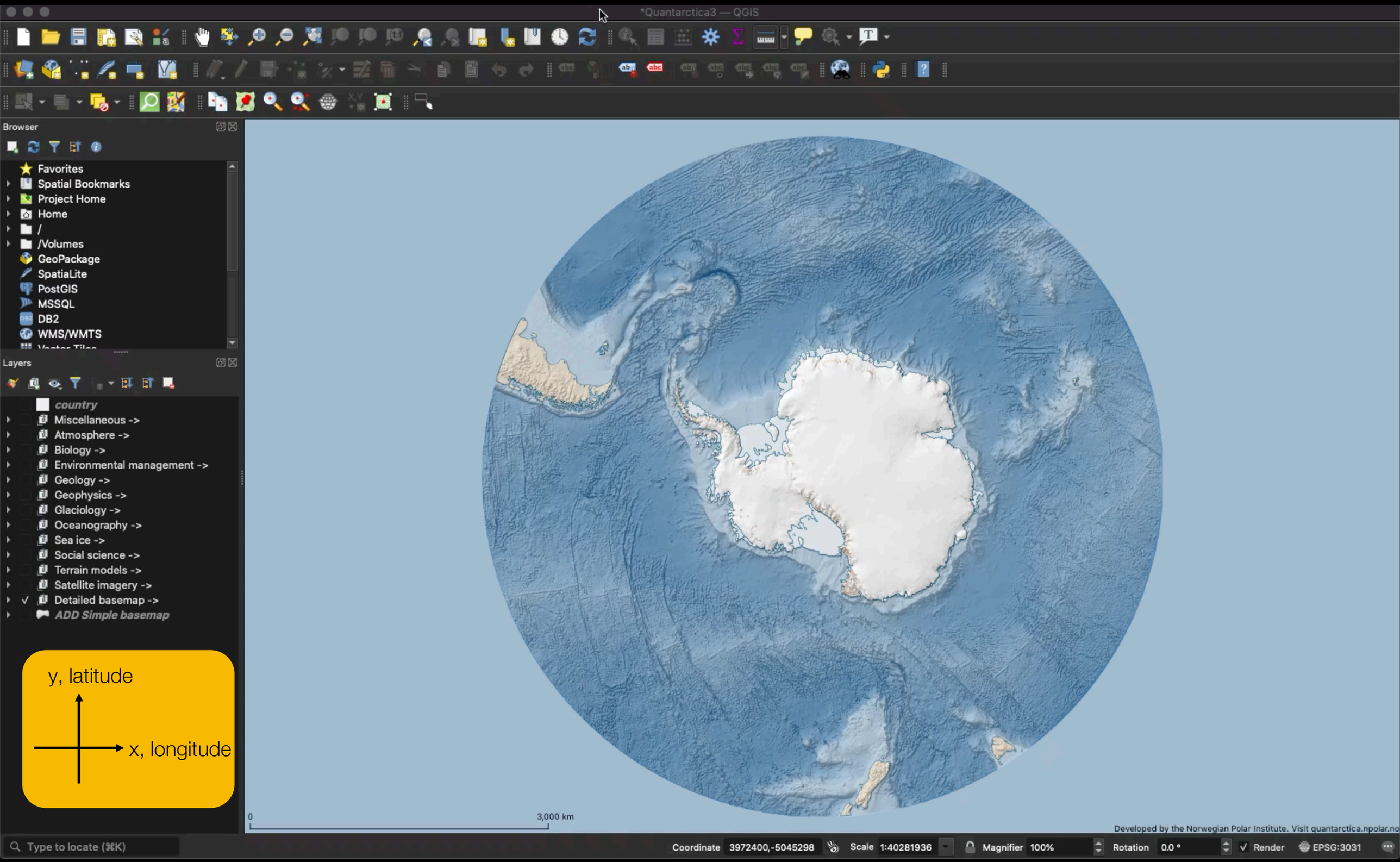
Modified structure

verbatimCoordinates	verbatimLongitude	verbatimLatitude	**decimal**Longitude	**decimal**Latitude
-67° 50' 9.71", -67° 18' 3.85"	-67° 50' 9.71"	-67° 18' 3.85"	-67.83603	-67.30107

Provided value

Modified values

Visualize with Quantarctica



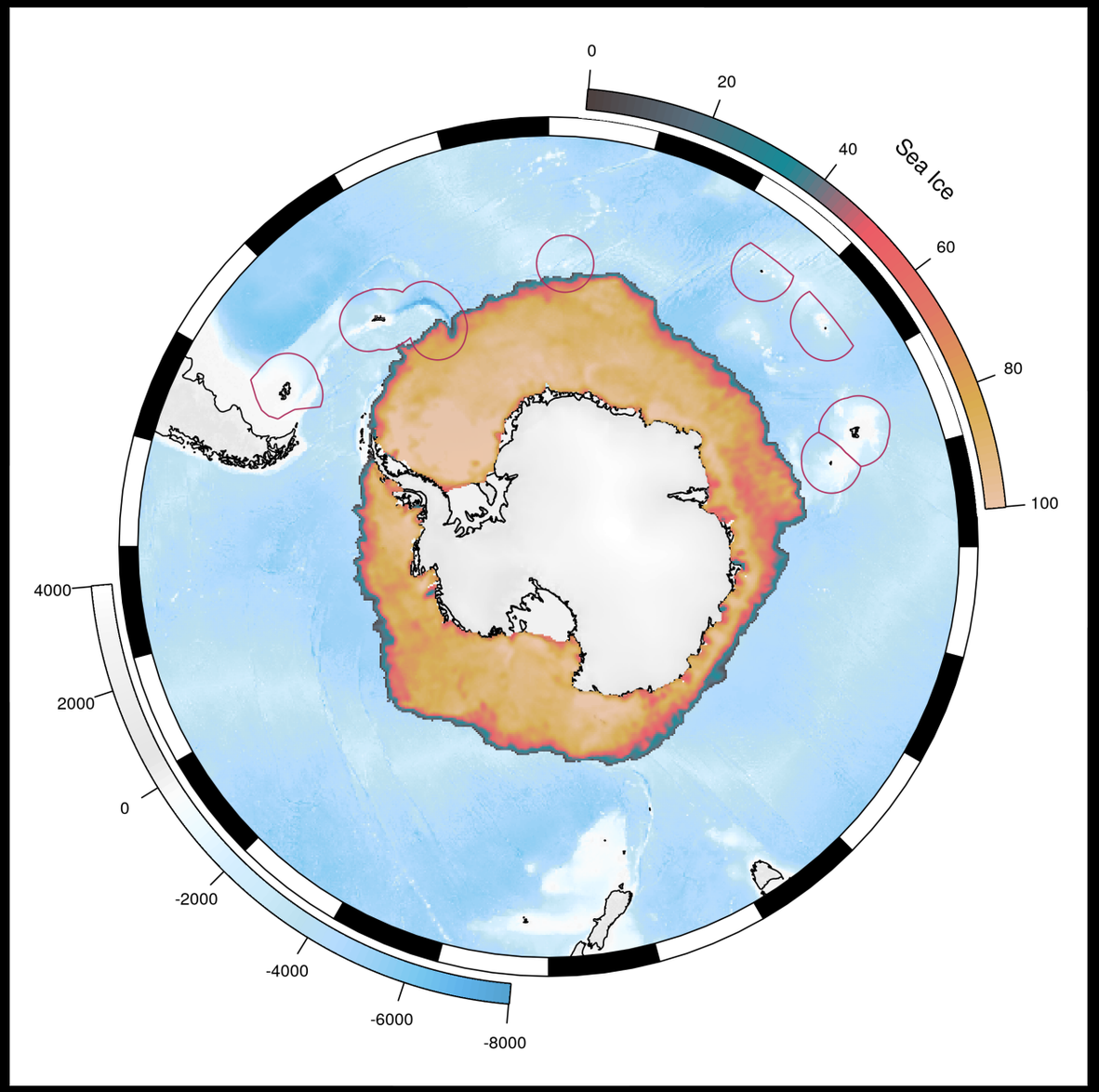
QGIS: https://www.qgis.org/en/site/, Quantarctica: https://www.npolar.no/quantarctica/

