Data cleaning & standardisation in R

Laura Márquez

Salvador Fernández

September 27th, 2022

VLIZ pioneer in marine science

LifeWatch BELGIUM

What we will learn in this session

- Importance of clean data
- Intro to standards in biodiversity & its importance
- Exercises!

VLIZ pioneer in marine science

Sources for this session

- R binder Binder (mybinder.org)

- Rstudio

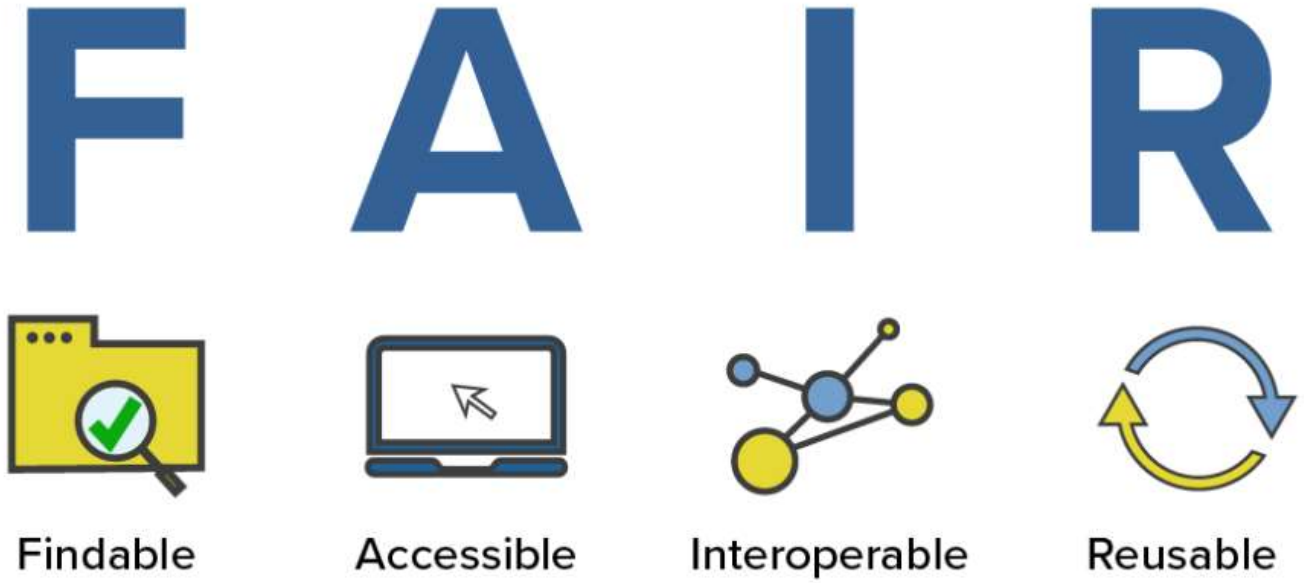
github.com/lifewatch/ebr-2022-data-cleaning-standardization

