TRYdb plant trait info

Desi Quintans

3 June 2018

```
librarian::shelf(dplyr, janitor, DesiQuintans/desiderata, readr, readxl, tidyr, visdat,
                 ggplot2, stringr, skimr, magrittr, knitr, kableExtra, stringdist, purrr)

knitr::opts_chunk$set(echo = TRUE)
```

Goal of this document

I have 3 sets of TRYdb https://www.try-db.org requests:

1. A set requested by Paul Corragio (`pc`)
2. A set requested by me early in my project (`d1`)
3. A set requested by me based on the plants whose pods I had collected (`d2`)

I want to gather them all up and see if they cover all of my species and all of the traits that I want.

PS: This is starting to look like fertile ground for another package.

TRYdb request details

I kept the details I submitted for the `d2` request:

Trait codes

```
requested_traits <- c(2940, 917, 918, 919, 2939, 2941, 99, 605, 585, 33, 1099, 3041, 26,
                      1101, 596, 34, 353, 1102, 27, 2946, 132, 350, 351, 1103, 1104, 131,
                      2944, 2945, 3000, 3045, 349, 98, 1108, 3043, 238, 1109, 3042, 239,
                      3044, 359, 611, 865, 3105, 3103, 207, 3008, 210, 2937, 205, 213,
                      212, 211, 2935, 206, 215, 597, 2999, 2956, 920, 130, 2817, 684,
                      688, 923, 685, 2934, 2947, 1111, 2809, 2807, 2808, 159, 22, 3027,
                      1024, 346, 345, 1026, 700, 119, 117, 1255, 1027, 197, 42, 587, 59,
                      788, 599, 668, 1033, 155, 1035, 1036, 679, 3028, 335, 602, 819,
                      681, 30, 318, 1251)
```

Species codes

These are all of the species that I collected pods from throughout my PhD.

```
requested_plants <- c(188, 198, 213, 219, 253, 279, 304, 324, 329, 358, 373, 381, 430,
                      434, 486, 498, 510, 516, 517, 539, 581, 617, 631, 641, 647, 7801,
                      7804, 7808, 7811, 17063, 17067, 17071, 17073, 17080, 17484, 18193,
                      18194, 18201, 26503, 26511, 26628, 26629, 26630, 26632, 26635,
                      27880, 27882, 29607, 29609, 30433, 36936, 36940, 42624, 42971,
                      45070, 45074, 45089, 45096, 45102, 56395, 67322, 76780)
```

Data import

```
pc_raw <-
    read_excel("Plant traits/TRYdb/Paul C TRYdb/TRY_db_all data PEAS.xlsx",
               col_types = c("text", "text", "numeric", "text", "text", "numeric", "text",
                             "numeric", "numeric", "numeric", "text", "numeric", "text",
                             "text", "text", "text", "text", "text", "text", "text",
                             "numeric", "text", "text", "text", "numeric", "text")) %>%
    mutate(Comment = NA)  # New TRYdb files have a 'Comment' column, which this one lacks.

d1_raw <-
    read_tsv("Plant traits/TRYdb/1st TRY retrieval/3869.txt",
             col_types = "ccicciciiiciccccccccncccnccc") %>%
    select(-X28)  # This column was created by unwanted trailing tabs on some lines.

d2_raw <-
    read_tsv("Plant traits/TRYdb/2nd TRY retrieval/4655.txt",
             col_types = "ccicciciiiciccccccccncccnccc") %>%
    select(-X28)

identical(colnames(pc_raw), colnames(d1_raw))
```

```
## [1] TRUE
```

```
identical(colnames(d1_raw), colnames(d2_raw))
```