R packages

SOmap: Plot Antarctic map effortlessly

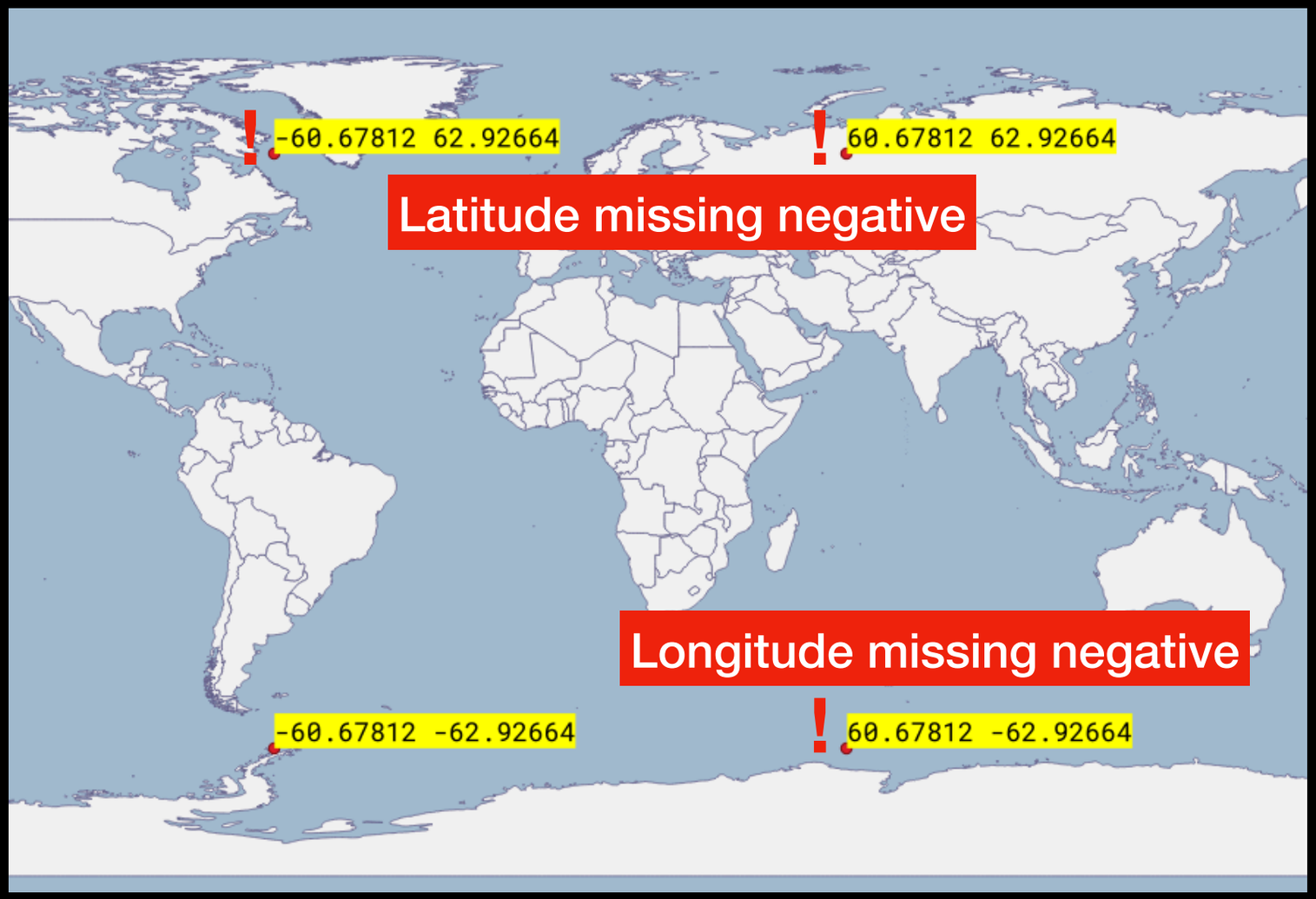
Development phase

SOmap: https://github.com/AustralianAntarcticDivision/SOmap



Visualize data to detect outliers

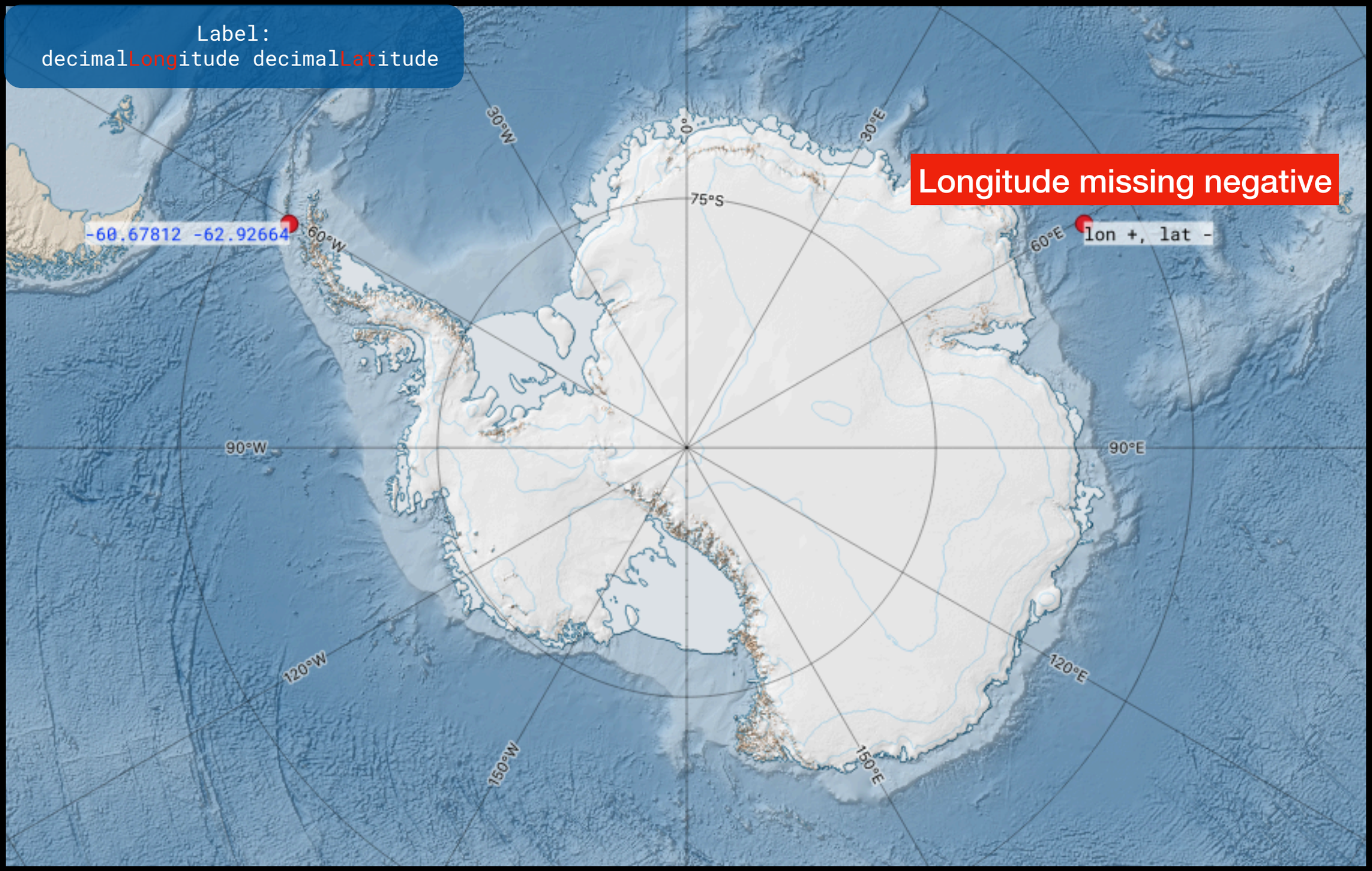
decimalLongitude	decimalLatitude
-60.67812	-62.92664
! 60.67812	-62.92664
! -60.67812	62.92664
! 60.67812	62.92664



Label:
decimalLongitude decimalLatitude

Visualize data to detect outliers

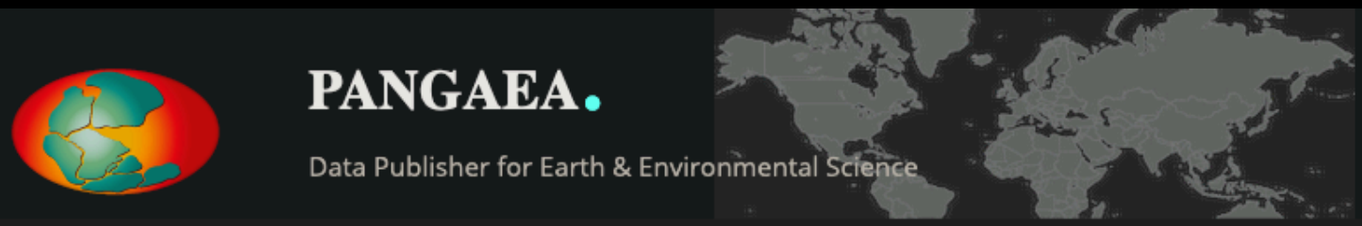
Negative sign of latitude and longitude



Visualize data to detect outliers

Negative sign of latitude and longitude

Cross check data with metadata



PANGAEA.

Data Publisher for Earth & Environmental Science

Not logged in

SEARCH SUBMIT ABOUT CONTACT

Event List of PS118

Download as tab-delimited text

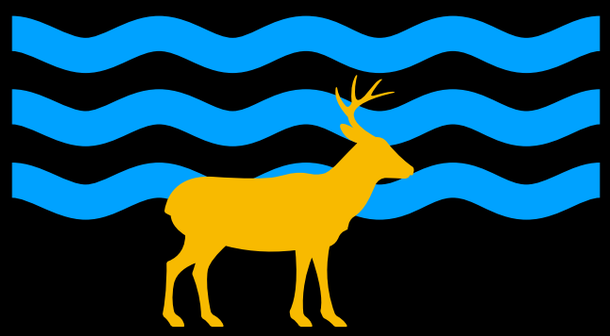
Event label	Optional label	Method/Device	Sensor URI	Date/Time	Latitude	Longitude	Elevation	Date/Time end	Latitude end	Longitude end	Elevation end	Comment
PS118-track		Underway cruise track measurements		2019-02-09T00:00:00	-53.14470	-70.90910		2019-04-08T00:00:00	-53.14470	-70.90910		Punta Arenas - Punta Arenas
PS118_0_underway-12		Ship Weather Station		2019-02-18T00:04:00	-53.12538	-70.85843	-9.7	2019-04-08T17:50:00	-53.18601	-70.90040	-12.8	
PS118_0_underway-14		HydroSweep		2019-02-20T08:50:00	-57.89118	-61.54603	-1.6	2019-04-05T15:02:00	-57.64487	-60.12226	-3012	
PS118_0_underway-11		Thermosalinograph		2019-02-20T08:54:00	-57.89708	-61.53389		2019-04-05T15:01:00	-57.64543	-60.12109	-3012	Keel 2
PS118_0_underway-10		Thermosalinograph		2019-02-20T08:54:00	-57.89733	-61.53338		2019-04-05T15:01:00	-57.64579	-60.12033	-3012	Keel
PS118_0_underway-6		Underway pCO2 measurements		2019-02-20T08:55:00	-57.89832	-61.53160		2019-02-25T12:00:00	-65.14284	-57.17910	-463	
PS118_0_underway-9		Sound velocity profiler		2019-02-20T08:55:00	-57.89790	-61.53234		2019-04-05T15:01:00	-57.64626	-60.11935	-3012	
PS118_0_underway-5		Underway pCO2 measurements		2019-02-20T08:56:00	-57.89883	-61.53069		2019-02-25T12:00:00	-65.14284	-57.17910	-463	
PS118_0_underway-3		FerryBox		2019-02-20T08:56:00	-57.89918	-61.52994		2019-04-05T15:00:41	-57.64814	-60.11549	-3012.1	
PS118_0_underway-8		Gravimetry		2019-02-20T09:02:00	-57.90768	-61.51589		2019-04-05T15:00:00	-57.64860	-60.11458	-3012	
PS118_0_underway-13		Magnetometer		2019-02-20T09:03:00	-57.90864	-61.51411	-3240	2019-04-05T15:00:00	-57.64918	-60.11342	-3012	
PS118_0_underway-1		Vessel mounted Acoustic Doppler Current Profiler 150 kHz	https://hdl.handle.net/10013/sensor.7bdbc478-a4c9-4eb9-917d-9a89ee098774#subItemID=625&subItemEventID=3634	2019-02-20T09:15:00	-57.92407	-61.48704	-2904	2019-04-05T15:00:00	-57.64980	-60.11216	-3039	

