

# Data format for punching alien stowaway data

*Jens Åström*

*18 September, 2020*

## Contents

<b>Intro</b>	<b>1</b>
<b>Container table</b>	<b>1</b>
<b>Insect container records</b>	<b>2</b>
<b>Insect species names</b>	<b>3</b>
<b>Plant records</b>	<b>5</b>
<b>Plant species names</b>	<b>6</b>

## Intro

This script creates examples of the tables in the planteimport-database, which can be used for punching new data. The idea is that importing the data could be faster if the data was punched in this format from the start. Currently I have to do a lot of rearranging and changing of names manually.

## Container table

This table holds the information of the sampled containers. This is already punched in a very similar format to the table in the database, so we could just continue using this format. Whenever there's a new locality or new exporter, we also need to import these into their respective lookup tables. This happens so infrequently that we probably can do this manually. NB! Don't use the same locality name if the locality has changed. For example "Blomsteringen" is one unique location now. If blomsteringen uses different sample locations, we should split them up with separate names.

This is one row in the container table.

```

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

containers %>%
  print(n = 1,
        width = Inf)

## # Source:   table<common.containers> [?? x 27]
## # Database: postgres [jens.astrom@ninradardata01.nina.no:5432/planteimport]
##   id                      container subsample locality
##   <chr>                  <int>      <int> <chr>
## 1 a88aa624-2200-11e8-8b3b-001cc4ddf696      29      5 Plantasjen Skedsmo
##   date_sampled date_in   date_out   netting_type species_latin
##   <date>       <date>   <date>   <chr>        <chr>
## 1 2015-04-26   2015-04-27 2015-04-30 F      Cardamine pratensis
##   plant_comment wet_volume wet_weight dry_volume dry_weight exporter
##   <chr>          <dbl>    <dbl>    <dbl>    <dbl> <chr>
## 1 viftelønn      2        554      1.5      198 Plantagen source
##   country transport_type pdf_present mattilsynet comment_certificate oh_recieved
##   <chr>    <chr>         <lgl>    <lgl>    <chr>                <chr>
## 1 Germany <NA>         TRUE      FALSE    <NA>                <NA>
##   container_weight number_of_articles number_of_species number_total
##   <dbl>              <int>          <int>          <int>
## 1      NA              7              82             3517
##   volume_per_crate comment_contents
##   <chr>            <chr>
## 1 <NA>            Litt usikker på sertifikatet her
## # ... with more rows

```

When punching, you don't need to fill the "id" column. That is internal to the database. I'll write out a small sample that could be used for punching (but the one we already use is fine!).

```

containers %>%
  select(-id) %>% #internal primary key in database
  collect() %>%
  slice(1:5) %>%
  write_csv(path = "out/example_containers.csv")

```

## Insect container records

This table holds all invertebrate data from the containers. So far, we have only imported the container records in the database, and we will make similar tables for the other collected invertebrates.

```
insect_container_records <- tbl(con, in_schema("insects", "container_records"))
```

There are some columns that are filled automatically in the database; id, latinsknavnid, last\_updated\_by, last\_updated. Project\_id is left blank for now, but could be specified if we want to separate the findings from different projects (contracts?). So these columns are the only ones that we need to punch (project id can be left blank):

```
insect_container_records %>%
  select(-c(id, latinsknavnid, last_updated_by, last_updated)) %>%
  arrange(container, subsample, species_latin)
```

```
## # Source:      lazy query [?? x 5]
## # Database:    postgres [jens.astrom@ninradardata01.nina.no:5432/planteimport]
## # Ordered by:  container, subsample, species_latin
##   container subsample projectid species_latin      amount
##   <int>      <int> <chr>      <chr>          <int>
## 1          1          1 <NA>      Amischa analis      1
## 2          1          1 <NA>      Bourletiella sp. juv. 1
## 3          1          1 <NA>      Desoria grisea       1
## 4          1          1 <NA>      Isotomurus palustris 10
## 5          1          1 <NA>      Proisotoma minuta    1
## 6          1          1 <NA>      Tomocerus vulgaris    1
## 7          1          4 <NA>      Bourletiella sp. juv. 3
## 8          1          4 <NA>      Cartodere bifasciata  1
## 9          1          4 <NA>      Isotomurus palustris 18
## 10         1          4 <NA>      Isotomurus sp. juv.   6
## # ... with more rows
```

I'll write out a short sample that could be used for future punching.

```
insect_container_records %>%
  select(-c(id, latinsknavnid, last_updated_by, last_updated)) %>%
  arrange(container, subsample, species_latin) %>%
  collect() %>%
  slice(1:5) %>%
  write_csv(path = "out/example_insect_container_records.csv")
```

## Insect species names

All invertebrate records need to conform to a list of species names, stored in another table. As new species are discovered, we will add these to the list of names. This list of names is compared to artsnavneliste from artsdatabanken. In the future we will add the black list to the database to be able to update the black list categories automatically (as automatic as possible). Other alien

species lists are also possible to import. We have both a column called “native” and one called “alien”. They are not 100% exclusive, as it appears that some species may not be found earlier in Norway, but still isn’t known to be alien.

Note that we use a single column called **species\_latin** for the unique names of the species or “taxa” we find. This is the columns that are used to match species between tables. In case of juveniles, we specify the juvenile status at the end of the species names, as can be seen in the example. Yes, there still is some cleaning up to do here.