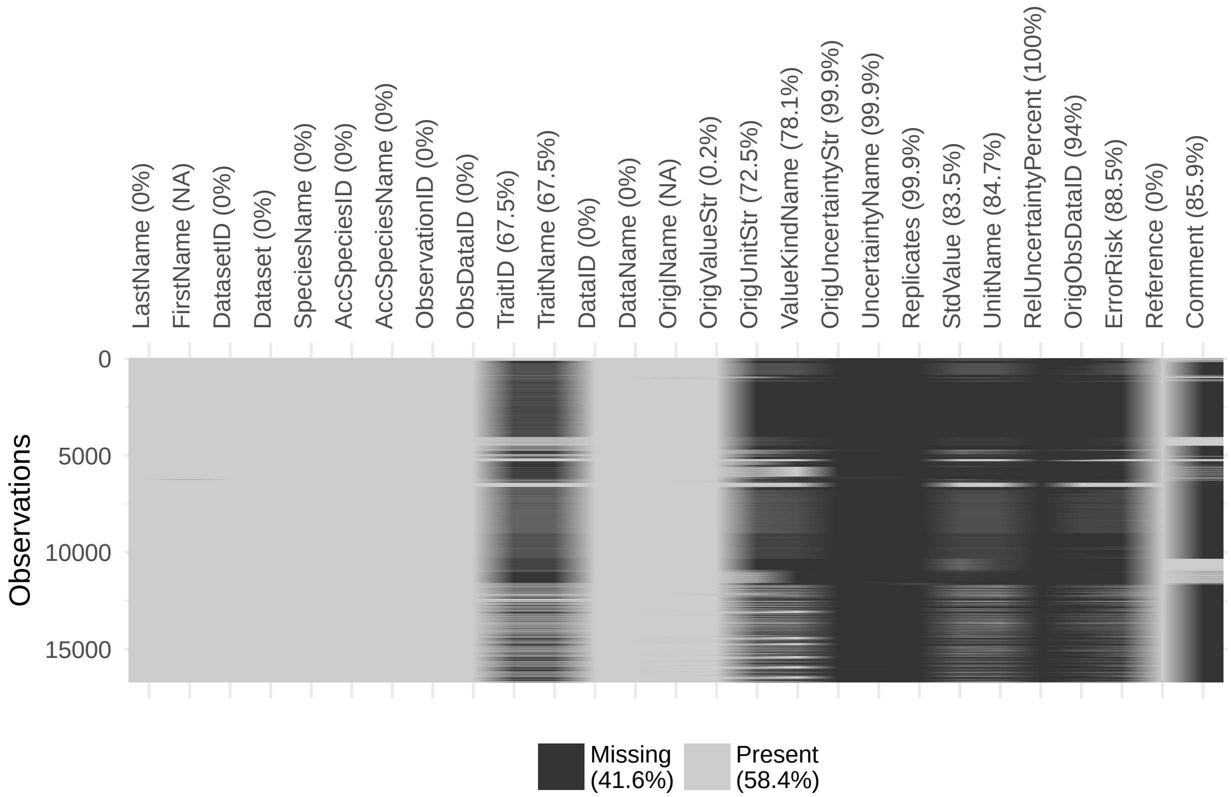
```
## [1] TRUE
```

```
# I can row-bind safely because the columns are identical.

trydb_raw <-
    bind_rows(d1_raw, d2_raw, pc_raw)
```

Missing data

```
vis_miss(trydb_raw) + rotate_x_text(angle = 90, align = 0)
```



Four columns are 100% missing and will be removed. SpeciesName and OrigObsDataID are also of little value now.

The patterns of missing data in TraitID and TraitName are caused by the data in these fields being broken into multiple rows, but with the TraitID and TraitName not being repeated:

TraitID	TraitName	DataID	DataName
26	Seed dry mass	30	Seed dry mass
		298	Weight precision
		183	Seed mass comment
		113	Reference / source
26	Seed dry mass	30	Seed dry mass
		298	Weight precision
		183	Seed mass comment
		235	Comment
		113	Reference / source
98	Seed storage behaviour	299	Seed storage behaviour

