ALSIA Data Preparation

Elise

2024-06-08

Set the right working directory.

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

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] "ALSIA.Rmd"                 "Alsia_Initial Code Draft"
##  [3] "extracted_imaging.txt"     "filtered_enviromental.txt"
##  [5] "ISA_EPPN2020_ALSIA.xlsx"   "Nouveau dossier"
##  [7] "raw_destructive.txt"       "raw_enviromental.txt"
##  [9] "raw_imaging.txt"           "raw_water.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_ALSIA.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_pheno
- data_imaging
- data_environment

```
head(coordinates)
```

```
##   Sample.Name nrow ncol rep
## 1 DIC09D11A01    6   18   1
## 2 DIC09D11A02   15   18   2
## 3 DIC09D11A03    5   17   3
## 4 DIC09D11A04   20   17   4
## 5 DIC09D11A05    3   16   5
## 6 DIC09D11A06   19   16   6
```

```
head(data)
```

```
##
## 1 function (..., list = character(), package = NULL, lib.loc = NULL,
## 2     verbose = getOption("verbose"), envir = .GlobalEnv, overwrite = TRUE)
## 3 {
## 4     fileExt <- function(x) {
## 5         db <- grepl("\\\\.[^.]+\\\\.(gz|bz2|xz)$", x)
## 6         ans <- sub(".*\\\\.", "", x)
```

```
head(data_imaging)
```

```
##             timestamp       date plantbarcode genotype   type Column Row
## 1 2020-09-10 09:10:51 2020-09-10  DIC09D11G09  EPPN3_H Hybrid     14   1
## 2 2020-09-10 09:11:06 2020-09-10  DIC09D11F09  EPPN2_H Hybrid     14   2
## 3 2020-09-10 09:12:36 2020-09-10  DIC09D11C09  EPPN3_L   Line     14   5
## 4 2020-09-10 09:13:14 2020-09-10  DIC09D11D09  EPPN4_L   Line     14   6
## 5 2020-09-10 09:13:52 2020-09-10  DIC09D11H09  EPPN4_H Hybrid     14   7
## 6 2020-09-10 09:14:30 2020-09-10  DIC09D11E09  EPPN1_H Hybrid     14   8
##   replica potId   area.S   area.T convex_hull_area.S convex_hull_area.T
## 1       9 c1r14 5.732298 14.51799           19.20558           25.69446
## 2       9 c2r14 8.855819 27.38742           49.34901           62.02855
## 3       9 c5r14 8.901922 12.57987           21.06872           16.49839
## 4       9 c6r14 3.435569 14.87307           11.27200           20.61771
## 5       9 c7r14 6.415449 14.13333           28.78837           22.18702
## 6       9 c8r14 8.873346 24.59541           32.53503           54.50113
##   solidity.S solidity.T height_above_reference.S projected_shoot_area wue
## 1  0.3013386  0.5650243                 6.982816             25.98259  NA
## 2  0.2199222  0.4415292                11.037797             45.09906  NA
## 3  0.4374041  0.7624904                 8.714774             30.38371  NA
## 4  0.4049420  0.7213737                 3.999999             21.74421  NA
## 5  0.2887331  0.6370088                11.051543             26.96423  NA
## 6  0.4510514  0.4512826                10.336766             42.34211  NA
```

```
head(data_environment)
```

```
##              date_time                sensorID          variable  value
## 1 2020-09-07T23:58:00Z       station_01/par_01               PAR    0.0
## 2 2020-09-08T00:17:00Z       station_01/co2_01               CO2  450.0
## 3 2020-09-08T00:33:00Z station_01/multisens_01       Temperature   22.9
## 4 2020-09-08T00:55:00Z station_01/multisens_01 Relative Humidity  100.0
## 5 2020-09-08T14:55:00Z       station_01/par_01               PAR 1015.0
## 6 2020-09-08T15:00:00Z       station_01/co2_01               CO2  672.0
```

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, reference$Sample.Name)]

################################################################################
# DATA_PHENO
################################################################################

# Time, Date and Timestamp
data_pheno$Timestamp <- as.POSIXct(data_pheno$timestamp, format = "%Y-%m-%d %H:%M:%S")
data_pheno$Date <- as.Date(data_pheno$date, format = "%Y-%m-%d")
data_pheno$Time <- sapply(strsplit(as.character(data_pheno$timestamp), split = " "), '[', 2)

# Name of the platform
data_pheno$Platform <- "ALSIA"

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

# Rename the columns for the template
data_pheno <- rename(data_pheno,
                     Genotype = genotype,
                     Replication = replica,
                     FW_shoot_g = fresh_weight,
                     Plant_height_cm = manual_plant_height
                     )

################################################################################
# DATA_IMAGING
################################################################################
head(data_imaging)
```