There are quite a few columns in this table that is either filled automatically, or could be filled once and for all by me, using matches from the artsnavnebase. For new species, we will add new rows to this table. Currently, it would be good to fill out all these lines:

```
insect_species <- tbl(con, in_schema("insects", "species"))
```

```
insect_species %>%
  collect() %>%
  group_by(stadium) %>%
  slice(1) %>%
  select(species_latin,
         stadium,
         indetermined,
         autorstring,
         native,
         alien,
         blacklist_cat)
```

```
## # A tibble: 6 x 7
## # Groups:   stadium [6]
##   species_latin stadium indetermined autorstring native alien blacklist_cat
##   <chr>          <chr>    <lgl>         <chr>      <lgl>  <lgl> <chr>
## 1 Supraphorura fur~ adult    FALSE        <NA>      FALSE FALSE <NA>
## 2 Thysanoptera spp~ juvenile TRUE        <NA>      FALSE FALSE <NA>
## 3 Elateridae spp. ~ larvae    TRUE        <NA>      FALSE FALSE <NA>
## 4 Diptera spp. juv. larvae ~ TRUE        <NA>      FALSE FALSE <NA>
## 5 Sternorrhyncha s~ nymph    TRUE        <NA>      FALSE FALSE <NA>
## 6 Aphidoidea spp.~ nymph a~ TRUE        <NA>      FALSE FALSE <NA>
```

I here include the complete list of the “species names” known in the database as of today. **All new records should use these names** if there isn’t a true new “species”.

```
insect_species %>%
  arrange(species_latin) %>%
  select(species_latin,
         stadium,
```

```

        indetermined,
        autorstring,
        native,
        alien,
        blacklist_cat) %>%
collect() %>%
write_csv(path = "out/example_insect_species_names.csv")

```

## Plant records

The plant records are a little bit simpler, but here we need to make a note of if they were found before or after vernalisation.

```

plant_container_records <- tbl(con, in_schema("plants", "container_records"))

plant_container_records %>%
  select(-c(id,
            last_updated_by,
            last_updated))

## # Source:   lazy query [?? x 7]
## # Database: postgres [jens.astro@niradardata01.nina.no:5432/planteimport]
##   container subsample latinsknavnid projectid species_latin amount
##   <int>      <int> <int64>         <chr>      <chr>          <int>
## 1         1         1 NA             <NA>      Cardamine hi~      4
## 2         1         1 NA             <NA>      Cardamine hi~     79
## 3         1         1 NA             <NA>      Cerastium gl~      3
## 4         1         1 NA             <NA>      Cerastium gl~      4
## 5         1         1 NA             <NA>      Senecio vulg~     22
## 6         1         1 NA             <NA>      Stellaria me~      1
## 7         1         1 NA             <NA>      Stellaria me~     17
## 8         1         2 NA             <NA>      Cardamine hi~      1
## 9         1         2 NA             <NA>      Juncus bulbo~      1
## 10        1         3 NA             <NA>      Cardamine hi~      1
## # ... with more rows, and 1 more variable: vernalisation <lgl>

```

I'll write out a short sample that could be used for future punching.

```

plant_container_records %>%
  select(-c(id,
            latinsknavnid,
            projectid,
            last_updated_by,
            last_updated)) %>%
collect() %>%

```

```
slice(1:5) %>%
write_csv(path = "out/example_plant_container_records.csv")
```

## Plant species names

The plant species names are stored in a simpler table than the insects. That's because the original insect table was more complex to start with.

```
plant_species <- tbl(con, in_schema("plants", "species"))
```

For new plant species, it would be good to record at least the columns listed below. NB, autorstring is not important for me, but could be added if you like. Like for insects, the separate columns for native and alien is useful when the status is “complicated” or we haven’t identified the specimen to species. I will add a column “indetermined” here as well (need to get datahjelp to change the ownership of the table).

```
plant_species %>%
  select(species_latin,
         species_norsk,
         autorstreng,
         native,
         alien)
```

```
## # Source:   lazy query [?? x 5]
## # Database: postgres [jens.astro@nirradardata01.nina.no:5432/planteimport]
##   species_latin      species_norsk      autorstreng native alien
##   <chr>              <chr>              <chr>      <lgl>  <lgl>
## 1 Panicum capillare Buskkinapalme/dvergpalme <NA>      FALSE  TRUE
## 2 Erica cf. ciliaris duskamarant           <NA>      FALSE  TRUE
## 3 Humulus cf. yunnanensis sommerfuglbusk        <NA>      FALSE  TRUE
## 4 Panicum capillaceum Dvergsypress         <NA>      FALSE  TRUE
## 5 Panicum repens    Dvergpalme           <NA>      FALSE  TRUE
## 6 Howea sp.         <NA>                 <NA>      FALSE  TRUE
## 7 Chaenorhinum minus småtorskemunn        <NA>      FALSE  TRUE
## 8 Chenopodiastrum murale gatemelde           <NA>      FALSE  TRUE
## 9 Chenopodium ficifolium fikenmelde          <NA>      FALSE  TRUE
## 10 Chenopodium hircinum bukkemelde           <NA>      FALSE  TRUE
## # ... with more rows
```

I’ll write out a complete list of all plant species known in the database so far. Also here, there are some cleaning up to do.

```
plant_species %>%
  arrange(species_latin) %>%
  select(species_latin,
```

```
        species_norsk,  
        autorstreng,  
        native,  
        alien) %>%  
collect() %>%  
write_csv(path = "out/example_plant_species_names.csv")
```

Happy punching!