It seems that the most expedient way to get the actual trait values and avoid the extra guff is to just filter out rows that don't have a `TraitID`. Using `tidyr::fill()` to fill down the `NA`s is possible, but then I have to use `DataName` and other columns to decide which ones to actually keep.

```r
trydb <-
    trydb_raw %>%
    select(-(OrigUncertaintyStr:Replicates), -RelUncertaintyPercent,
           -SpeciesName, -OrigObsDataID) %>%
    filter(TraitID %in% requested_traits,
           AccSpeciesID %in% requested_plants) %>%
    distinct(DatasetID, AccSpeciesID, ObservationID, ObsDataID, TraitID,
             DataID, .keep_all = TRUE)

unique(trydb$DataName)
```

```
##   [1] "Seedbank location"
##   [2] "Germination stimulation"
##   [3] "Seedbank longevity"
##   [4] "Flower color"
##   [5] "Fruit/Seed Color"
##   [6] "Cold Stratification Required"
##   [7] "Fertility Requirement"
##   [8] "Fruit/Seed Abundance"
##   [9] "Seedling Vigor"
##  [10] "Seed dry mass"
##  [11] "Seed storage behaviour"
##  [12] "Plant growth form"
##  [13] "Plant longevity, plant maximum age"
##  [14] "Whole plant dry mass of reproductive structures per individual plant"
##  [15] "Seed number per plant"
##  [16] "Onset of flowering (First Flowering Date, Fowering beginning)"
##  [17] "Onset of seed maturation"
##  [18] "Shoot growth form"
##  [19] "Fruit type"
##  [20] "Resprouting capacity after fire (fire response)"
##  [21] "Post fire seedling emergence"
##  [22] "Post fire seedlings survival"
##  [23] "Fire regeneration category"
##  [24] "Regeneration after fire by seeds and/or resprouts"
##  [25] "Tolerance to frequent fires"
##  [26] "Tolerance to infrequent fires"
##  [27] "Plant age at first flowering (primary juvenil period)"
##  [28] "Age of maturity of resprouts"
##  [29] "Flowering season"
##  [30] "Recommended minimimum fire interval"
##  [31] "Plant photosynthetic pathway"
##  [32] "End of flowering"
##  [33] "Climbing mode"
##  [34] "Flower sexual system"
##  [35] "Self-incompatibility"
##  [36] "Flower pollen ovule ratio"
##  [37] "Flower: pollinator and type of reward"
##  [38] "Flower type"
##  [39] "Flower sex timing"
##  [40] "Seed thickness"
##  [41] "Seed width (middle dimension length)"
##  [42] "Seed length (largest dimension length)"
##  [43] "Flower UV reflectance in periphery"
##  [44] "Flower UV reflectance in centrum"
##  [45] "Flower UV reflectance pattern"
##  [46] "Active Growth Period"
##  [47] "Bloat"
##  [48] "Fire Resistant"
##  [49] "Growth Form"
##  [50] "Plant Growth Rate"
##  [51] "Known Allelopath"
##  [52] "Leaf Retention"
##  [53] "Plant life span"
##  [54] "Low Growing Grass"
##  [55] "Plant resprouting capacity after disturbance"
##  [56] "Shape and Orientation"
##  [57] "Toxicity"
##  [58] "Tolerance to drought"
##  [59] "Fire Tolerance"
##  [60] "Moisture Use"
##  [61] "Precipitation, Minimum"
##  [62] "Precipitation, Maximum"
##  [63] "Fruit/Seed Period Begin"
##  [64] "Fruit/Seed Period End"
##  [65] "Fruit/Seed Persistence"
##  [66] "Palatable Browse Animal"
##  [67] "Palatable Human"
##  [68] "Palatable Graze Animal"
##  [69] "Total plant nitrogen content per dry mass"
##  [70] "Growth per nitrogen content in whole plant"
##  [71] "Growth per nitrogen content in foliage"
##  [72] "Seasonality of growth"
##  [73] "Seed shedding season (time of seed dispersal)"
##  [74] "Dormancy"
##  [75] "Fire mortality"
##  [76] "Plant palatability"
##  [77] "Spinescence / thorniness"
##  [78] "Secondary compounds"
##  [79] "Whole Plant Dry Mass"
##  [80] "Plant functional type PFT (Sheffield DGVM 1)"
##  [81] "Plant functional type PFT (Sheffield DGVM 2)"
##  [82] "Plant functional type PFT (Sheffield DGVM 3)"
##  [83] "Plant functional type PFT (Biome-BGC 1)"
##  [84] "Plant functional type PFT (Biome-BGC 2)"
##  [85] "Plant functional type PFT (LPJ DGVM 1)"
##  [86] "Plant functional type PFT (LPJ DGVM 2)"
##  [87] "Light requirement"
##  [88] "Perennation 1 (plant age)"
##  [89] "Time (season) of germination (seedling emergence)"
##  [90] "Germination type"
##  [91] "Seedbank type"
##  [92] "Seedbank density"
##  [93] "Physical defences on leaves"
##  [94] "BudBank >0 to 10 cm: seasonality"
##  [95] "BudBank soil surface: seasonality"
##  [96] "BudBank <0 to -10 cm: seasonality"
##  [97] "BudBank >10 cm: seasonality"
##  [98] "Seed longevity"
##  [99] "Seed longevity: Fraction of plots with persistent seeds"
## [100] "Annual radial growth"
## [101] "Seed morphology"
## [102] "Buds seasonality aboveground"
## [103] "Buds seasonality at soil surface"
## [104] "Buds seasonality belowground"
## [105] "Buds seasonality belowground (at soil layer -10 - 0 cm))"
## [106] "Seed structure hooked"
## [107] "Seedbank number of layers"
## [108] "Seedbank thickness of top layer (cm)"
## [109] "Seedbank duration (month)"
## [110] "Seed longevity (max possible)"
## [111] "Seedbank longevity index"
## [112] "Seed number per ramet/tussock or individual plant"
## [113] "Seed number per single flower inflorescence"
## [114] "Plant functional type (JULES)"
## [115] "Plant functional type (Ian Wright)"
## [116] "Plant functional type (JULES plus ev/dec)"
## [117] "Plant functional type (Sitch, Harper, Mercado)"
## [118] "Plant growth form: climber"
## [119] "Plant functional type PFT (Poulter, B)"
## [120] "Plant growth form standardized"
## [121] "Plant growth form attributed diversity"
## [122] "Plant growth form consensus"
## [123] "Photosynthetic pathway"
## [124] "Gymnosperm/Angiosperm/Tree/Herb"
```

```r
unique(trydb$AccSpeciesName)
```

```
##  [1] "Bossiaea ensata"           "Bossiaea heterophylla"
##  [3] "Bossiaea obcordata"        "Daviesia ulicifolia"
##  [5] "Gompholobium grandiflorum" "Hardenbergia violacea"
##  [7] "Hovea linearis"            "Indigofera australis"
##  [9] "Mirbelia platylobioides"   "Acacia brownii"
## [11] "Acacia elongata"           "Acacia fimbriata"
## [13] "Acacia floribunda"         "Acacia gunnii"
## [15] "Acacia hispidula"          "Acacia linifolia"
## [17] "Acacia myrtifolia"         "Acacia obtusifolia"
## [19] "Acacia oxycedrus"          "Acacia parramattensis"
## [21] "Acacia suaveolens"         "Acacia terminalis"
## [23] "Acacia trinervata"         "Acacia ulicifolia"
## [25] "Daviesia alata"            "Daviesia corymbosa"
## [27] "Daviesia latifolia"        "Daviesia mimosoides"
## [29] "Dillwynia retorta"         "Gompholobium glabratum"
## [31] "Gompholobium latifolium"   "Mirbelia rubiifolia"
## [33] "Podolobium ilicifolium"    "Pultenaea scabra"
## [35] "Acacia longifolia"         "Vicia sativa"
## [37] "Acacia implexa"            "Acacia rubida"
## [39] "Acacia dealbata"           "Acacia buxifolia"
## [41] "Acacia podalyriifolia"     "Hardenbergia comptoniana"
## [43] "Acacia baileyana"          "Acacia conferta"
## [45] "Glycine clandestina"       "Dillwynia floribunda"
## [47] "Dillwynia brunioides"      "Pultenaea daphnoides"
## [49] "Platylobium formosum"      "Bossiaea rhombifolia"
## [51] "Glycine tabacina"          "Hovea heterophylla"
## [53] "Acacia binervia"           "Desmodium varians"
## [55] "Gompholobium huegelii"     "Gompholobium pinnatum"
## [57] "Pultenaea parviflora"      "Acacia parvipinnula"
## [59] "Dillwynia elegans"         "Pultenaea tuberculata"
## [61] "Pultenaea ferruginea"
```