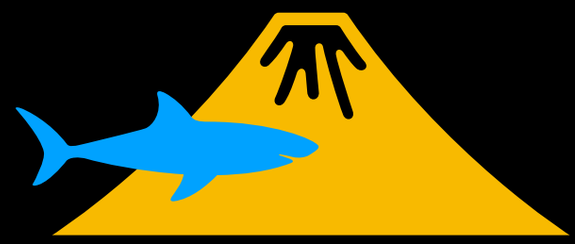
Example cruise events of expedition PS118:
https://www.pangaea.de/expeditions/events/PS118

Geographic coordinate

Points on land or in water

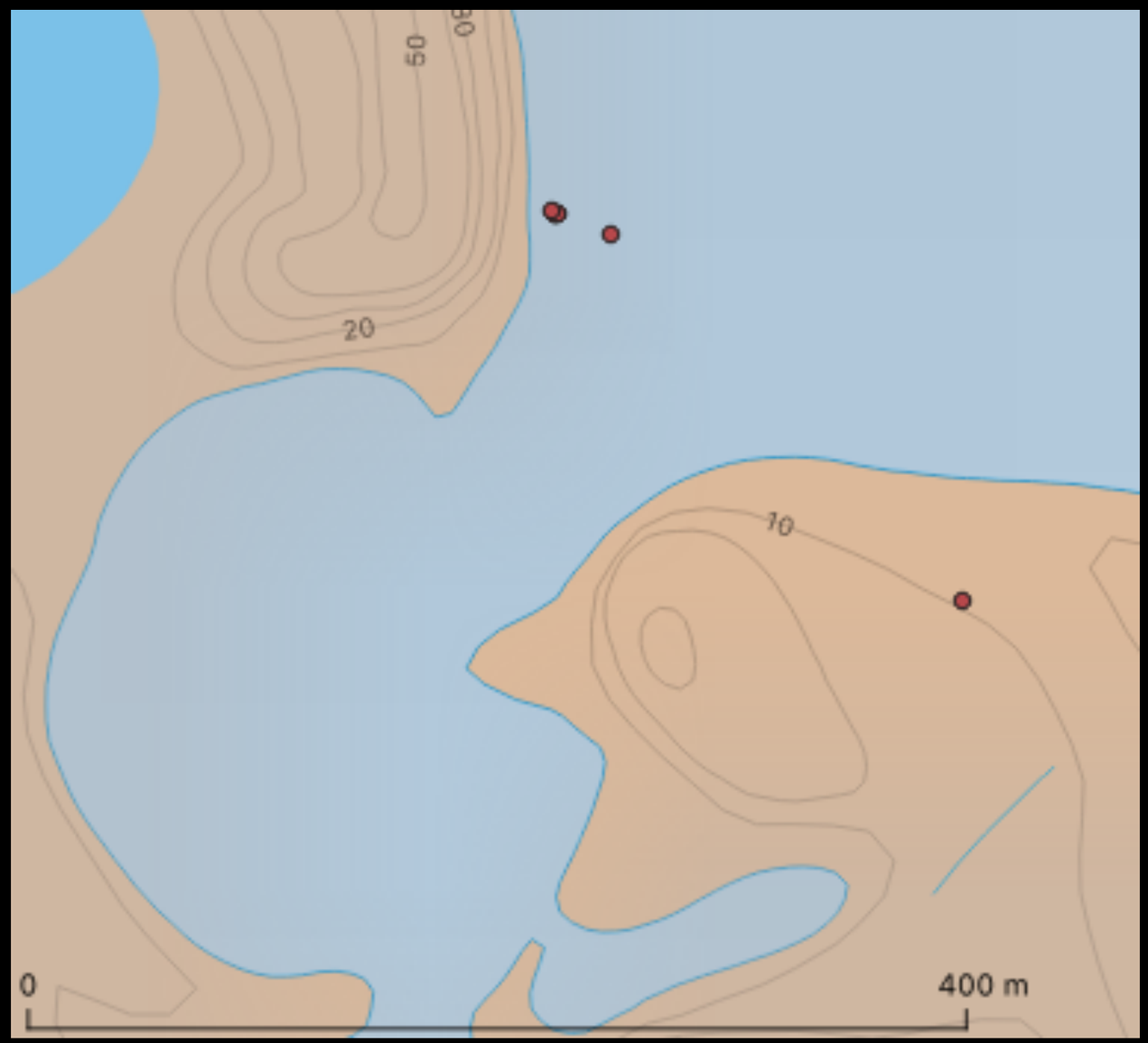
Terrestrial species in water, marine species on land





Check if points are on land or in water

Visualize data points with QGIS

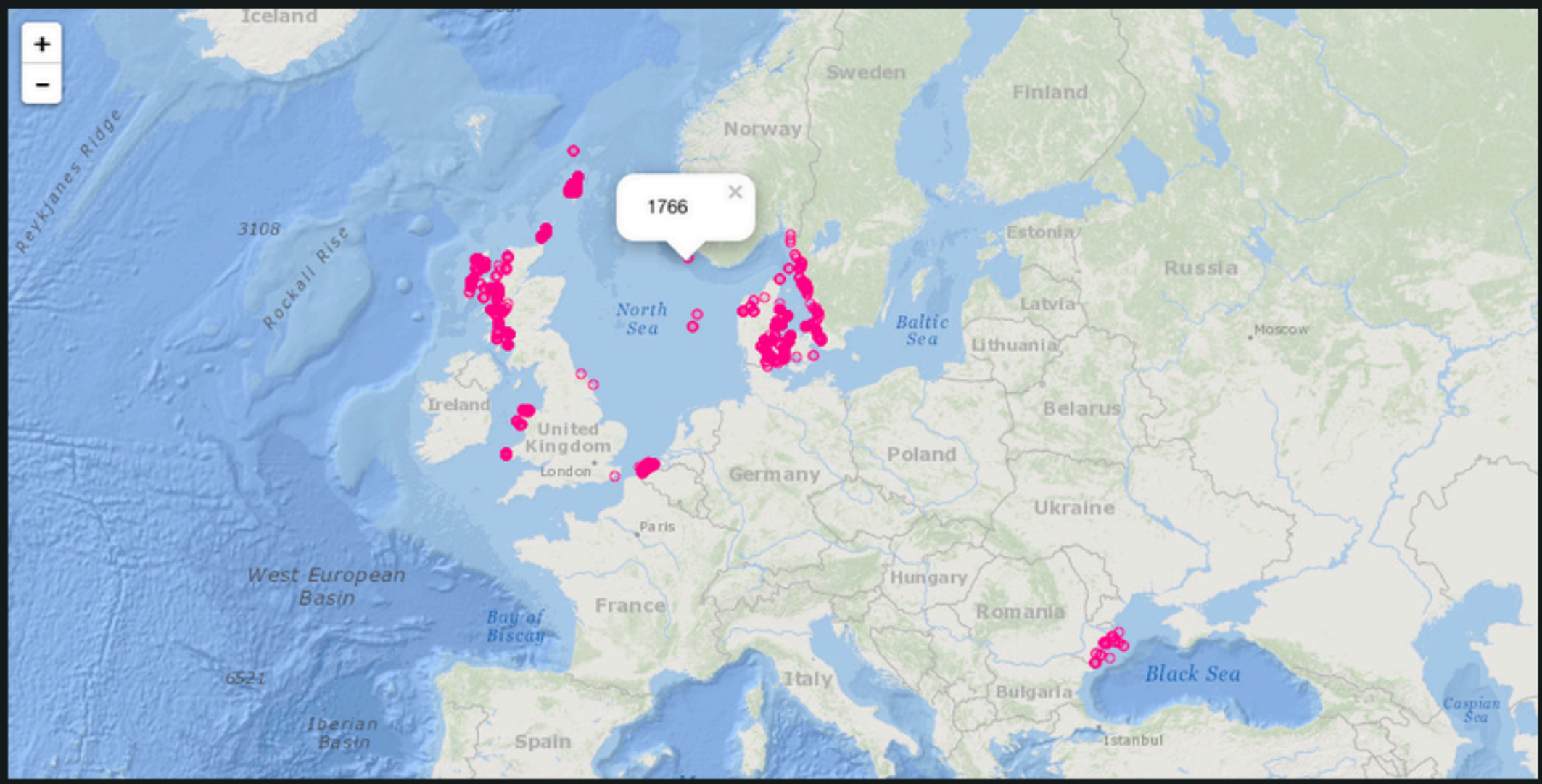


Software: QGIS 3.14, Layer: Quantarctica 3 - Detailed basemap, Antarctic polar stereographic projection

R packages

Plot points using obistools

```
plot_map_leaflet(abra)
```



https://github.com/iobis/obistools

Check if points are on land or in water

Use land polygons (obistools)

Check points on land

`check_onland()` uses the xylookup web service which internally uses land polygons from OpenStreetMap to check if any points are located on land. Other shapefiles can be used as well.

