4PMI Data Preparation

Elise

2024-06-09

Set the right working directory.

```
setwd("C:/Users/elise/Documents/Mémoire/Main/Data/Drive/4PMI")
```

Packages importation

1. Data importation

The first step in this data preparation process involves importing all the pertinent datasets listed in the Google Sheets “Variables template” document. Fist we find the files, then import them.

```
##  [1] "endpoint.txt"
##  [2] "ISA_EPPN2020_4PMI.xlsx"
##  [3] "plant_info.txt"
##  [4] "PLAT015_harvest_phenotyping_datas.txt"
##  [5] "PLAT015_Pot_environmental_datas.txt"
##  [6] "PLAT015_Roots_Traits_datas.txt"
##  [7] "PLAT015_Study_environmental_datas.txt"
##  [8] "S_timeseries.txt"
##  [9] "T_timeseries.txt"
## [10] "timeseries.txt"
```

We can extract the coordinates of each plant with the ISA_EPPN.xlsx dataset, using a made-up function “coordinates_isaTAB”.

```
# Get the coordinates
isaTAB <- read_excel("ISA_EPPN2020_4PMI.xlsx", sheet = "s_exp")
```

```
## New names:
## • `Unit` -> `Unit...9`
## • `Term Source REF` -> `Term Source REF...10`
## • `Term Accession Number` -> `Term Accession Number...11`
## • `Unit` -> `Unit...13`
## • `Term Source REF` -> `Term Source REF...14`
## • `Term Accession Number` -> `Term Accession Number...15`
## • `Unit` -> `Unit...22`
## • `Term Source REF` -> `Term Source REF...23`
## • `Term Accession Number` -> `Term Accession Number...24`
## • `Unit` -> `Unit...26`
## • `Term Source REF` -> `Term Source REF...27`
## • `Term Accession Number` -> `Term Accession Number...28`
```

```
coordinates <- coordinates_isaTAB(isaTAB)
```

A. Datasets structures

We can take a quick look at all the datasets.

- coordinates
- data

```
head(coordinates)
```

```
##                                         Sample.Name nrow ncol rep
## 1  PLAT015_U4L22R1_1_RT_6087_1pl_EPPN2020_7L_Prel05    1   22   5
## 2  PLAT015_U4L22R2_2_RT_5137_1pl_EPPN2020_8H_Prel05    2   22   5
## 3 PLAT015_U4L22R3_3_RT_5251_1pl_EPPN2020_15H_Prel05    3   22   5
## 4 PLAT015_U4L22R4_4_RT_5294_1pl_EPPN2020_14H_Prel05    4   22   5
## 5  PLAT015_U4L22R5_5_RT_5500_1pl_EPPN2020_8L_Prel05    5   22   5
## 6 PLAT015_U4L22R6_6_RT_5515_1pl_EPPN2020_13H_Prel05    6   22   5
```

```
head(data)
```

```
##         date                                      sample_Name
## 1 2019-11-25  PLAT015_U4L7R1_331_RT_5436_1pl_Calib_21T_Prel01
## 2 2019-11-25  PLAT015_U4L7R2_332_RT_6164_1pl_Calib_21T_Prel01
## 3 2019-11-25  PLAT015_U4L7R3_333_RT_5112_1pl_Calib_21T_Prel01
## 4 2019-11-25 PLAT015_U4L7R10_340_RT_5455_1pl_Calib_22T_Prel01
## 5 2019-11-25 PLAT015_U4L7R11_341_RT_5328_1pl_Calib_22T_Prel01
## 6 2019-11-25 PLAT015_U4L7R12_342_RT_6146_1pl_Calib_22T_Prel01
##                 trait Scale_name raw_value OutLier Corrected_Value
## 1 Visible_Leaf_number   Unitless         2      NA              NA
## 2 Visible_Leaf_number   Unitless         1      NA              NA
## 3 Visible_Leaf_number   Unitless         2      NA              NA
## 4 Visible_Leaf_number   Unitless         2      NA              NA
## 5 Visible_Leaf_number   Unitless         2      NA              NA
## 6 Visible_Leaf_number   Unitless         2      NA              NA
```

B. Data manipulation

This next step standardizes diverse datasets by renaming variables for consistency, converting data into appropriate units, adding necessary columns, and merging the datasets.

```
################################################################################
# COORDINATES
################################################################################
# Unit.ID
coordinates$Unit.ID <- seq_len(nrow(coordinates))
# Reference for Sample.Name et Unit.ID
reference <- coordinates[, c("Sample.Name", "Unit.ID")]
## We can then copy dataset2$Unit.ID <- reference$Unit.ID[match(dataset2$Sample.Name, r
eference$Sample.Name)]
# Genotype
coordinates$Genotype <- sapply(strsplit(as.character(coordinates$Sample.Name), "_"), `
[`, 8)
coordinates <- coordinates %>%
  slice(1:330) # filter the calibration step

################################################################################
# DATA
################################################################################
# Time, Date and Timestamp
data$Date <- as.Date(data$date)
data <- filter(data, Date > as.Date('2019-12-03')) # filter the calibration step


# Name of the platform
data$Platform <- "4PMI"

# Unit.ID
data$Unit.ID <- reference$Unit.ID[match(data$sample_Name, reference$Sample.Name)]

# Genotype
data$Genotype <- sapply(strsplit(as.character(data$sample_Name), "_"), `[`, 8)



# Rename the columns for the template
Visible_Leaf_number <- data[data$trait == "Visible_Leaf_number", ]
Shoot_dry_biomass <- data[data$trait == "Shoot_dry_biomass", ]
Root_dry_biomass <- data[data$trait == "Root_dry_biomass", ]
Seed_dry_biomass <- data[data$trait == "Seed_dry_biomass", ]
```

2. Data template

A. Data template: plant_info

This dataset contains information about the plant: Unit.ID, genotype, replication, row and column location in the greenhouse, and soil treatment.