Remove traits that are invariant or not represented in enough species

```
trait_variation <-
    trydb %>%
    select(TraitName, DataName, OrigValueStr, AccSpeciesID) %>%
    group_by(TraitName, DataName) %>%
    summarise(unique_values = length(unique(OrigValueStr)),
              species_covered = length(unique(AccSpeciesID))) %>%
    arrange(unique_values, species_covered)

length(unique(trydb$DataName))
```

```
## [1] 124
```

```
trydb <-
    trydb %>%
    semi_join(filter(trait_variation, unique_values > 1),
              by = c("TraitName", "DataName"))

length(unique(trydb$DataName))
```

```
## [1] 70
```

Excluding traits with lots of missing species

I requested 62 species but there's not a single trait that is recorded for every one of them. I will only keep traits that are recorded for at least 85 % of species (this cut-off is arbitrary).

```
trait_record <-
    trydb %>%
    distinct(AccSpeciesID, TraitID, DataID) %>%
    count(TraitID, DataID) %>%
    left_join(select(distinct(trydb, TraitID, DataID, .keep_all = TRUE),
                     TraitID, DataID, TraitName, DataName),
              by = c("TraitID", "DataID"))

percentile(trait_record$n)
```

```
##    0%   10%   20%   25%   33%   50%   66%   75%   80%   85%   90%   95%
##  1.00  1.00  1.00  1.25  2.00  2.00  4.00  5.75 14.20 23.95 47.50 52.00
##   99%  100%
## 53.00 53.00
```

```
trait_record <-
    trait_record %>%
    filter(n > percentile(trait_record$n)["85%"])

length(unique(trydb$DataName))
```

```
## [1] 70
```

```
trydb <-
    trydb %>%
    semi_join(trait_record, by = c("TraitID", "TraitName", "DataID", "DataName"))

length(unique(trydb$DataName))
```

```
## [1] 11
```

Table 2: Unique DataName values that are present for at least 85 percent of species.

TraitID	DataID	n	TraitName	DataName
26	30	53	Seed dry mass	Seed dry mass
159	480	53	Seedbank type	Seedbank location
34	38	52	Seed germination stimulation	Germination stimulation
318	776	52	Plant tolerance to fire	Regeneration after fire by seeds and/or resprouts
318	780	52	Plant tolerance to fire	Tolerance to frequent fires
318	781	52	Plant tolerance to fire	Tolerance to infrequent fires
318	790	52	Plant tolerance to fire	Fire regeneration category
335	786	47	Plant reproductive phenology timing	Flowering season
42	47	32	Plant growth form	Plant growth form
155	782	28	Plant ontogeny: age of maturity (first flowering)	Plant age at first flowering (primary juvenil period)
33	481	25	Seed (seedbank) longevity	Seedbank longevity