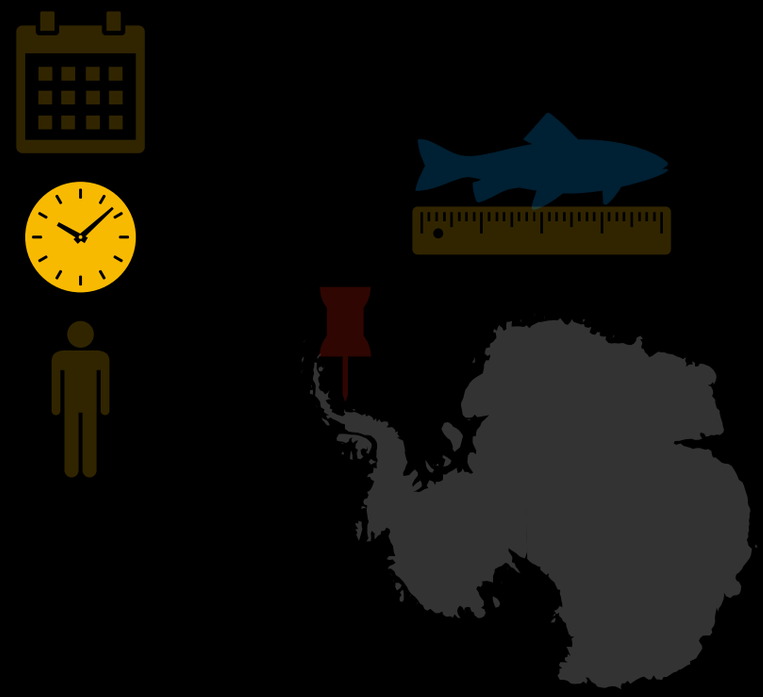
```
check_onland(abra)
```

```
          id decimalLongitude decimalLatitude basisOfRecord           eventDate
31 365512845       -0.9092748        54.57467    Occurrence 2011-09-03 10:00:00
                                  institutionCode collectionCode catalogNumber
31 Yorkshire Naturalists' Union Marine and Coastal Se          60051     261729389 Skinningrove. Cati
                                                  datasetName   phylum    order    family genus s
31 Yorkshire Naturalists Union Marine and Coastal Section Records Mollusca Cardiida Semelidae  Abra
   originalScientificName scientificNameAuthorship obisID resourceID yearcollected   species
31             Abra alba          (W. Wood, 1802) 395450       3083          2011 Abra alba 1073216
   speciesID continent coordinateUncertaintyInMeters       datasetID            modified
31    395450    Europe                         707.0 IMIS:dasid:3182 2014-04-16 16:16:43
                                                                 occurrenceID    recordedBy
31 urn:catalog:Yorkshire Naturalists' Union Marine and Coastal Se:60051:261729389 Adrian Norris
                           scientificNameID    class lifestage  sex individualCount eventID depth
31 urn:lsid:marinespecies.org:taxname:141433 Bivalvia      <NA> <NA>             NA    <NA>    NA
   minimumDepthInMeters maximumDepthInMeters fieldNumber occurrenceRemarks eventTime footprintWKT ide
31                   NA                   NA        <NA>              <NA>      <NA>         <NA>
```

```
check_onland(abra, report = TRUE)
```

```
  field   level row                         message
1    NA warning  31 Coordinates are located on land
```

obistools: https://github.com/iobis/obistools



Where, When, What, Who, How

Date & Time

ISO 8601 standard

Date

Confusing information

Often due software settings, habit

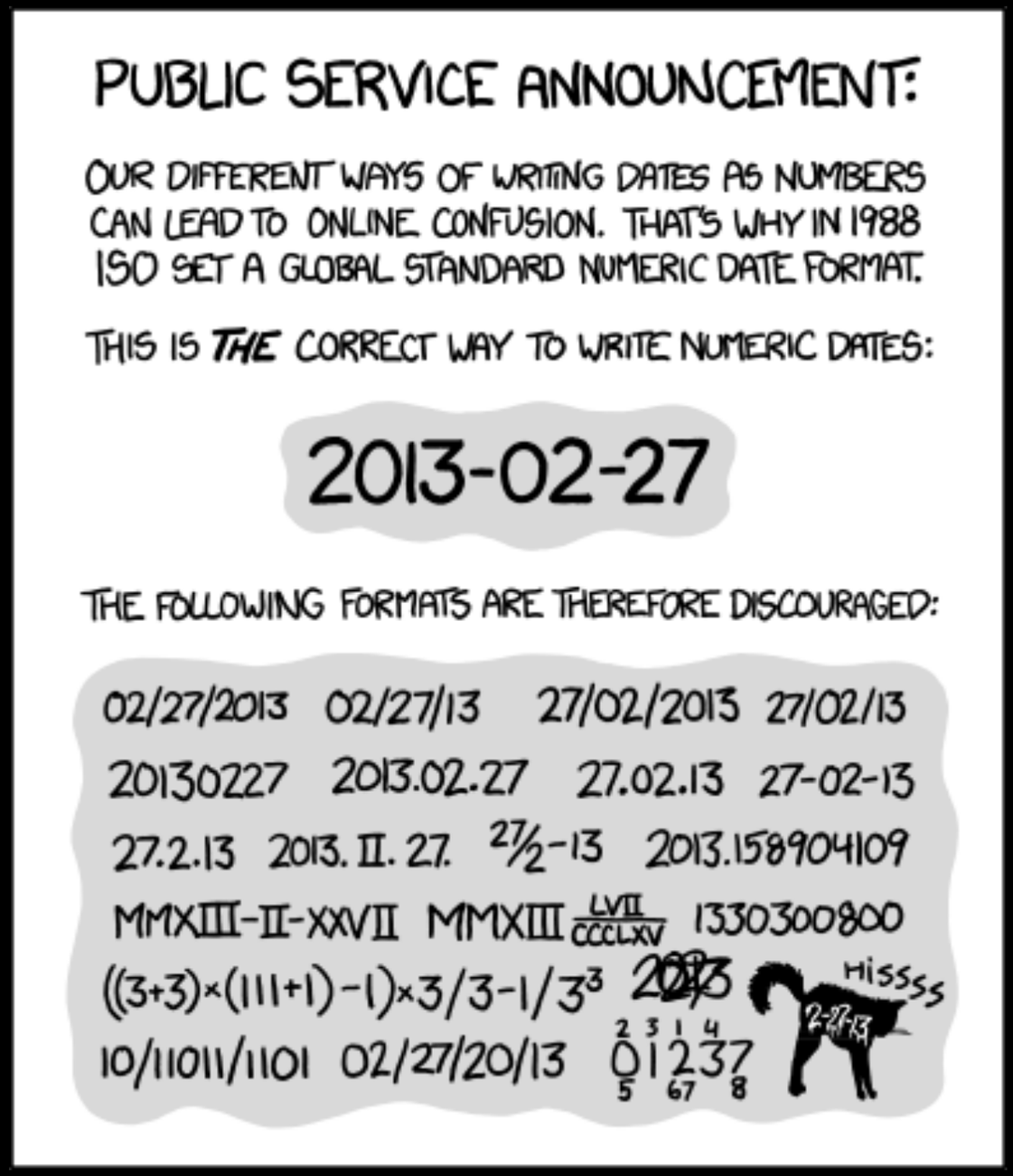
date
1/7/20
1/7/20
1/8/20
1/9/20
1/9/20

DD / MM / YY ?

MM / DD / YY ?

Date

ISO 8601 standard



https://xkcd.com/1179/

Verify information with metadata

e.g. refer to cruise events

PANGAEA.

Data Publisher for Earth & Environmental Science

Not logged in

SEARCH SUBMIT ABOUT CONTACT

Event List of PS118

Download as tab-delimited text

Event label	Optional label	Method/Device	Sensor URI	Date/Time	atitude	Longitude	Elevation	Date/Time end	Latitude end	Longitude end	Elevation end	Comment
PS118-track		Underway cruise track measurements		2019-02-09T00:00:00	3.14470	-70.90910		2019-04-08T00:00:00	-53.14470	-70.90910		Punta Arenas - Punta Arenas
PS118_0_underway-12		Ship Weather Station		2019-02-18T00:04:00	3.12538	-70.85843	-9.7	2019-04-08T17:50:00	-53.18601	-70.90040	-12.8	
PS118_0_underway-14		HydroSweep		2019-02-20T08:50:00	7.89118	-61.54603	-1.6	2019-04-05T15:02:00	-57.64487	-60.12226	-3012	
PS118_0_underway-11		Thermosalinograph		2019-02-20T08:54:00	7.89708	-61.53389		2019-04-05T15:01:00	-57.64543	-60.12109	-3012	Keel 2
PS118_0_underway-10		Thermosalinograph		2019-02-20T08:54:00	7.89733	-61.53338		2019-04-05T15:01:00	-57.64579	-60.12033	-3012	Keel
PS118_0_underway-6		Underway pCO2 measurements		2019-02-20T08:55:00	7.89832	-61.53160		2019-02-25T12:00:00	-65.14284	-57.17910	-463	
PS118_0_underway-9		Sound velocity profiler		2019-02-20T08:55:00	7.89790	-61.53234		2019-04-05T15:01:00	-57.64626	-60.11935	-3012	
PS118_0_underway-5		Underway pCO2 measurements		2019-02-20T08:56:00	7.89883	-61.53069		2019-02-25T12:00:00	-65.14284	-57.17910	-463	
PS118_0_underway-3		FerryBox		2019-02-20T08:56:00	7.89918	-61.52994		2019-04-05T15:00:41	-57.64814	-60.11549	-3012.1	
PS118_0_underway-8		Gravimetry		2019-02-20T09:02:00	7.90768	-61.51589		2019-04-05T15:00:00	-57.64860	-60.11458	-3012	
PS118_0_underway-13		Magnetometer		2019-02-20T09:03:00	7.90864	-61.51411	-3240	2019-04-05T15:00:00	-57.64918	-60.11342	-3012	
PS118_0_underway-1		Vessel mounted Acoustic Doppler Current Profiler 150 kHz	https://hdl.handle.net/10013/sensor.7bdbc478-a4c9-4eb9-917d-9a89ee098774#subItemID=625&subItemEventID=3634	2019-02-20T09:15:00	7.92407	-61.48704	-2904	2019-04-05T15:00:00	-57.64980	-60.11216	-3039	

