# Dataimport of 2019 data - Planteimport

*Jens Åström*

*25 September, 2020*

## Contents



**To-do**

- Check/ask if pdf_present is Correct for 2018 and 2019 data. Only "x" for 2019 data, but True for most in database. Tidy up the exporter names, there's a mix aof names with and without company type abbrevations, sometimes the same company both with and without abbrevation. Remove all abbreviations seems to be the best.
- Check the transport type for the rest of the containers. Says boat for a few, and NA for the rest.

## Intro

This script imports the 2019 alien stowaway data to the database.

So far, we only deal with the container samples, where we have both plant and invertebrate data from soil samples of imported pots.

## Container data

This will add new rows to the table `common.containers` (schema common, table containers).

We first see what the latest sample in the database currently is.

```
conTab <- tbl(con, in_schema("common", "containers"))

conTab %>%
  arrange(desc(container),
          subsample
          ) %>%
  select(1:5)
```

So we have 87 containers and the last one is sampled in April 2018. Time to load the new data.

```
newConDataRaw <- read.xlsx("../../rawData/data_2019/Database2018 med 2019-data konteinere.xl
                           detectDates = T) %>%
  as_tibble()

newConDataRaw %>%
  select(1:8)
```

So we subset the new data to contain only 2019 data.

```
conData2019 <- newConDataRaw %>%
  filter(container > 87) #found above

conData2019 %>%
  select(1:8)
```

And test an import, find out what's wrong and fix it.

```
dbBegin(con)
dbWriteTable(con,
             Id(schema = "common", table = "containers"),
             conData2019,
             append = T)

dbRollback(con)
```

## Fixes to the container data

This came up through the quality check.

### Change na/x to logical (yes/no).

```
conData2019 <- conData2019 %>%
  mutate(pdf_present = as.logical(ifelse(is.na(pdf_present), FALSE, TRUE)),
         mattilsynet = as.logical(ifelse(is.na(mattilsynet), FALSE, TRUE)))


# tt %>%
#   select(pdf_present,
#          mattilsynet) %>%
#   print(n = Inf)
```

### Change the missing country to unknown and translate country to English. Also change the almost duplicate names to one single (Tyskland/Nederland vs Nederland/Tyskland etc.)

```
conData2019 <- conData2019 %>%
  mutate(country = ifelse(is.na(country), "Unknown", country)) %>%
  mutate(country = ifelse(country == "Nederland", "Netherlands", country),
         country = ifelse(country == "Tyskland", "Germany", country),
         country = ifelse(country == "Tyskland/Nederland", "Germany, Netherlands", country),
         country = ifelse(country == "Litauen/Tyskland", "Germany, Lithuania", country),
         country = ifelse(country == "Italia", "Italy", country),
         country = ifelse(country == "Nederland/Tyskland", "Germany, Netherlands", country),
         country = ifelse(country == "Italia/Nederland", "Netherlands, Italy", country),
         country = ifelse(country == "Spania", "Spain", country),
         country = ifelse(country == "Danmark", "Denmark", country),
         country = ifelse(country == "Tyskland/Litauen", "Germany, Lithuania", country))


conData2019 %>%
  select(country) %>%
  distinct()
```

Add the new "countries" to the lookup table of countries.

```
dbWriteTable(con,
             Id(schema = "common", table = "country"),
             tibble(country = c("Unknown",
                                "Germany, Netherlands",
                                "Germany, Lithuania",
                                "Netherlands, Italy",
```

```
                                    "Spain")),
              append = T)
```

## Tidy up exporter names and add new exporters to the lookup table.

```
conData2019 %>%
  select(exporter) %>%
  arrange(exporter) %>%
  distinct()


exporter <- tbl(con,
                in_schema("common", "exporter"))
exporter %>%
  select(exporter) %>%
  print(n = Inf)
```