B. Data template: endpoint

This datasets contains information of the end of the experiment (variables at harvest). It is then linked by the Unit.ID to the plant_info data template.

C. Data template: timeseries

This section in divided in three data templates:

- timeseries
- S_timeseries (variables computed from sideview imaging or image processing)
- T_timeseries (variables computed from topview imaging or image processing)

The time interval between data timestamps varies in each platform. They are then linked by the Unit.ID to the plant_info data template.

D. NaPPI data templates

- plant_info
- endpoint
- timeseries
- S_timeseries
- T_timeseries

```
##   Unit.ID Genotype Soil Replication Row Column Platform
## 1       1       7L   NA           5   1     22     4PMI
## 2       2       8H   NA           5   2     22     4PMI
## 3       3      15H   NA           5   3     22     4PMI
## 4       4      14H   NA           5   4     22     4PMI
## 5       5       8L   NA           5   5     22     4PMI
## 6       6      13H   NA           5   6     22     4PMI
##   Unit.ID Time       Date Timestamp DW_shoot_g FW_shoot_g DW_root_g FW_root_g
## 1     166   NA 2019-12-05        NA     0.0593         NA    0.0462        NA
## 2     167   NA 2019-12-05        NA     0.4302         NA    0.3027        NA
## 3     168   NA 2019-12-05        NA     0.2923         NA    0.2075        NA
## 4     169   NA 2019-12-05        NA     0.3760         NA    0.3035        NA
## 5     170   NA 2019-12-05        NA     0.1642         NA    0.0832        NA
## 6     171   NA 2019-12-05        NA     0.2525         NA    0.2168        NA
##   Leaf_number Plant_height_cm DW_plant_g Root_length_cm Root_number Root_angle
## 1           3              NA         NA             NA          NA         NA
## 2           5              NA         NA             NA          NA         NA
## 3           4              NA         NA             NA          NA         NA
## 4           4              NA         NA             NA          NA         NA
## 5           4              NA         NA             NA          NA         NA
## 6           5              NA         NA             NA          NA         NA
##   Total_wu DW_seed_g FW_seed_g Leaf_area_cmsquared Genotype Soil Replication
## 1       NA    0.0346        NA                  NA       4L   NA           8
## 2       NA    0.0230        NA                  NA      11H   NA           8
## 3       NA    0.0313        NA                  NA       7H   NA           8
## 4       NA    0.0380        NA                  NA      14H   NA           8
## 5       NA    0.0495        NA                  NA      12L   NA           8
## 6       NA    0.0307        NA                  NA       6L   NA           8
##   Row Column Platform
## 1  12     15     4PMI
## 2  13     15     4PMI
## 3  14     15     4PMI
## 4  15     15     4PMI
## 5  16     15     4PMI
## 6  17     15     4PMI
```

```
##   Unit.ID Time Date Timestamp Manual_Plant_height_cm Leaf_number Wue
## 1    <NA>   NA   NA        NA                     NA          NA  NA
##   Plant_biomass Ligulated_leaf_number Plant_emergence Plant_transpiration
## 1            NA                    NA              NA                  NA
##   Daily_wu Soil_water_potential Genotype Soil Replication  Row Column Platform
## 1       NA                   NA     <NA>   NA        <NA> <NA>   <NA>     <NA>
```

```
##   Unit.ID Timestamp Date Time S_Height_cm S_Height_pixel S_Area_cmsquared
## 1    <NA>        NA   NA   NA          NA             NA               NA
##   S_Area_pixel S_Perimeter_cm S_Perimeter_pixel S_Convex_hull_area_cmsquared
## 1           NA             NA                NA                           NA
##   S_Solidity S_Compactness S_Width_cm S_Width_pixel S_Leaf_area_cmsquared
## 1         NA            NA         NA            NA                    NA
##   Genotype Soil Replication  Row Column Platform
## 1     <NA>   NA        <NA> <NA>   <NA>     <NA>
```

```
##   Unit.ID Time Date Timestamp T_Area_cm_squared T_Area_pixel T_Perimeter_cm
## 1    <NA>   NA   NA        NA                NA           NA             NA
##   T_Perimeter_pixel T_Convex_hull_area_cmsquared T_Solidity T_Compactness
## 1                NA                           NA         NA            NA
##   T_Roundness T_Roundness2 T_Isotropy T_Eccentricity T_Rms T_Sol Genotype Soil
## 1          NA           NA         NA             NA    NA    NA     <NA>   NA
##   Replication  Row Column Platform
## 1        <NA> <NA>   <NA>     <NA>
```

3. Export the data templates in .txt

Stock the new data sets in a new folder.

```
setwd("C:/Users/elise/Documents/Mémoire/Main/Data/Templates/4PMI")

write.table(plant_info, file = "plant_info.txt", sep = "\t", row.names = FALSE, quote = FALSE)
write.table(endpoint, file = "endpoint.txt", sep = "\t", row.names = FALSE, quote = FALSE)
write.table(timeseries, file = "timeseries.txt", sep = "\t", row.names = FALSE, quote = FALSE)
write.table(S_timeseries, file = "S_timeseries.txt", sep = "\t", row.names = FALSE, quote = FALSE)
write.table(T_timeseries, file = "T_timeseries.txt", sep = "\t", row.names = FALSE, quote = FALSE)
```