-> R/Datacleaning_standardization.R

-> data/DataAbundancew.csv

VLIZ pioneer in marine science

FAIR principles

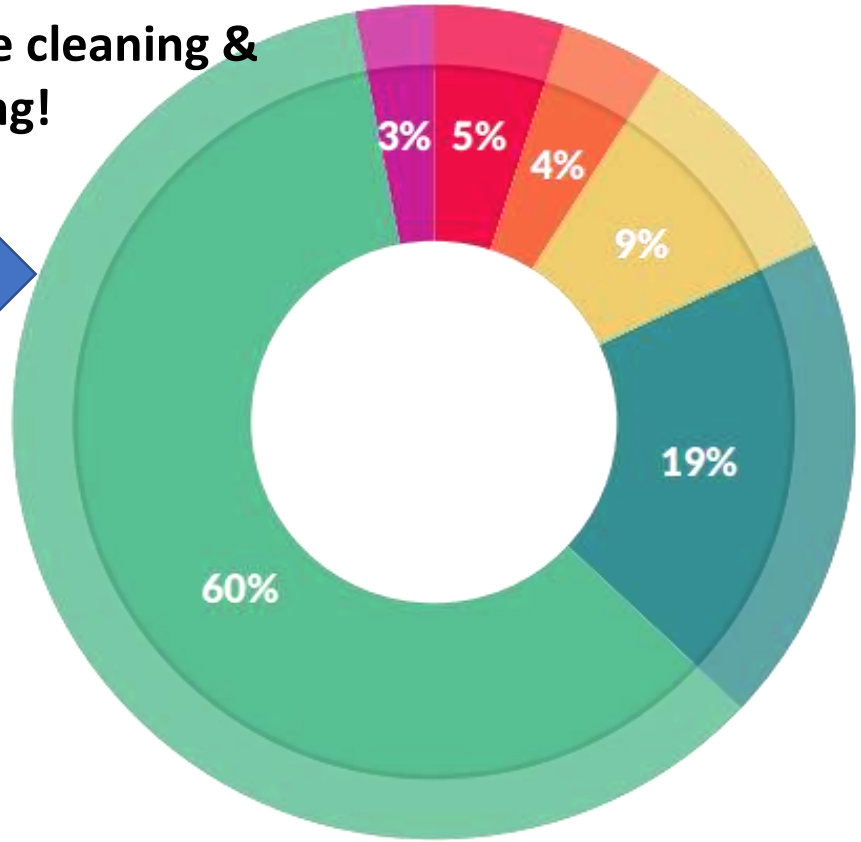


By Common Data Elements: Standardizing Data Collection (nih.gov)

VLIZ pioneer in marine science

Most time goes to data wrangling

60% time cleaning & organizing!



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

CrowdFlower_DataScienceReport_2016

VLIZ pioneer in marine science

More on data cleaning time

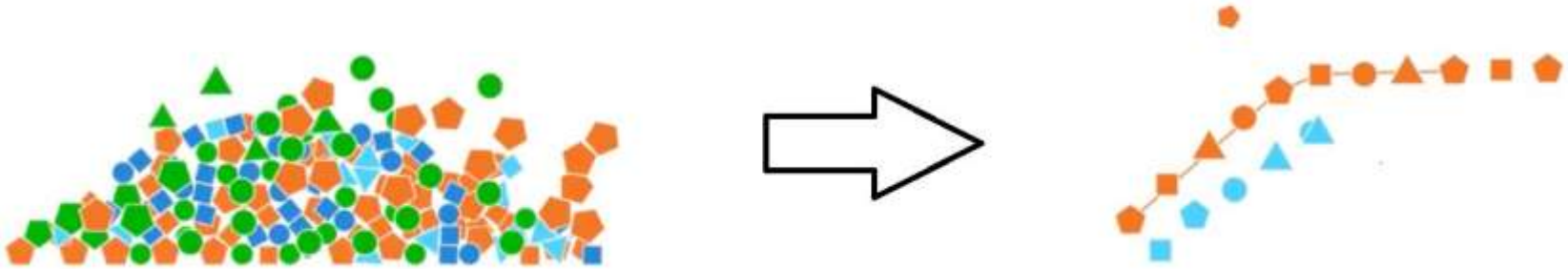
"Students in PhD programmes spend up to **80% of their time** on **'data munging'**, **fixing formatting** and **minor mistakes** to make data suitable for analysis — wasting time and talent."

Invest 5% of research funds in ensuring data are reusable (nature.com)

VLIZ pioneer in marine science

Data cleaning

1. Spell checking
2. Finding & replacing errors
3. Handling missing values
4. Merging and splitting columns
5. Joining tables
6. Tidy data



By ChrisJen517/ CCBY

Sources for this session

- R binder Binder (mybinder.org)