Example cruise events of expedition PS118:

https://www.pangaea.de/expeditions/events/PS118

R package for handling date-time data

lubridate

part of the tidyverse

1.7.9.9000

Features

```
library(lubridate, warn.conflicts = FALSE)
```

- Easy and fast parsing of date-times: `ymd()`, `ymd_hms`, `dmy()`, `dmy_hms`, `mdy()`, ...

```
ymd(20101215)
#> [1] "2010-12-15"
mdy("4/1/17")
#> [1] "2017-04-01"
```

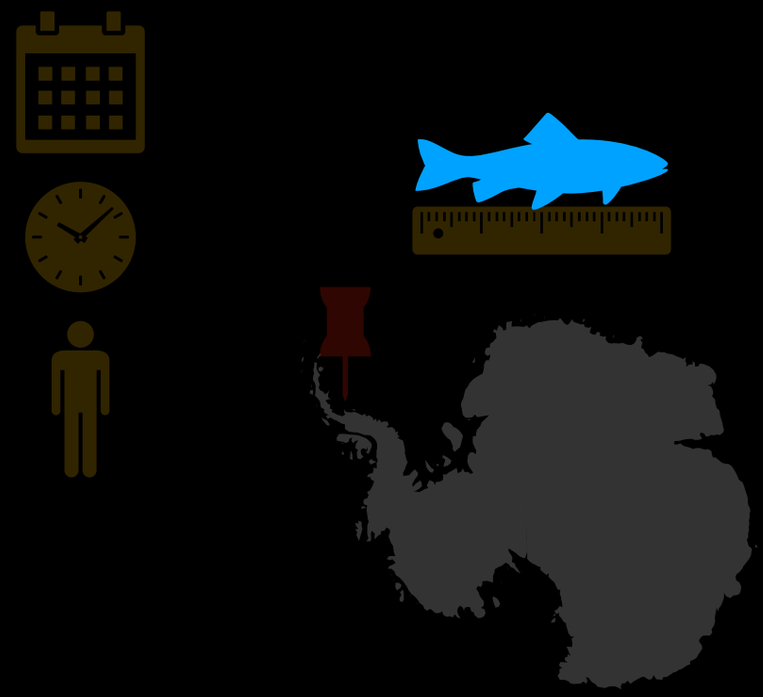
- Simple functions to get and set components of a date-time, such as `year()`, `month()`, `mday()`, `hour()`, `minute()` and `second()`:

```
bday <- dmy("14/10/1979")
month(bday)
#> [1] 10
wday(bday, label = TRUE)
#> [1] Sun
#> Levels: Sun < Mon < Tue < Wed < Thu < Fri < Sat

year(bday) <- 2016
wday(bday, label = TRUE)
#> [1] Fri
#> Levels: Sun < Mon < Tue < Wed < Thu < Fri < Sat
```

- Helper functions for handling time zones: `with_tz()`, `force_tz()`

https://lubridate.tidyverse.org/index.html



Where, When, What, Who, How

Taxonomy

Taxonomy

Homonyms - same name, different taxon

Only 1 of the 2 names can stay “valid”,
the other becomes “invalid”.

Genus: *Alebion*

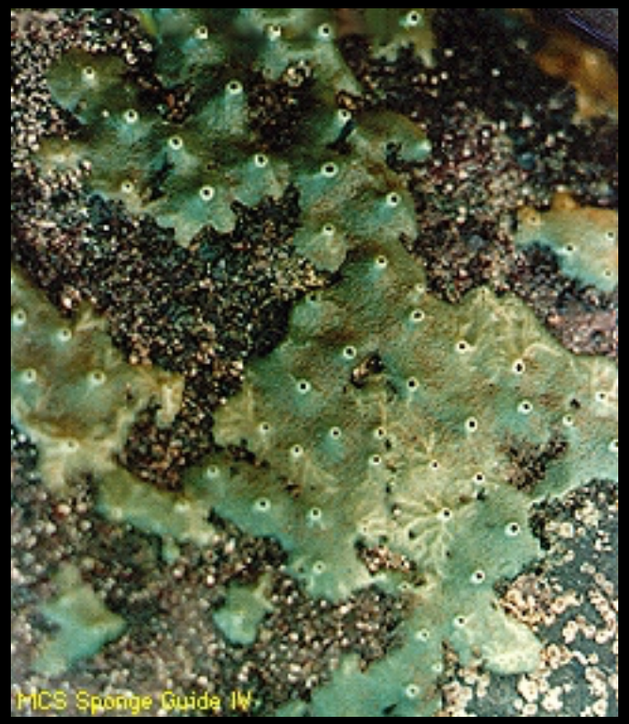
Alebion Krøyer, 1863
- Animalia, Crustacea
- parasitic copepods

Alebion Gray, 1867
- Animalia, Porifera
- Accepted as *Iophon Gray, 1867*

Vandepitte, L. (2017) Biodiversity data standards: Taxonomy. OBIS nodes training course.
https://classroom.oceanteacher.org/mod/resource/view.php?id=7439

Taxonomy

Synonyms - different names, same taxon



"Halichondria panicea" by Rob van Soest CC BY-NC-SA 4.0.

Halichondria (Halichondria) panicea (Pallas, 1766)

> 60 synonyms

- *Hymeniacidon fallaciosus* Bowerbank, 1866 (genus transfer and junior synonym)
- *Hymeniacidon firmus* Bowerbank, 1874 (genus transfer)
- *Hymeniacidon fragilis* Bowerbank, 1866 (genus transfer and junior synonym)
- *Hymeniacidon lactea* Bowerbank, 1866 (genus transfer and junior synonym)
- *Hymeniacidon membrana* Bowerbank, 1866 (genus transfer and junior synonym)
- *Hymeniacidon parfitti* Parfitt, 1868 (genus transfer and junior synonym)
- *Hymeniacidon reticulatus* Bowerbank, 1866 (genus transfer and junior synonym)
- *Hymeniacidon solidus* Bowerbank, 1882 (genus transfer and junior synonym)
- *Hymeniacidon tegeticula* Bowerbank, 1874 (genus transfer and junior synonym)
- *Hymeniacidon thomasii* Bowerbank, 1866 (genus transfer and junior synonym)
- *Isodictya crassa* Bowerbank, 1882 (genus transfer and junior synonym)
- *Isodictya perplexa* Bowerbank, 1882 (genus transfer and junior synonym)
- *Menanetia minchini* Topsent, 1896 (genus transfer and junior synonym)
- *Microciona tumulosa* Bowerbank, 1882 (genus transfer and junior synonym)
- *Pellina bibula* Schmidt, 1870 (genus transfer and junior synonym)
- *Scypha papillaris f. sowerbeii* Gray, 1821 (genus transfer and junior synonym)
- *Seriatula seriata* (Grant, 1826) (genus transfer and junior synonym)
- *Spongia albescens* Rafinesque, 1818 (genus transfer and junior synonym)
- *Spongia compacta* Sowerby, 1806 (genus transfer and junior synonym)
- *Spongia cristata* Ellis & Solander, 1786 (genus transfer and junior synonym)
- *Spongia panicea* Pallas, 1766 (genus transfer)
- *Spongia seriata* Grant, 1826 (genus transfer and junior synonym)
- *Spongia tomentosa* Linnaeus, 1767 (genus transfer and junior synonym)
- *Spongia urens* Ellis & Solander, 1786 (genus transfer and junior synonym)
- *Spuma borealis var. convoluta* Miklucho-Maclay, 1870 (genus transfer & junior synonym)
- *Spuma borealis var. tuberosa* Miklucho-Maclay, 1870 (genus transfer & junior synonym)
- *Alcyonium manusdiaboli* sensu Esper, 1794 (genus transfer and junior synonym)
- *Alcyonium medullare* Lamarck, 1815 (genus transfer & junior synonym)
- *Alcyonium paniceum* (Pallas, 1766) (genus transfer)
- *Amorphina appendiculata* Schmidt, 1875 (genus transfer and junior synonym)
- *Amorphina coccinea* (Bowerbank, 1861) (genus transfer & junior synonym)
- *Amorphina grisea* Fristedt, 1887 (genus transfer and junior synonym)
- *Amorphina paciscens* Schmidt, 1875 (genus transfer and junior synonym)
- *Amorphina panicea* (Pallas, 1766) (genus transfer)
- *Clathria (Microciona) seriata* (Grant, 1826) (genus transfer and junior synonym)
- *Clathria (Microciona) tumulosa* (Bowerbank, 1882) (genus transfer and junior synonym)
- *Clathria seriata* (Grant, 1826) (genus transfer and junior synonym)
- *Eumastia appendiculata* (Schmidt, 1875) (genus transfer and junior synonym)
- *Halichondria albescens* (Rafinesque, 1818) (junior synonym)
- *Halichondria ambigua* Bowerbank, 1874· accepted, alternate representation (junior synonym)
- *Halichondria bibula* (Schmidt, 1870) (junior synonym)
- *Halichondria brettii* (Bowerbank, 1866)· accepted, alternate representation (subgenus assignment)
- *Halichondria caduca* Bowerbank, 1866· accepted, alternate representation (junior synonym)
- *Halichondria coccinea* Bowerbank, 1861· accepted, alternate representation (junior synonym)
- *Halichondria coralloides* Bowerbank, 1882· accepted, alternate representation (junior synonym)
- *Halichondria edusa* Bowerbank, 1874· accepted, alternate representation (junior synonym)
- *Halichondria firmus* (Bowerbank, 1874)· accepted, alternate representation (junior synonym)
- *Halichondria glabra* Bowerbank, 1866· accepted, alternate representation (junior synonym)
- *Halichondria grisea* (Fristedt, 1887)· accepted, alternate representation (junior synonym)
- *Halichondria incerta* Bowerbank, 1866· accepted, alternate representation (junior synonym)
- *Halichondria lactea* (Bowerbank, 1866)· accepted, alternate representation (junior synonym)
- *Halichondria membrana* (Bowerbank, 1866)· accepted, alternate representation (junior synonym)
- *Halichondria paciscens* (Schmidt, 1875)· accepted, alternate representation (junior synonym)
- *Halichondria panacea* (misspelling of species name)
- *Halichondria panicea* (Pallas, 1766)· accepted, alternate representation (subgenus assignment)
- *Halichondria pannosus* Verrill, 1874· accepted, alternate representation (junior synonym)
- *Halichondria papillaris* (Linnaeus, 1791)· accepted, alternate representation (junior synonym)
- *Halichondria reticulata* Lieberkühn, 1859· accepted, alternate representation (junior synonym)
- *Halichondria reticulata* (Bowerbank, 1866) (junior synonym)
- *Halichondria sevosa* Johnston, 1842· accepted, alternate representation (junior synonym)
- *Halichondria topsenti* Laubenfels, 1936· accepted, alternate representation (junior synonym)
- *Halichondriella corticata* Burton, 1931 (genus transfer and junior synonym)
- *Halina panicea* (Pallas, 1766) (genus transfer)
- *Halina papillaris* (Pallas, 1766) (genus transfer and junior synonym)
- *Halispongia papillaris* (Pallas, 1766) (genus transfer and junior synonym)
- *Hymeniacidon brettii* Bowerbank, 1866 (genus transfer and junior synonym)
- *Hymeniacidon coccinea* (Bowerbank, 1861) (genus transfer and junior synonym)