```
##             timestamp       date plantbarcode genotype   type Column Row
## 1 2020-09-10 09:10:51 2020-09-10  DIC09D11G09  EPPN3_H Hybrid     14   1
## 2 2020-09-10 09:11:06 2020-09-10  DIC09D11F09  EPPN2_H Hybrid     14   2
## 3 2020-09-10 09:12:36 2020-09-10  DIC09D11C09  EPPN3_L   Line     14   5
## 4 2020-09-10 09:13:14 2020-09-10  DIC09D11D09  EPPN4_L   Line     14   6
## 5 2020-09-10 09:13:52 2020-09-10  DIC09D11H09  EPPN4_H Hybrid     14   7
## 6 2020-09-10 09:14:30 2020-09-10  DIC09D11E09  EPPN1_H Hybrid     14   8
##   replica potId   area.S   area.T convex_hull_area.S convex_hull_area.T
## 1       9 c1r14 5.732298 14.51799           19.20558           25.69446
## 2       9 c2r14 8.855819 27.38742           49.34901           62.02855
## 3       9 c5r14 8.901922 12.57987           21.06872           16.49839
## 4       9 c6r14 3.435569 14.87307           11.27200           20.61771
## 5       9 c7r14 6.415449 14.13333           28.78837           22.18702
## 6       9 c8r14 8.873346 24.59541           32.53503           54.50113
##   solidity.S solidity.T height_above_reference.S projected_shoot_area wue
## 1  0.3013386  0.5650243                 6.982816             25.98259  NA
## 2  0.2199222  0.4415292                11.037797             45.09906  NA
## 3  0.4374041  0.7624904                 8.714774             30.38371  NA
## 4  0.4049420  0.7213737                 3.999999             21.74421  NA
## 5  0.2887331  0.6370088                11.051543             26.96423  NA
## 6  0.4510514  0.4512826                10.336766             42.34211  NA
```

```
# Time, Date and Timestamp
data_imaging$Timestamp <- as.POSIXct(data_imaging$timestamp, format = "%Y-%m-%d %H:%M:%
S")
data_imaging$Date <- as.Date(data_imaging$date, format = "%Y-%m-%d")
data_imaging$Time <- sapply(strsplit(as.character(data_imaging$timestamp), split = "
"), '[', 2)

# Name of the platform
data_imaging$Platform <- "ALSIA"

# Unit.ID
data_imaging$Unit.ID <- reference$Unit.ID[match(data_imaging$plantbarcode, reference$Sa
mple.Name)]

# Rename the columns for the template
data_imaging <- rename(data_imaging,
                       Genotype = genotype,
                       Replication = replica,
                       S_Area_cmsquared = area.S,
                       T_Area_cmsquared = area.T,
                       S_Convex_hull_area_cmsquared = convex_hull_area.S,
                       T_Convex_hull_area_cmsquared = convex_hull_area.T,
                       S_Solidity = solidity.S,
                       T_Solidity = solidity.T,
                       S_Height_cm = height_above_reference.S,
                       S_Leaf_area_cmsquared = projected_shoot_area,
                       Wue = wue
                       )
```

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. ALSIA data templates

- plant_info
- endpoint
- timeseries
- S_timeseries

