REDCap Moonpie Vignette

2022-05-16

Motivation

REDCap is a data collection tool. Data can be exported and read into R as a CSV file. Some basic manipulation (like setting variable labels) can be performed by downloading an R script also provided through the REDCap system. Parsing this R script can provide a data dictionary, which the package will provide in additional formats (YAML, CSV). This package will allow a user to modify the data dictionary which will update the R script used for importing the REDCap data.

Functionality

Use `script2info` to build a data dictionary (list) from the REDCap provided R script. While you can examine the list created (for example with `str`), you will likely want to use the `dd2df` function to convert the list to a data.frame.

```r
library(rcmoonpie)
dd <- script2info(system.file("examples", "ex_script.R", package = "rcmoonpie"))
dd2df(dd)
```

```
##    name             variable                label exclude    level label.factor exclude.factor
## 1  data            record_id            Record ID   FALSE     <NA>         <NA>           TRUE
## 2  data    redcap_event_name           Event Name   FALSE baseline     Baseline          FALSE
## 3  data    redcap_event_name           Event Name   FALSE followup    Follow-up          FALSE
## 4  data           visit_date           Visit Date   FALSE     <NA>         <NA>           TRUE
## 5  data        randomization        Randomization   FALSE        0      Control          FALSE
## 6  data        randomization        Randomization   FALSE        1    Treatment          FALSE
## 7  data                  sex                  Sex   FALSE        0         Male          FALSE
## 8  data                  sex                  Sex   FALSE        1       Female          FALSE
## 9  data                  sex                  Sex   FALSE      666    Ambiguous          FALSE
## 10 data                  sex                  Sex   FALSE      999      Missing          FALSE
## 11 data                  age                  Age   FALSE     <NA>         <NA>           TRUE
## 12 data social_connectedness Social Connectedness   FALSE        0            0          FALSE
## 13 data social_connectedness Social Connectedness   FALSE        1            1          FALSE
## 14 data social_connectedness Social Connectedness   FALSE        2            2          FALSE
## 15 data social_connectedness Social Connectedness   FALSE        3            3          FALSE
## 16 data social_connectedness Social Connectedness   FALSE        4            4          FALSE
## 17 data social_connectedness Social Connectedness   FALSE        5            5          FALSE
## 18 data social_connectedness Social Connectedness   FALSE        6            6          FALSE
## 19 data social_connectedness Social Connectedness   FALSE        7            7          FALSE
## 20 data social_connectedness Social Connectedness   FALSE      666    Ambiguous          FALSE
## 21 data social_connectedness Social Connectedness   FALSE      999      Missing          FALSE
## 22 data             comments             Comments   FALSE     <NA>         <NA>           TRUE
```

You can export your data dictionary to YAML or CSV with the `dd2yaml` and `dd2csv` functions. Set the "file" argument to a file location, or alternatively allow it to print to standard output.

YAML and CSV can be imported back into a data dictionary with the `yaml2info` and `csv2info` functions.

```r
dd2yaml(dd)
```

```
## dataset:
## - name: data
##   variables:
##   - name: record_id
##     label: Record ID
##   - name: redcap_event_name
##     label: Event Name
##     factor:
##     - label: Baseline
##       level: baseline
##     - label: Follow-up
##       level: followup
##   - name: visit_date
##     label: Visit Date
##   - name: randomization
##     label: Randomization
##     factor:
##     - label: Control
##       level: '0'
##     - label: Treatment
##       level: '1'
##   - name: sex
##     label: Sex
##     factor:
##     - label: Male
##       level: '0'
##     - label: Female
##       level: '1'
##     - label: Ambiguous
##       level: '666'
##     - label: Missing
##       level: '999'
##   - name: age
##     label: Age
##   - name: social_connectedness
##     label: Social Connectedness
##     factor:
##     - label: '0'
##       level: '0'
##     - label: '1'
##       level: '1'
##     - label: '2'
##       level: '2'
##     - label: '3'
##       level: '3'
##     - label: '4'
##       level: '4'
##     - label: '5'
##       level: '5'
##     - label: '6'
##       level: '6'
##     - label: '7'
##       level: '7'
##     - label: Ambiguous
##       level: '666'
##     - label: Missing
##       level: '999'
##   - name: comments
##     label: Comments
```

```r
td <- tempdir()
dd2csv(dd, file.path(td, 'test.csv'))
dd_alt <- csv2info(file.path(td, 'test.csv'))
# exporting to CSV and back will have equivalent objects
all.equal(dd, dd_alt)
```