Manually edit traits

Some traits that supposedly have multiple values actually only have one, but it's modified in some basic ways. For example, `299, Seed storage behaviour` has 2 values, but those values are "Orthodox" and "Orthodox?". I kept these because they actually are informative, they just need to be edited.

```
trydb <-
    trydb %>%
    # These two traits need to be made uniform.
    mutate(OrigValueStr = case_when(DataName == "Germination stimulation" ~ "heat",
                                    DataName == "Seedbank location" ~ "soil",
                                        TRUE ~ OrigValueStr)) %>%
    # This trait needs to be made numeric.
    mutate(OrigValueStr = ifelse(DataName == "Seedbank longevity",
                                 case_when(OrigValueStr == "persistent/unk" ~ "55",
                                           OrigValueStr == ">55" ~ "55",  # Mode is 55.
                                           OrigValueStr == "greater than 5 years." ~ "55",
                                           TRUE ~ OrigValueStr),
                                 OrigValueStr)) %>%
    filter(OrigValueStr != "not available")  # One row in 'Flowering season'.
```

Final filtered dataset

Out of 103 requested traits for 62 requested plant species, I am left with 8 traits (with 11 sub-traits) for 60 species.

Table 3: Traits and sub-traits that survived filtering.

Trait	Sub-trait	Spp recorded
Seed dry mass	Seed dry mass	53
Seedbank type	Seedbank location	53
Plant tolerance to fire	Fire regeneration category	52
Plant tolerance to fire	Regeneration after fire by seeds and/or resprouts	52
Plant tolerance to fire	Tolerance to frequent fires	52
Plant tolerance to fire	Tolerance to infrequent fires	52
Seed germination stimulation	Germination stimulation	52
Plant reproductive phenology timing	Flowering season	47
Plant growth form	Plant growth form	32
Plant ontogeny: age of maturity (first flowering)	Plant age at first flowering (primary juvenil period)	28
Seed (seedbank) longevity	Seedbank longevity	25

```
##  [1] "Acacia baileyana"          "Acacia binervia"
##  [3] "Acacia brownii"            "Acacia buxifolia"
##  [5] "Acacia conferta"           "Acacia dealbata"
##  [7] "Acacia elongata"           "Acacia fimbriata"
##  [9] "Acacia floribunda"         "Acacia gunnii"
## [11] "Acacia hispidula"          "Acacia implexa"
## [13] "Acacia linifolia"          "Acacia longifolia"
## [15] "Acacia myrtifolia"         "Acacia obtusifolia"
## [17] "Acacia oxycedrus"          "Acacia parramattensis"
## [19] "Acacia parvipinnula"       "Acacia podalyriifolia"
## [21] "Acacia rubida"             "Acacia suaveolens"
## [23] "Acacia terminalis"         "Acacia trinervata"
## [25] "Acacia ulicifolia"         "Bossiaea ensata"
## [27] "Bossiaea heterophylla"     "Bossiaea obcordata"
## [29] "Bossiaea rhombifolia"      "Daviesia alata"
## [31] "Daviesia corymbosa"        "Daviesia latifolia"
## [33] "Daviesia mimosoides"       "Daviesia ulicifolia"
## [35] "Desmodium varians"         "Dillwynia brunioides"
## [37] "Dillwynia elegans"         "Dillwynia floribunda"
## [39] "Dillwynia retorta"         "Glycine clandestina"
## [41] "Glycine tabacina"          "Gompholobium glabratum"
## [43] "Gompholobium grandiflorum" "Gompholobium huegelii"
## [45] "Gompholobium latifolium"   "Gompholobium pinnatum"
## [47] "Hardenbergia comptoniana"  "Hardenbergia violacea"
## [49] "Hovea linearis"            "Indigofera australis"
## [51] "Mirbelia platylobioides"   "Mirbelia rubiifolia"
## [53] "Platylobium formosum"      "Podolobium ilicifolium"
## [55] "Pultenaea daphnoides"      "Pultenaea ferruginea"
## [57] "Pultenaea parviflora"      "Pultenaea scabra"
## [59] "Pultenaea tuberculata"     "Vicia sativa"
```

```r
write_csv(trydb, make_path("_compiled/try.csv"))
write_rds(trydb, make_path("_compiled/try.rds"))
```