- Rstudio

github.com/lifewatch/ebr-2022-data-cleaning-standardization

-> R/Datacleaning_standardization.R

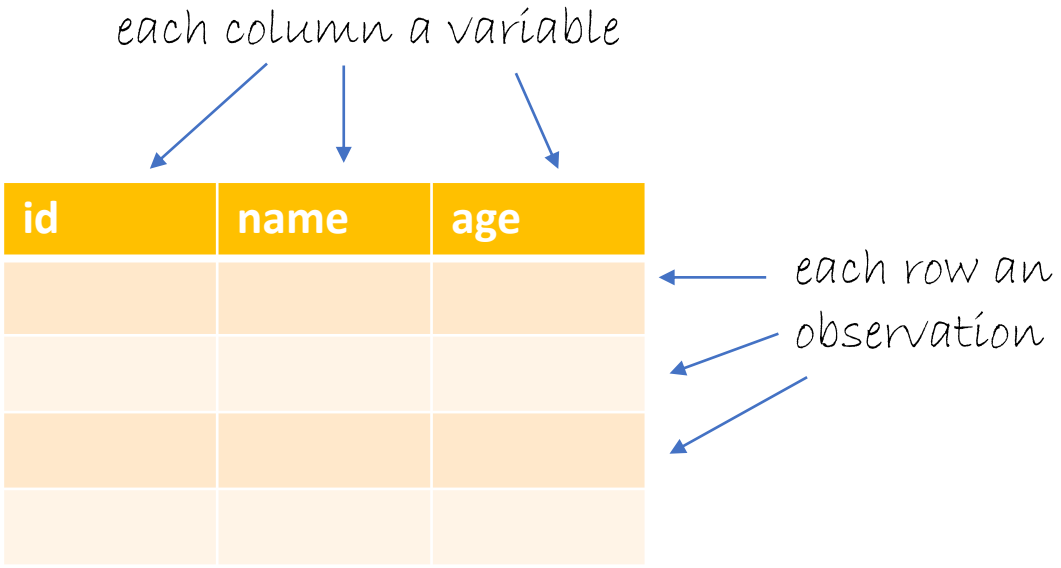
-> data/DataAbundancew.csv

VLIZ pioneer in marine science

Tidy data

Quick reminders on tidy structure:

- Consistent names & good null values
- No spaces or $peci@l characters
- No colors, fonts, italics
- Dates YYYY-mm-dd
- Avoid using multiple chunks of data
- Easy to read formats



INTEROPERABILITY

VLIZ pioneer in marine science

Exercise dataset

- Dataset title: Explore the shore example
- Geographical coverage: Ireland
- Parameters: Abundance
- Taxonomic coverage: Mammalia
- Citation: Fernández S.; Márquez L.; Flanders Marine Institute (VLIZ), Belgium; (2022): Explore the shore example. https://doi.org/10.xxxxx/xxx

VLIZ pioneer in marine science

Time for exercise!

VLIZ pioneer in marine science

Exercises part A – 10 min

1. Open the file DataAbundancew.csv and look at the scope of the data, something strange with the site names? *use unique()
2. Identify mistyping errors in the site names and correct them *use str_replace()
3. Delete special characters and spaces from the data *use str_replace()

VLIZ pioneer in marine science

Key R functions to use in data standardisation

- %>%

- select()
- filter()
- mutate()
- transmute()

VLIZ pioneer in marine science

Time for exercise!

VLIZ pioneer in marine science

Exercises part B – 15 min

1. Standardize the dates in IS0 8601 standard (YYYY-mm-dd) *use separate() & mutate()
2. Use pivot function to get your data tidy (one observation per row) *use pivot_longer()
3. Drop the unnecessary columns *use select()

country	1999	2000
A	0.7K	2K
B	37K	80K
C	212K	213K

→

country	year	cases
A	1999	0.7K
B	1999	37K
C	1999	212K
A	2000	2K
B	2000	80K
C	2000	213K

VLIZ pioneer in marine science

Recap

Data cleaning is key before handling, analyzing, etc.

Making your data TIDY is a great way to make data interoperable (and to share it with colleagues)

In R, tidyverse is a useful (collection) package to help you handling your data.

Upcoming…standardization in biodiversity!