These are the fixes I made.

```
conData2019 <- conData2019 %>%
  mutate(exporter = ifelse(exporter == "Aris B.V.", "Aris", exporter),
         exporter = ifelse(exporter == "Aris B V", "Aris", exporter),
         exporter = ifelse(exporter == "Aris B V / Plantagen Source GmbH", "Plantagen source
         exporter = ifelse(exporter == "Floranordic bv .", "Floranordic", exporter),
         exporter = ifelse(exporter == "Floranordic bv,", "Floranordic", exporter),
         exporter = ifelse(exporter == "Noviflora Holland B.V.", "Noviflora Holland", export
         exporter = ifelse(exporter == "Az agr Anania Patrizia - ME/19/0953", "Anania Patriz
         exporter = ifelse(exporter == "CATTANEO BRUNO S.r.l,", "Cattaneo Bruno", exporter),
         exporter = ifelse(exporter == "Elbers Export GmbH", "Elbers Export", exporter),
         exporter = ifelse(exporter == "3IAMBO' PIANTE Dl VITO GIAMBO' (cod. reg. ME/19/1627
         exporter = ifelse(exporter == "Noviflora Holland B.V.", "Noviflora Holland", export
         exporter = ifelse(exporter == "PIantagen Source GmbH", "Plantagen source", exporter
         exporter = ifelse(exporter == "PIantagen Source GmbH/JSC \"Eglesakis\"", "Plantager
         exporter = ifelse(exporter == "Plantagen Source GmbH", "Plantagen source", exporter
         exporter = ifelse(exporter == "Plantagen Source GmbH", "Plantagen source", exporter
         exporter = ifelse(exporter == "PROVAL, S.A.T. N° 362 C.V", "SAT N 362 CV PROVAL", e
         exporter = ifelse(exporter == "Plantagen Source GmbH/Aris B V", "Plantagen source,
         exporter = ifelse(exporter == "GIAMBO' PIANTE Dl VITO GIAMBO' (cod. reg. ME/19/1627
         exporter = ifelse(exporter == "JSC \"Eglesakis\"/Plantagen Source GmbH", "Plantager
         exporter = ifelse(exporter == "Feldborg A/S", "Feldborg", exporter),
         exporter = ifelse(exporter == "Ikke i sertifikatet?", "Unknown", exporter),
         exporter = ifelse(exporter == "Sjekk - ikke i sertifikat", "Unknown", exporter)
```

```
         )


conData2019 %>%
  select(exporter) %>%
  arrange(exporter) %>%
  distinct()
```

We compare this more compact list of exporters to what is already stored in the database.

```
expInData <-
  exporter %>%
  select(id,
         expInData = exporter) %>%
  collect()

conData2019 %>%
  select(exporter) %>%
  distinct() %>%
  left_join(expInData,
            by = c("exporter" = "expInData"))
```

And update the table with the missing exporters.

```
expToImport <- conData2019 %>%
  select(exporter) %>%
  distinct() %>%
  left_join(expInData,
            by = c("exporter" = "expInData")) %>%
  filter(is.na(id)) %>%
  select(-id)

expToImport
```

```
dbWriteTable(con,
             Id(schema = "common", table = "exporter"),
             expToImport,
             append = T)
```

## Update the missing transport_types. (I've added the "Boat" type to the lookup table)

```
conData2019 %>%
  select(transport_type) %>%
```

```
  distinct()

conData2019 <- conData2019 %>%
  mutate(transport_type = ifelse(transport_type == "Båt", "Boat", transport_type))
```

## Update the species list with the new import species

Fix obvious errors, like trim white spaces and uppercase genus.

```
conData2019 <- conData2019 %>%
  mutate(species_latin = str_trim(species_latin)) %>%
  mutate(species_latin = str_to_sentence(species_latin))
```

Change some naming errors.

```
conData2019 <- conData2019 %>%
  mutate(species_latin = ifelse(species_latin == "Juniperus chi", "Juniperus chinensis", spe
         species_latin = ifelse(species_latin == "Salix capra", "Salix caprea", species_lati
         species_latin = ifelse(species_latin == "Festicua glauca", "Festuca glauca", specie
         species_latin = ifelse(species_latin == "Fragesia", "Fargesia", species_latin),
         species_latin = ifelse(species_latin == "Deschamsia caespitosa", "Deschamsia cespit
```

```
impSpecies2019 <- conData2019 %>%
  select(species_latin) %>%
  distinct() %>%
  arrange()

plantSpeciesInDatabase <- tbl(con, in_schema("plants", "species"))

plantSpeciesInDatabase <- plantSpeciesInDatabase %>%
  select(id, species_latin) %>%
  distinct() %>%
  arrange() %>%
  collect()


plantsNotInDatabase <- impSpecies2019 %>%
  left_join(plantSpeciesInDatabase,
            by = c("species_latin" = "species_latin"))
```

These are the new "species" in the container data for 2019. Some trailing white spaces and some genus names not capitalized.

```
plantsNotInDatabase %>%
  filter(is.na(id)) %>%
```

```
  arrange() %>%
  print(n = Inf)
```

## Import new species names to the lookup table

```
plantsToImport <- plantsNotInDatabase %>%
  filter(is.na(id)) %>%
  arrange(species_latin)  %>%
  select(-id)
```

```
dbWriteTable(con, Id(schema = "plants", table = "species"),
             plantsToImport,
             append = T)
```

Some of the new names had matches in artsdatabankens artsnavnebase, and those that didn't seemed to have the correct spelling.

## Import the container data

After this there is no further complaints from the database and we can import the data.

```
dbWriteTable(con,
             Id(schema = "common", table = "containers"),
             conData2019,
             append = T)
```