UCL Data Preparation

Elise

2024-06-09

Set the right working directory.

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

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] "env-multisensor.txt"
## [2] "img_raw.txt"
## [3] "ISA_EPPN2020_UCLouvain - data_multisensor.csv"
## [4] "ISA_EPPN2020_UCLouvain.xlsx"
```

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_UCLouvain.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...21`
## • `Term Source REF` -> `Term Source REF...22`
## • `Term Accession Number` -> `Term Accession Number...23`
## • `Unit` -> `Unit...25`
## • `Term Source REF` -> `Term Source REF...26`
## • `Term Accession Number` -> `Term Accession Number...27`
```

```
coordinates <- coordinates_isaTAB(isaTAB)

# For the UCL platform, the data is not saved with the same format
coordinates$box <- coordinates$nrow
coordinates$nrow <- coordinates$ncol
coordinates$ncol <- coordinates$rep
coordinates$rep <- NA
```

A. Datasets structures

We can take a quick look at all the datasets.

- coordinates
- data_pheno
- data_environment

```
head(coordinates)
```

```
##   Sample.Name nrow ncol rep box
## 1     A_01_1    1    1  NA   A
## 2     A_02_1    2    1  NA   A
## 3     A_03_1    3    1  NA   A
## 4     A_04_1    4    1  NA   A
## 5     A_05_1    5    1  NA   A
## 6     A_06_1    6    1  NA   A
```

```
head(data_pheno)
```

```
##   Unit.ID Timestamp      Device.ID        Variable.ID Value
## 1  A_01_1        NA SCALE_SEFY_01 HarvestRootFreshMass  3.94
## 2  A_02_1        NA SCALE_SEFY_01 HarvestRootFreshMass  3.17
## 3  A_03_1        NA SCALE_SEFY_01 HarvestRootFreshMass  2.61
## 4  A_05_1        NA SCALE_SEFY_01 HarvestRootFreshMass  4.32
## 5  A_06_1        NA SCALE_SEFY_01 HarvestRootFreshMass  1.28
## 6  A_07_1        NA SCALE_SEFY_01 HarvestRootFreshMass  3.90
```

```
head(data_environment)
```

```
##                            Unit.ID        Timestamp  Device.ID
## 1 EPPN2020_JRA1.4_OBJ1_ROOTPHAIR 01/12/2020 00:02 PAR_GH_EXT
## 2 EPPN2020_JRA1.4_OBJ1_ROOTPHAIR 01/12/2020 00:02    T_GH_S22
## 3 EPPN2020_JRA1.4_OBJ1_ROOTPHAIR 01/12/2020 00:02   RH_GH_S22
## 4 EPPN2020_JRA1.4_OBJ1_ROOTPHAIR 01/12/2020 00:02      SUPLED
## 5 EPPN2020_JRA1.4_OBJ1_ROOTPHAIR 01/12/2020 00:02  SHADINGNET
## 6 EPPN2020_JRA1.4_OBJ1_ROOTPHAIR 01/12/2020 00:03 PAR_GH_EXT
##              Variable.ID Value
## 1          AirTemperature   0.0
## 2 VisibleSpectrumRadiation  19.9
## 3      AirRelativeHumidity  55.0
## 4     SupplementalRadiation   0.0
## 5      PlantLevelRadiation   0.0
## 6          AirTemperature   0.0
```

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.

```r
################################################################################
# COORDINATES
################################################################################
# Unit.ID
coordinates$Unit.ID <- seq_len(nrow(coordinates))
# Genotype
coordinates$Genotype <- isaTAB$`Source Name`
# Reference for Sample.Name et Unit.ID
reference <- coordinates[, c("Sample.Name", "Unit.ID")]
## We can then copy dataset2$Unit.ID <- reference$Unit.ID[match(dataset2$Sample.Name, r
eference$Sample.Name)]


################################################################################
# DATA_PHENO
################################################################################
# Time, Date and Timestamp
data_pheno$Date <- as.Date("2020-12-23")

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

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

# Rename the columns for the template
DW_shoot_g <- data_pheno %>%
  filter(Variable.ID == "HarvestShootDryMass") %>%
  select(Unit.ID, Value) %>%
  rename(DW_shoot_g = Value)

FW_shoot_g <- data_pheno %>%
  filter(Variable.ID == "HarvestShootFreshMass") %>%
  select(Unit.ID, Value) %>%
  rename(FW_shoot_g = Value)

DW_root_g <- data_pheno %>%
  filter(Variable.ID == "HarvestRootDryMass") %>%
  select(Unit.ID, Value) %>%
  rename(DW_root_g = Value)

FW_root_g <- data_pheno %>%
  filter(Variable.ID == "HarvestRootFreshMass") %>%
  select(Unit.ID, Value) %>%
  rename(FW_root_g = Value)

data <- list(DW_shoot_g, FW_shoot_g, DW_root_g, FW_root_g)
data <- reduce(data, function(x, y) merge(x, y, by = "Unit.ID", all = TRUE))
```

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. UCL data templates

- plant_info
- endpoint
- timeseries
- S_timeseries
- T_timeseries

```
##   Unit.ID Genotype Soil Replication Row Column Platform
## 1       1 EPPN14_H   NA          NA   1      1      UCL
## 2       2  EPPN3_L   NA          NA   2      1      UCL
## 3       3 EPPN12_H   NA          NA   3      1      UCL
## 4       4 EPPN12_L   NA          NA   4      1      UCL
## 5       5 EPPN10_H   NA          NA   5      1      UCL
## 6       6    Local   NA          NA   6      1      UCL
```

```
##   Unit.ID Time       Date Timestamp DW_shoot_g FW_shoot_g DW_root_g FW_root_g
## 1       1   NA 2020-12-23        NA     0.3837       5.47    0.1703      3.94
## 2       2   NA 2020-12-23        NA     0.3246       4.78    0.1279      3.17
## 3       3   NA 2020-12-23        NA     0.2416       3.53    0.1040      2.61
## 4       5   NA 2020-12-23        NA     0.5105       7.63    0.1912      4.32
## 5       6   NA 2020-12-23        NA     0.2566       3.57    0.0624      1.28
## 6       7   NA 2020-12-23        NA     0.3246       4.91    0.1696      3.90
##   Leaf_number Plant_height_cm DW_plant_g Root_length_cm Root_number Root_angle
## 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
##   Total_wu DW_seed_g FW_seed_g Leaf_area_cmsquared Genotype Soil Replication
## 1       NA        NA        NA                  NA EPPN14_H   NA          NA
## 2       NA        NA        NA                  NA  EPPN3_L   NA          NA
## 3       NA        NA        NA                  NA EPPN12_H   NA          NA
## 4       NA        NA        NA                  NA EPPN10_H   NA          NA
## 5       NA        NA        NA                  NA    Local   NA          NA
## 6       NA        NA        NA                  NA  EPPN6_H   NA          NA
##   Row Column Platform
## 1   1      1      UCL
## 2   2      1      UCL
## 3   3      1      UCL
## 4   5      1      UCL
## 5   6      1      UCL
## 6   7      1      UCL
```

```
##   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/UCL")

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)
```