Vandepitte, L. (2017) Biodiversity data standards: Taxonomy. OBIS nodes training course. https://classroom.oceanteacher.org/mod/resource/view.php?id=7439

Taxonomy

Taxonomy is constantly being updated

Taxonomy

WoRMS

- Register of Antarctic Species

Global Names Index

Integrated Taxonomic Information System

Encyclopedia of Life

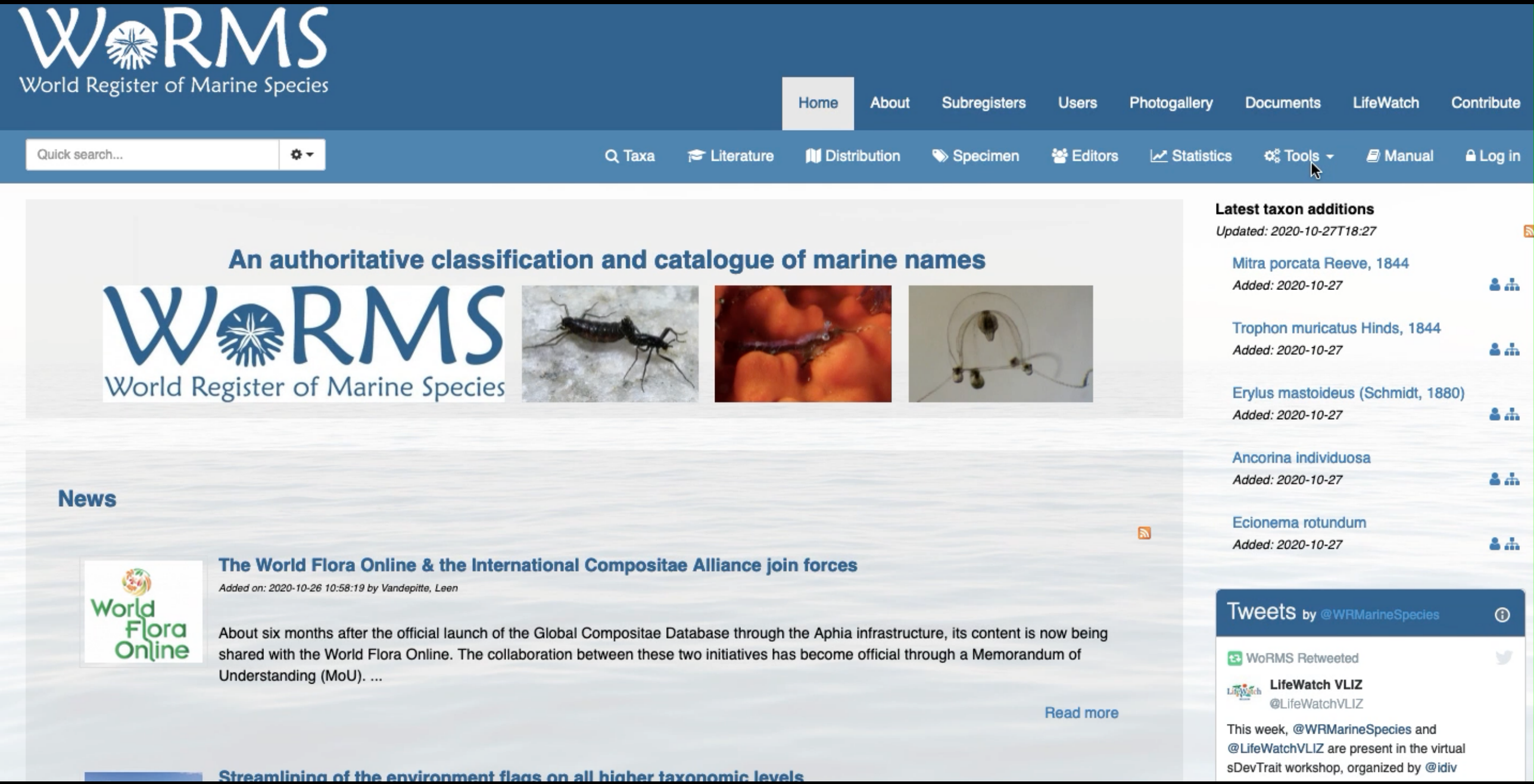
Taxonomy

Map taxa to a backbone - allow taxon from different datasets to be comparable

scientificName
Akanthophoreus antarctica
Akanthophoreus multiserratus

Taxonomy

Match scientific name to taxon using WoRMS taxon match service



World Register of Marine Species. Available from http://www.marinespecies.org at VLIZ. Accessed 2020-10-27

Taxonomy

Excel: Join tables using VLOOKUP

Matched scientific name in WoRMS

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	scientificName		AphiaID	Match typ	LSID	TSN	Qualitysta	Taxon stat	ScientificName	Authority	AphiaID_a	ScientificN	Authority_	Kingdom	Phylum	Cla
2	Akanthophoreus antarctica		448316	near_2	urn:lsid:marinespecies.org:taxname:448316		Checked by	unaccepted	Akanthophoreus antarcticus	(Vanhöffen	798737	Parakanthc	(Vanhöffen	Animalia	Arthropoda	Ma
3	Akanthophoreus multiserratus		136343	exact	urn:lsid:marinespecies.org:taxname:136343		Checked by	unaccepted	Akanthophoreus multiserratus	(Hansen, 1	798748	Parakanthc	(Hansen, 1	Animalia	Arthropoda	Ma
4					5th column				9th column							

File from WoRMS taxon match

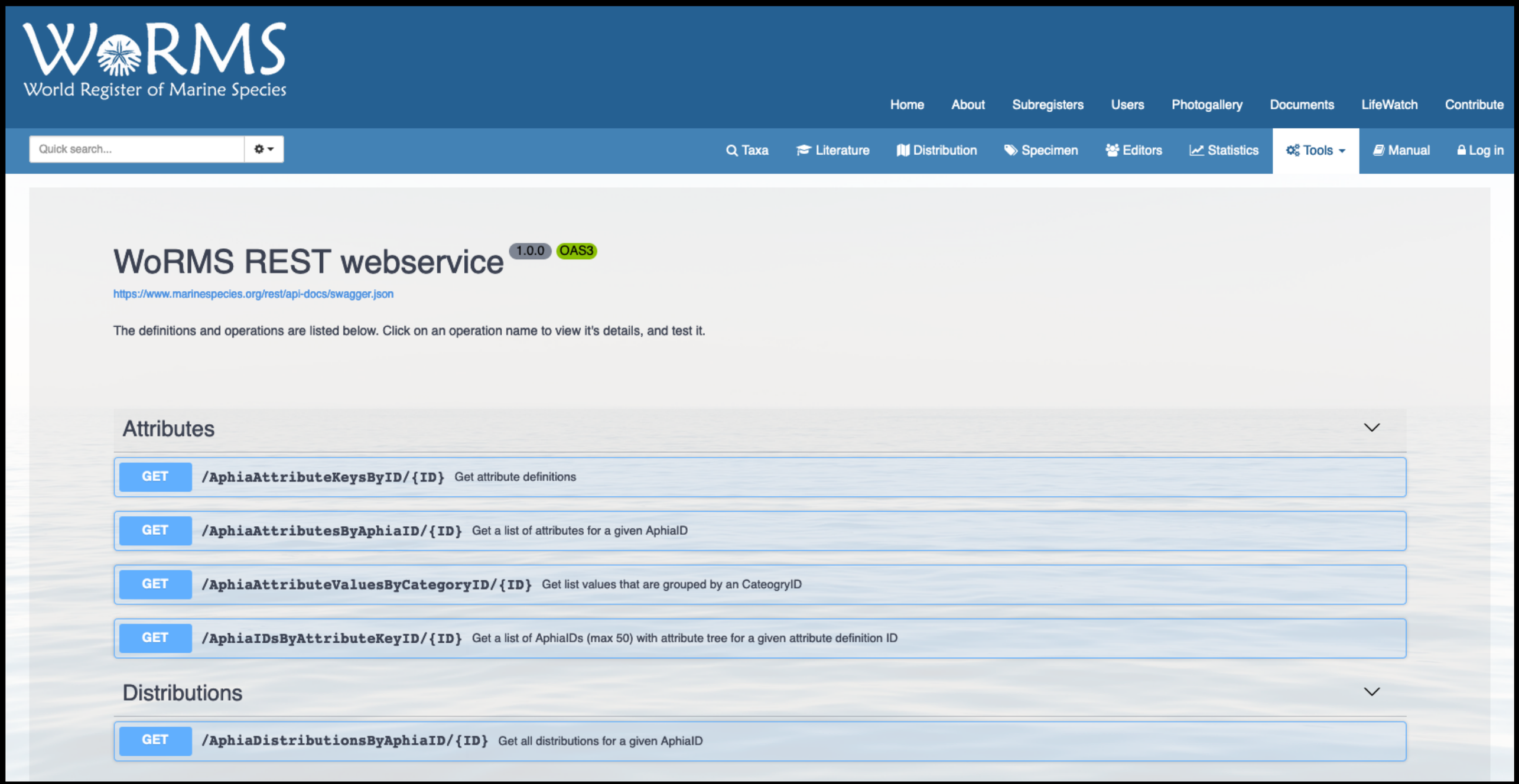
	A	B	C	D	E	F	G	H	I	J	K	L
1		scientificName	scientificNameID									
2	Akanthophoreus antarctica											
3	Akanthophoreus multiserratus											
4												
5												
6		VLOOKUP($A2,'WoRMS match'!$A$2:$I$3,9,FALSE)										
7												
8												
9												
10												
11												