Please install…

R packages **mregions2** & **worrms**

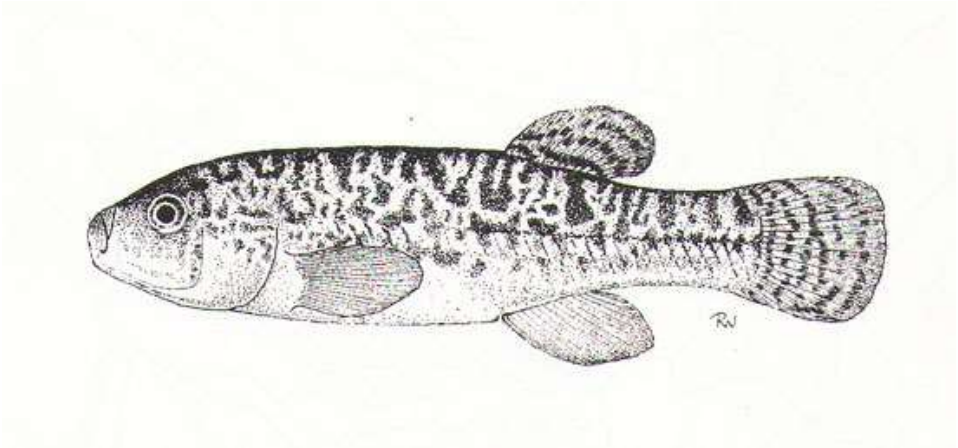
```
#RUN
devtools::install_github("lifewatch/mregions2", build_vignettes = TRUE)

install.packages("worrms")
```