```
##   Unit.ID Genotype Soil Replication Row Column Platform
## 1       1  EPPN1_L   NA           1   6     18    ALSIA
## 2       2  EPPN1_L   NA           2  15     18    ALSIA
## 3       3  EPPN1_L   NA           3   5     17    ALSIA
## 4       4  EPPN1_L   NA           4  20     17    ALSIA
## 5       5  EPPN1_L   NA           5   3     16    ALSIA
## 6       6  EPPN1_L   NA           6  19     16    ALSIA
##   Unit.ID     Time       Date           Timestamp DW_shoot_g FW_shoot_g
## 1       1 12:00:00 2020-10-03 2020-10-03 12:00:00         NA       69.3
## 2       2 12:00:00 2020-10-03 2020-10-03 12:00:00         NA       59.0
## 3       3 12:00:00 2020-10-03 2020-10-03 12:00:00         NA       98.3
## 4       4 12:00:00 2020-10-03 2020-10-03 12:00:00         NA       50.1
## 5       5 12:00:00 2020-10-03 2020-10-03 12:00:00         NA       37.7
## 6       6 12:00:00 2020-10-03 2020-10-03 12:00:00         NA       95.6
##   DW_root_g FW_root_g Leaf_number Plant_height_cm DW_plant_g Root_length_cm
## 1        NA        NA          NA             110         NA             NA
## 2        NA        NA          NA             105         NA             NA
## 3        NA        NA          NA             125         NA             NA
## 4        NA        NA          NA              99         NA             NA
## 5        NA        NA          NA             104         NA             NA
## 6        NA        NA          NA             115         NA             NA
##   Root_number Root_angle Total_wu DW_seed_g FW_seed_g Leaf_area_cmsquared
## 1          NA         NA       NA        NA        NA                  NA
## 2          NA         NA       NA        NA        NA                  NA
## 3          NA         NA       NA        NA        NA                  NA
## 4          NA         NA       NA        NA        NA                  NA
## 5          NA         NA       NA        NA        NA                  NA
## 6          NA         NA       NA        NA        NA                  NA
##   Genotype Soil Replication Row Column Platform
## 1  EPPN1_L   NA           1   6     18    ALSIA
## 2  EPPN1_L   NA           2  15     18    ALSIA
## 3  EPPN1_L   NA           3   5     17    ALSIA
## 4  EPPN1_L   NA           4  20     17    ALSIA
## 5  EPPN1_L   NA           5   3     16    ALSIA
## 6  EPPN1_L   NA           6  19     16    ALSIA
##   Unit.ID     Time       Date           Timestamp Manual_Plant_height_cm
## 1      69 09:10:51 2020-09-10 2020-09-10 09:10:51                     NA
## 2      59 09:11:06 2020-09-10 2020-09-10 09:11:06                     NA
## 3      29 09:12:36 2020-09-10 2020-09-10 09:12:36                     NA
## 4      39 09:13:14 2020-09-10 2020-09-10 09:13:14                     NA
## 5      79 09:13:52 2020-09-10 2020-09-10 09:13:52                     NA
## 6      49 09:14:30 2020-09-10 2020-09-10 09:14:30                     NA
##   Leaf_number Wue Plant_biomass Ligulated_leaf_number Plant_emergence
## 1          NA  NA            NA                    NA              NA
## 2          NA  NA            NA                    NA              NA
## 3          NA  NA            NA                    NA              NA
## 4          NA  NA            NA                    NA              NA
## 5          NA  NA            NA                    NA              NA
## 6          NA  NA            NA                    NA              NA
##   Plant_transpiration Daily_wu Soil_water_potential Genotype Soil Replication
## 1                  NA       NA                   NA  EPPN3_H   NA           9
## 2                  NA       NA                   NA  EPPN2_H   NA           9
## 3                  NA       NA                   NA  EPPN3_L   NA           9
## 4                  NA       NA                   NA  EPPN4_L   NA           9
## 5                  NA       NA                   NA  EPPN4_H   NA           9
## 6                  NA       NA                   NA  EPPN1_H   NA           9
##   Row Column Platform
## 1   1     14    ALSIA
## 2   2     14    ALSIA
## 3   5     14    ALSIA
## 4   6     14    ALSIA
## 5   7     14    ALSIA
## 6   8     14    ALSIA
##   Unit.ID           Timestamp       Date     Time S_Height_cm S_Height_pixel
## 1      69 2020-09-10 09:10:51 2020-09-10 09:10:51    6.982816             NA
## 2      59 2020-09-10 09:11:06 2020-09-10 09:11:06   11.037797             NA
## 3      29 2020-09-10 09:12:36 2020-09-10 09:12:36    8.714774             NA
## 4      39 2020-09-10 09:13:14 2020-09-10 09:13:14    3.999999             NA
## 5      79 2020-09-10 09:13:52 2020-09-10 09:13:52   11.051543             NA
## 6      49 2020-09-10 09:14:30 2020-09-10 09:14:30   10.336766             NA
##   S_Area_cmsquared S_Area_pixel S_Perimeter_cm S_Perimeter_pixel
## 1         5.732298           NA             NA                NA
## 2         8.855819           NA             NA                NA
## 3         8.901922           NA             NA                NA
## 4         3.435569           NA             NA                NA
## 5         6.415449           NA             NA                NA
## 6         8.873346           NA             NA                NA
##   S_Convex_hull_area_cmsquared S_Solidity S_Compactness S_Width_cm
## 1                     19.20558  0.3013386            NA         NA
## 2                     49.34901  0.2199222            NA         NA
## 3                     21.06872  0.4374041            NA         NA
## 4                     11.27200  0.4049420            NA         NA
## 5                     28.78837  0.2887331            NA         NA
## 6                     32.53503  0.4510514            NA         NA
##   S_Width_pixel S_Leaf_area_cmsquared Genotype Soil Replication Row Column
## 1            NA              25.98259  EPPN3_H   NA           9   1     14
## 2            NA              45.09906  EPPN2_H   NA           9   2     14
## 3            NA              30.38371  EPPN3_L   NA           9   5     14
## 4            NA              21.74421  EPPN4_L   NA           9   6     14
## 5            NA              26.96423  EPPN4_H   NA           9   7     14
## 6            NA              42.34211  EPPN1_H   NA           9   8     14
##   Platform
## 1    ALSIA
## 2    ALSIA
## 3    ALSIA
## 4    ALSIA
## 5    ALSIA
## 6    ALSIA
```

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/ALSIA")

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 = FAL
SE)
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, quo
te = FALSE)
write.table(T_timeseries, file = "T_timeseries.txt", sep = "\t", row.names = FALSE, quo
te = FALSE)
```