My data file

Sheet 1 cell B2:
Look for the value in A2 cell in the range specified and return the value of column 9 of the row that has exact match with A2

Web services

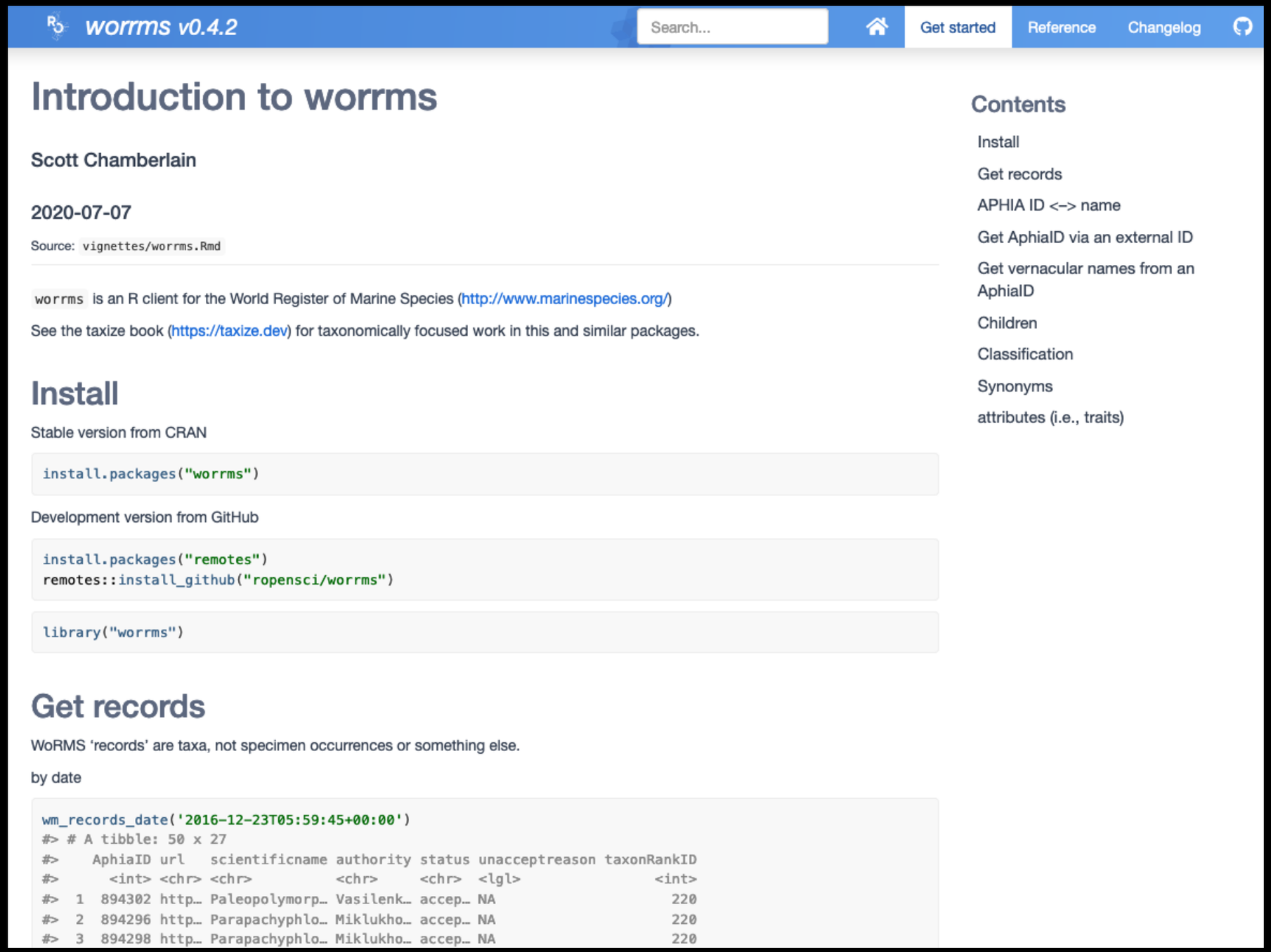
REST service for WoRMS



https://www.marinespecies.org/rest/

R packages

worrms: R client for the World Register of Marine Species



https://docs.ropensci.org/worrms/articles/worrms.html

Taxonomy

Keep original scientific name
Use persistent identifier to refer to the taxon in a checklist

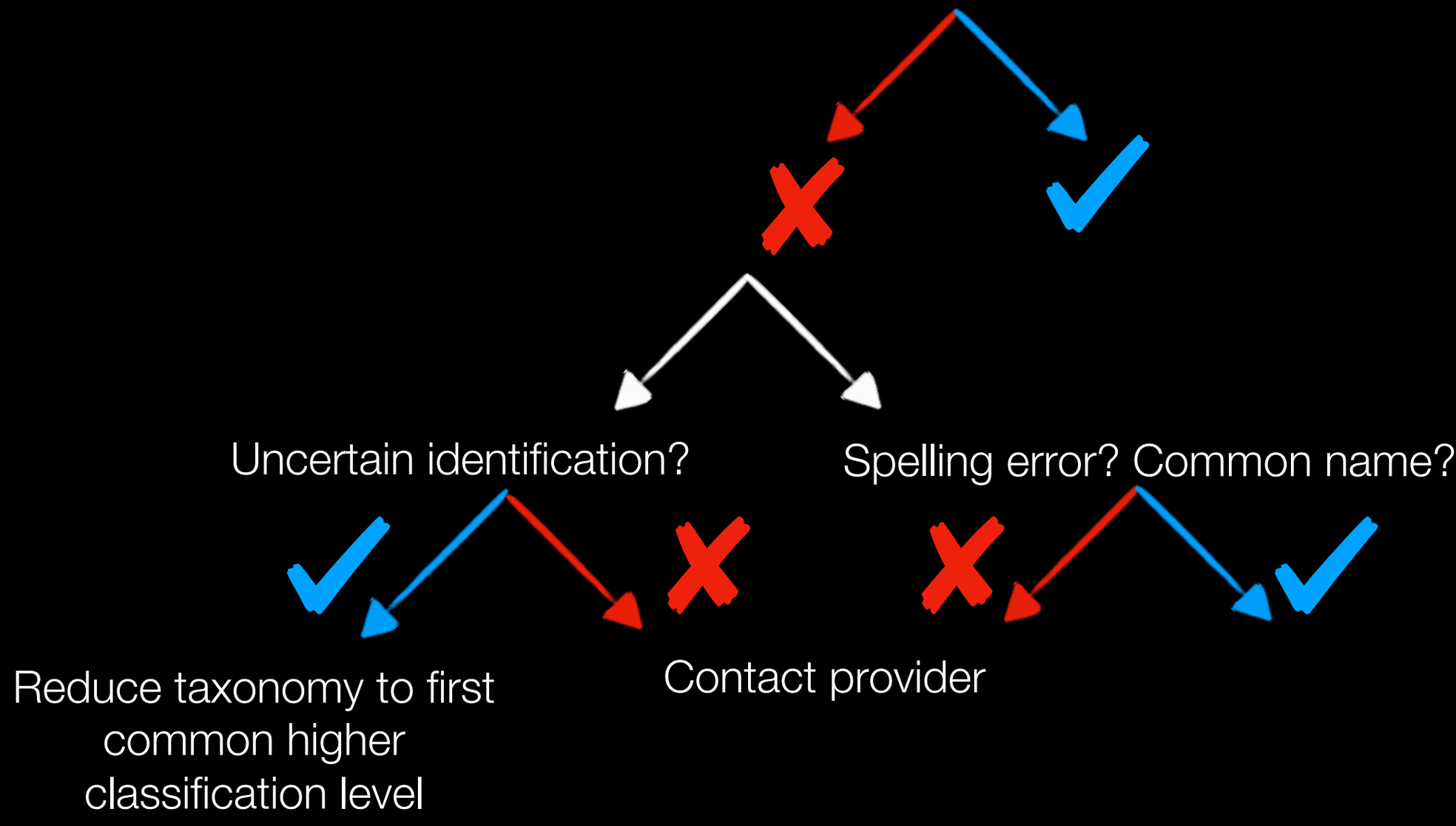
Do not change to accepted scientific name

scientificName	scientificName Authorship	scientificNameID	taxonomicStatus	taxonRank
Akanthophoreus antarcticus	(Vanhöffen, 1914)	urn:lsid:marinespecies.org:taxname:448316	unaccepted	species
Akanthophoreus multiserratus	(Hansen, 1913)	urn:lsid:marinespecies.org:taxname:136343	unaccepted	species