VLIZ pioneer in marine science

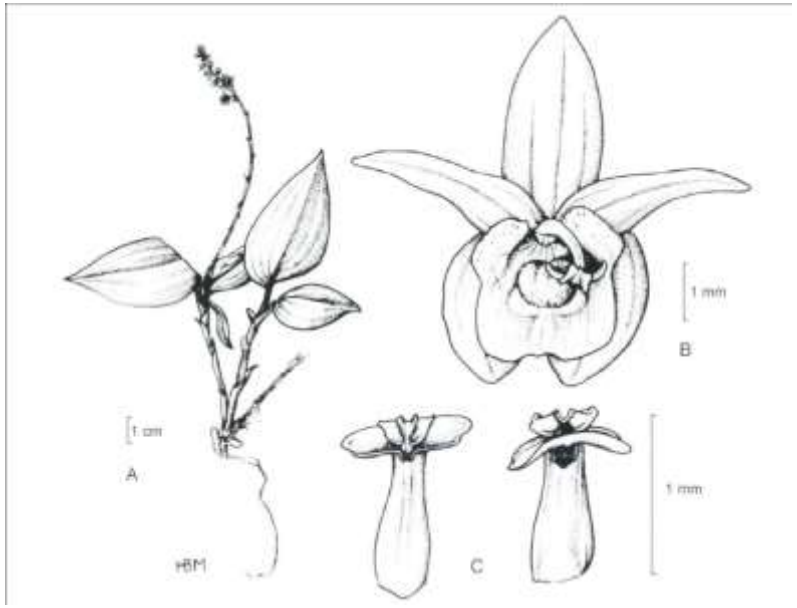
Lunch break

VLIZ pioneer in marine science

Standardisation



By S. Garman 1895



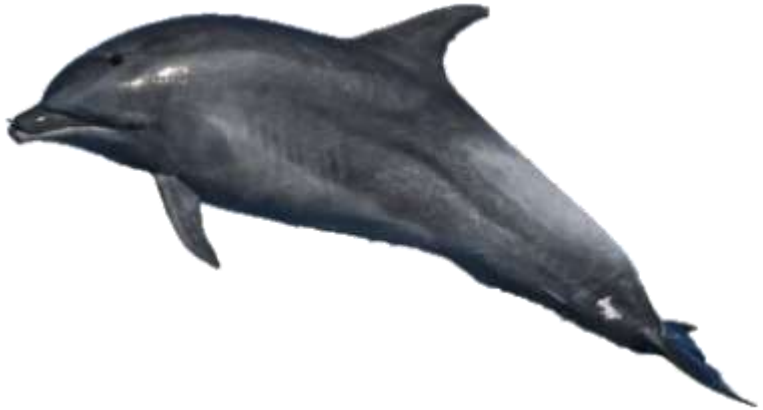
By Margonska and Szlachetko 2005

Orestias elegans

The World's Strangest Scientific Names| Smithsonian Magazine

Standardisation

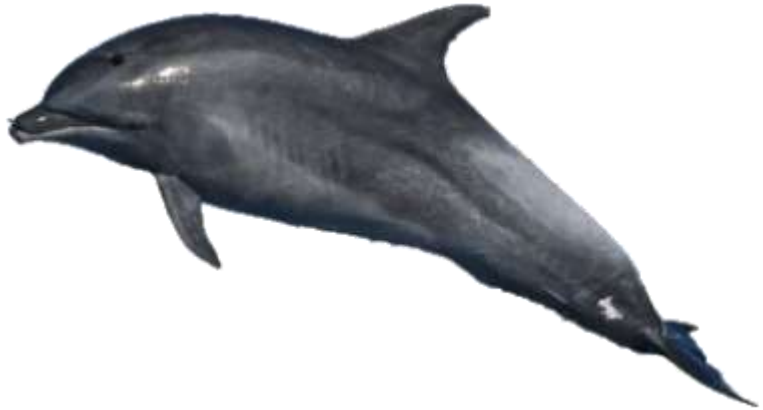
Derived from Laurent/ommag/ CCBY



Tursiops truncantus Montagu, 1821

231423 (urn:lsid:marinespecies.org:taxname:231423)

Derived from Laurent/ommag/ CCBY



Tursiops truncantus (Montagu, 1821)

137111 (urn:lsid:marinespecies.org:taxname:137111)

VLIZ pioneer in marine science

Standardisation

“Standardization is critical to scientists and regulators to ensure the quality and interoperability of research processes…”

Standardization and Omics Science: Technical and Social Dimensions Are Inseparable and Demand Symmetrical Study - PMC (nih.gov)

Standardisation

Standards in biodiversity

Darwin Core

What is in scope?

- Collections of any kind of biological objects or data.
- Terminology associated with biological collection data.
- Striving for compatibility with other biodiversity-related standards.
- Facilitating the addition of components and attributes of biological data.

Darwin Core quick reference guide

Access to Biological Collection Data (ABCD) Schema

Darwin Core

Check many more standards Standards - TDWG

VLIZ pioneer in marine science

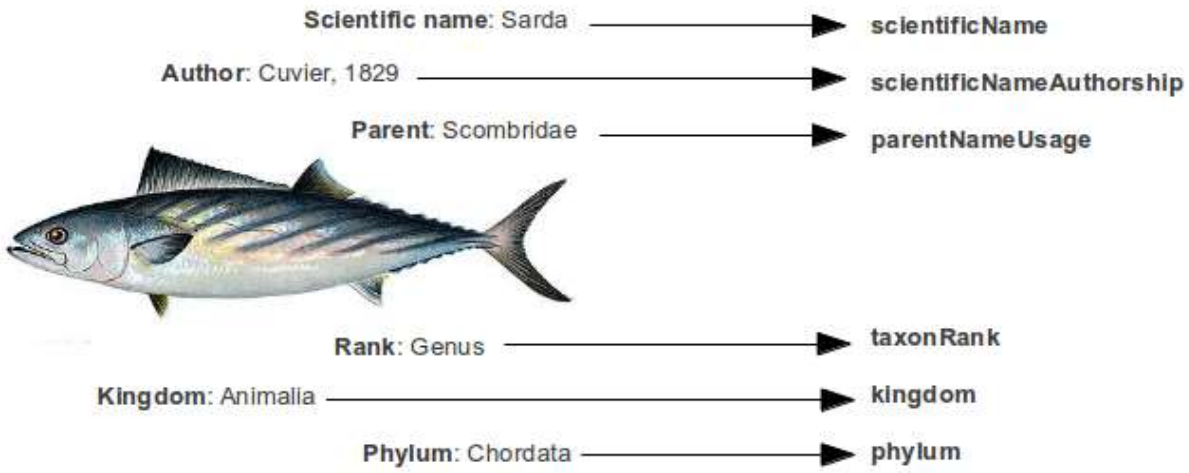
Standardisation

Standardizing the structure of the dataset: term names

Darwin Core

…facilitates the sharing of information about biological diversity

[Darwin Core quick reference guide]



VLIZ pioneer in marine science

Standardization

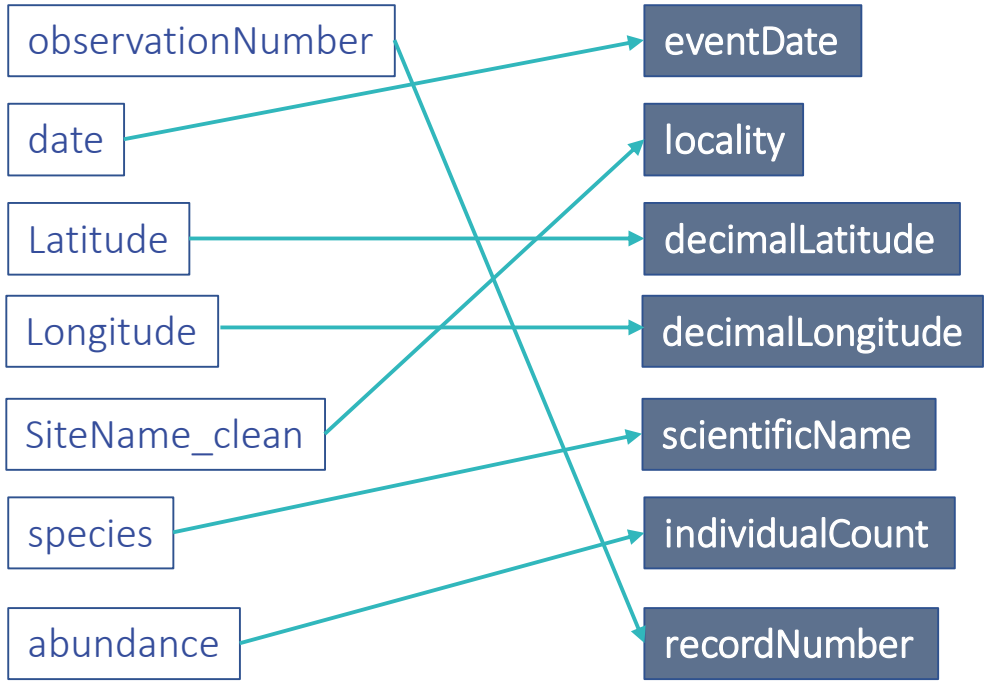
1. Standardizing the structure of the dataset: normative term names

Darwin Core

What is in scope?

- Collections of any kind of biological objects or data.
- Terminology associated with biological collection data.
- Striving for compatibility with other biodiversity-related standards.
- Facilitating the addition of components and attributes of biological data.

[Darwin Core quick reference guide]



VLIZ pioneer in marine science

Exercises part C – 5 min

1. Modify DataAbundancew column names to fit the Darwin Core column names * use rename()

Data standardisation

2. Standardizing data points: taxonomic quality control & georeferenced standards

WoRMS
World Register of Marine Species

Marineregions.org
a standard for georeferenced marine names

VLIZ pioneer in marine science

Data standardization: Taxonomic quality control

WoRMS
World Register of Marine Species

ITIS

THE PALEOBIOLOGY DATABASE
revealing the history of life

Catalogue of Life

GBIF Backbone Taxonomy

unite
community

VLIZ pioneer in marine science

To check more authorities: GBIF Backbone Taxonomy

Data standardization: Taxonomic quality control

Tursiops truncatus

WoRMS
World Register of Marine Species

AphiaID 137111

(urn:lsid:marinespecies.org:taxname:137111) ← LSID

WoRMS -Tursiops truncatus (Montagu, 1821)

VLIZ pioneer in marine science

Time for exercise!