Persistent identifiers

Taxonomy

QC in summary

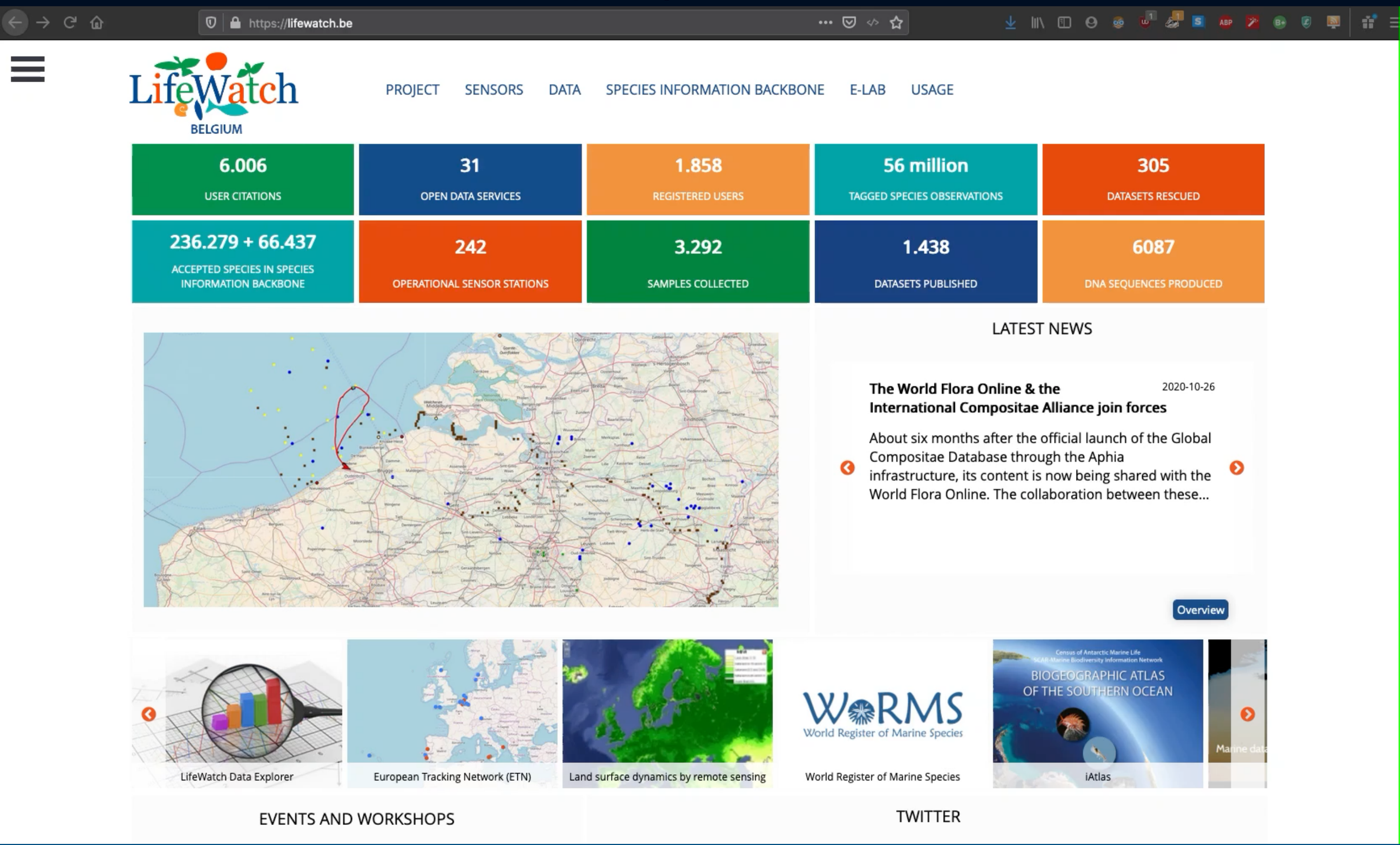


Vandepitte, L. (2017) Biodiversity data standards: Taxonomy. OBIS nodes training course.
https://classroom.oceanteacher.org/mod/resource/view.php?id=7439

Online tool for biodiversity data QC

Lifewatch e-lab services

Online quality control service



Summary

- Be organized
- Avoid changing raw data: set permission to read-only
- Ensure that mandatory fields are filled in with correct values
- Always keep original data
- Track changes – document the WHY
- Atomize columns/values
- Visualize your data
- Where possible transform data with functions
- Cross check data with metadata

Additional resources

- Georeferencing Best Practice from GBIF
  http://mb.gbif.org/documents/doc-georeferencing-best-practices/en/

- Ecology workshop from data carpentry
  https://datacarpentry.org/ecology-workshop/

Acknowledgement

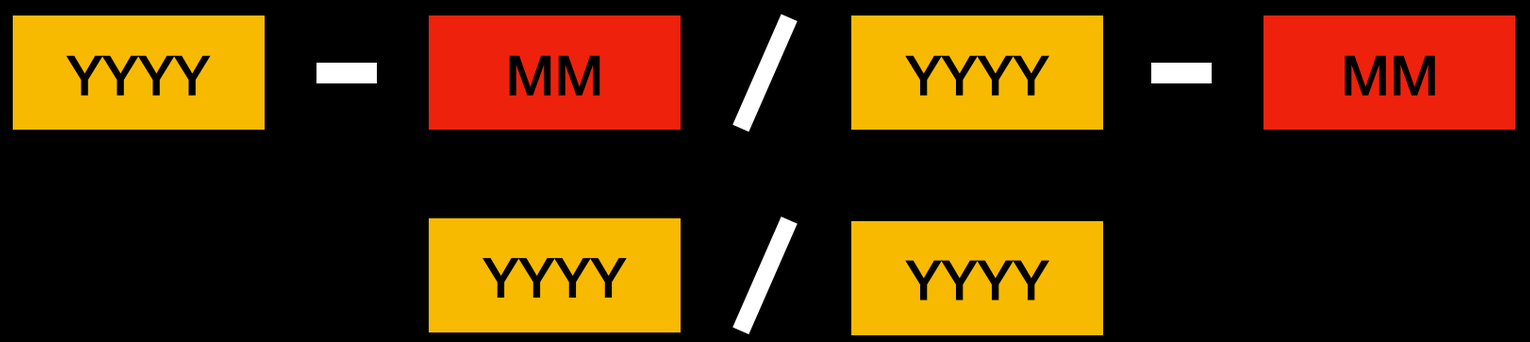
Early career scientists who provided valuable input:

Robyn Samuel, Raissa Meyer, Louraine Salabao,
Jasmine Lee, Kimberlee Bradley, Svenja Hafter

Date range

ISO 8601 standard

Start date / End date



etc etc …

Date range

ISO 8601 standard

This conforms to ISO 8601 standard too

2007 – 01 – 13 / 03 – 15

Some time in the interval between
13 January 2007
and
15 March 2007

Machine readable

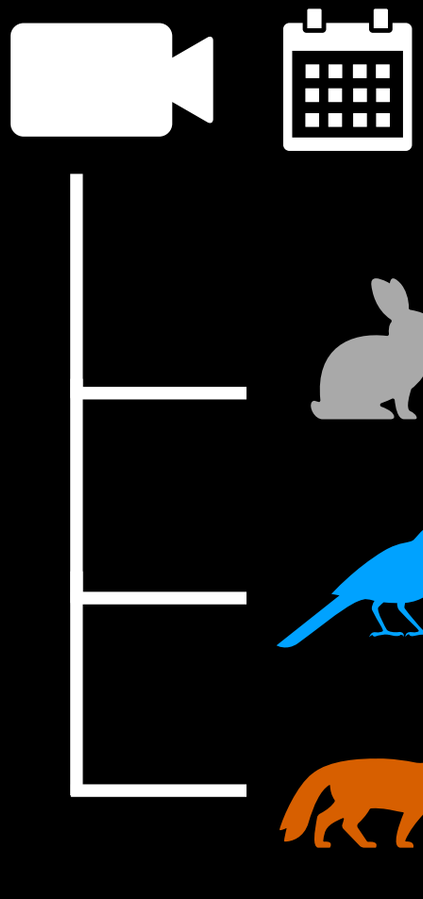
Recover metadata from file names with delimiters

```
> list.files() %>%
    stringr::str_split_fixed("[_\\.]", 4)
     [,1]         [,2]      [,3]             [,4]
[1,] "2018-12-15" "PS117"   "Cape-Town"      "pdf"
[2,] "2019-02-09" "PS118"   "Punta-Arenas"   "pdf"
[3,] "2019-04-13" "PS119"   "Punta-Arenas"   "pdf"
[4,] "2020-06-04" "PS122-4" "Arctic-Ocean"   "pdf"
[5,] "2020-08-12" "PS122-5" "Arctic-Ocean"   "pdf"
```

Date Campaign Location File type

Bryan, J. (2015) How to name files. Reproducible Science Workshop, Duke. https://speakerdeck.com/jennybc/how-to-name-files

eventDate and eventTime



2020-10-01/2020-10-02

2020-10-01T16:00:00Z

2020-10-02T08:00:00Z

2020-10-02T13:00:00Z

eventID	parentEventID	eventDate	eventTime
CAM_1		2020-10-01T15:00:00Z/2020-10-02T16:00:00Z	
CAM_1_2	CAM_1	2020-10-01	16:00:00Z
CAM_1_3	CAM_1	2020-10-02	08:00:00Z
CAM_1_4	CAM_1	2020-10-02	13:00:00Z