VLIZ pioneer in marine science

Exercises part D: taxon quality control – 10 min

1. Use **worrms** library to obtain the LSID of each species and put them in a new column named “scientificNameID”

* use wm_records_taxamatch (make sure to check the outcome of the search)

* once you obtained them, use “do.call(rbind, myspeciestable)” to transform the list of dataframes to a single

* use left_join to ONLY add the scientificNameID column

VLIZ pioneer in marine science

Data standardization: Geographical standards

Marineregions.org

a standard for georeferenced marine names

COMPOSITE GAZETTEER OF ANTARCTICA

ENEA – P.N.R.A.

GeoNames

ICSM

ANZLIC COMMITTEE ON SURVEYING & MAPPING

Elvis - Place Names - Foundation Spatial Data

VLIZ pioneer in marine science

Marine Regions

ICSM Composite Gazetteer of Australia

GeoNames

Place Names (fsdf.org.au)

Data standardization: Marine georeferencing

MRGID http://marineregions.org/mrgid/5681

Marineregions.org

a standard for georeferenced marine names

Marine Gazetteer Placedetails

Search
Browse
About
Tutorial
Webservices
Login

MRGID http://marineregions.org/mrgid/5681
Status Proposed standard
Name **Language** **Name** **Name source**
English Irish Exclusive economic Zone Flanders Marine Institute (2019). Maritime Boundaries Geodatabase: Maritime Boundaries and Exclusive Economic Zones (200NM), version 11. Available online at http://www.marineregions.org/. https://doi.org/10.14284/386 (look up in IMIS)
PlaceType EEZ
Latitude 52° 39' 9.7" N (52.65269°)
Longitude 11° 44' 46.7" W (-11.74631°)
Precision 583259 meter
Min. Lat 48° 10' 43.4" N (48.1787°)
Min. Long 16° 4' 26" W (-16.0739°)
Max. Lat 56° 42' 0" N (56.7°)
Max. Long 5° 16' 20.4" W (-5.2723°)
Source Flanders Marine Institute (2019). Maritime Boundaries Geodatabase: Maritime Boundaries and Exclusive Economic Zones (200NM), version 11. Available online at http://www.marineregions.org/. https://doi.org/10.14284/386 (look up in IMIS)
Links http://www.marineregions.org/eezdetails.php?mrgid=5681
Notes **Centroid calculation method** (en): Centroid
Relations Part of North Atlantic Ocean (IHO Sea Area) [view hierarchy]
Part of Ireland (Nation) [view hierarchy]

VLIZ pioneer in marine science

Marine regions Gazeetter

MarineRegions R client (github.com)

Time for exercise!

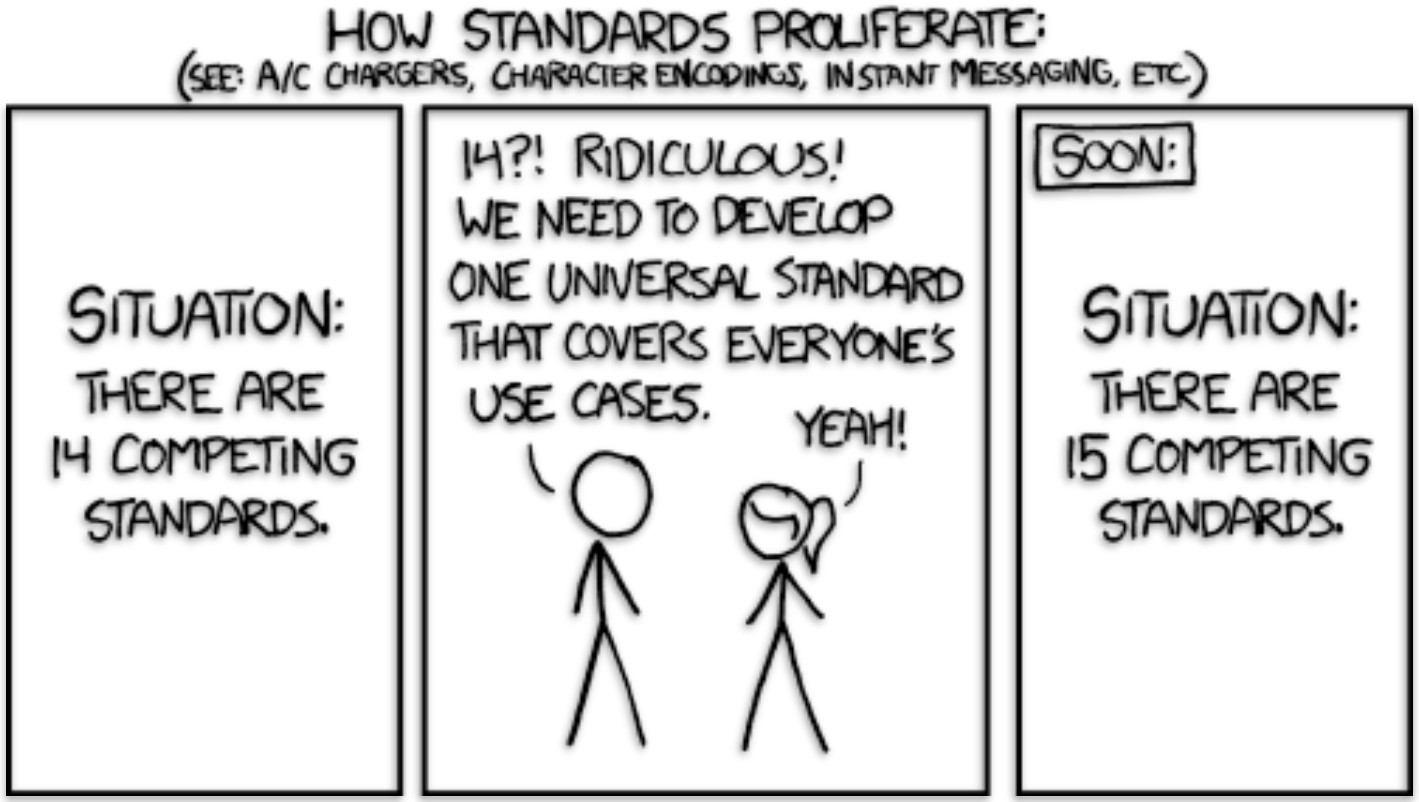
VLIZ pioneer in marine science

Exercises part E: Marine georeferencing – 10 min

1. Use **mregions2** to search for a unique MRGID that can fit all the points of DataAbundancew (make sure to check the outcome of the search) and put it in a new column named “locationID” * use mr_gaz_records_by_name
2. Save your file as a csv

VLIZ pioneer in marine science

Yes, many standards…



By xkcd/ CCBY

VLIZ pioneer in marine science

They work as containers where you can fit in your data!

his data

your data

Derived from Glass Jars With Food/ CC

her data

their data

Summary

- Data cleaning is an essential before handling, analyzing, etc., using R makes it easier
- Making your data “tidy” is a great way to make data interoperable (and to share it with colleagues)
- In R, tidyverse is a (collection of) package to help you handling data.
- In biodiversity, there are different standards useful for sharing your data and ensure its quality, such as Darwin Core or marine regions.

Standards are helpful to make sure we are all in the same page!

Further reading

www.lifewatch.be
www.marinespecies.org
www.marineregions.org
Course: Contributing datasets to EMODnet Biology
Best Practices in Publishing Species Checklists :: GBIF IPT User Manual
The OBIS manual
Introduction to rgbif • rgbif (ropensci.org)

VLIZ pioneer in marine science

LifeWatch BELGIUM

THANKS!

VLIZ pioneer in marine science

For further questions you can contact us:

Salvador Fernández
salvador.fernandez@vliz.be

Laura Márquez
laura.marquez@vliz.be

VLIZ pioneer in marine science