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

Variables and factor levels can be removed with `excludeVar` and `excludeLevel`. Excluding can be undone with `unexcludeVar` and `unexcludeLevel`. Note that you can also accomplish exclusion by modifying YAML or CSV output.

```r
dd <- excludeVar(dd, 'data', 'redcap_event_name')
dd <- excludeLevel(dd, 'data', 'sex', 666)
dd2df(dd)
```

```
##    name             variable                label exclude    level label.factor exclude.factor
## 1  data            record_id            Record ID   FALSE     <NA>         <NA>           TRUE
## 2  data    redcap_event_name           Event Name    TRUE baseline     Baseline          FALSE
## 3  data    redcap_event_name           Event Name    TRUE followup    Follow-up          FALSE
## 4  data           visit_date           Visit Date   FALSE     <NA>         <NA>           TRUE
## 5  data        randomization        Randomization   FALSE        0      Control          FALSE
## 6  data        randomization        Randomization   FALSE        1    Treatment          FALSE
## 7  data                  sex                  Sex   FALSE        0         Male          FALSE
## 8  data                  sex                  Sex   FALSE        1       Female          FALSE
## 9  data                  sex                  Sex   FALSE      666    Ambiguous           TRUE
## 10 data                  sex                  Sex   FALSE      999      Missing          FALSE
## 11 data                  age                  Age   FALSE     <NA>         <NA>           TRUE
## 12 data social_connectedness Social Connectedness   FALSE        0            0          FALSE
## 13 data social_connectedness Social Connectedness   FALSE        1            1          FALSE
## 14 data social_connectedness Social Connectedness   FALSE        2            2          FALSE
## 15 data social_connectedness Social Connectedness   FALSE        3            3          FALSE
## 16 data social_connectedness Social Connectedness   FALSE        4            4          FALSE
## 17 data social_connectedness Social Connectedness   FALSE        5            5          FALSE
## 18 data social_connectedness Social Connectedness   FALSE        6            6          FALSE
## 19 data social_connectedness Social Connectedness   FALSE        7            7          FALSE
## 20 data social_connectedness Social Connectedness   FALSE      666    Ambiguous          FALSE
## 21 data social_connectedness Social Connectedness   FALSE      999      Missing          FALSE
## 22 data             comments             Comments   FALSE     <NA>         <NA>           TRUE
```

Rather than passing `excludeVar` a variable name, a regular expression pattern can be used to exclude all variables matching the given pattern. You can check which columns a pattern includes with `findVariableByPattern`.

```r
# which variables include an underscore character
findVariableByPattern(dd, '_')
```

```
## [1] "record_id"            "redcap_event_name"    "visit_date"           "social_connectedness"
```

```r
# exclude all variables that have an "_" in the name
dd2df(excludeVar(dd, 'data', '_'))
```

```
##    name             variable                label exclude    level label.factor exclude.factor
## 1  data            record_id            Record ID    TRUE     <NA>         <NA>           TRUE
## 2  data    redcap_event_name           Event Name    TRUE baseline     Baseline          FALSE
## 3  data    redcap_event_name           Event Name    TRUE followup    Follow-up          FALSE
## 4  data           visit_date           Visit Date    TRUE     <NA>         <NA>           TRUE
## 5  data        randomization        Randomization   FALSE        0      Control          FALSE
## 6  data        randomization        Randomization   FALSE        1    Treatment          FALSE
## 7  data                  sex                  Sex   FALSE        0         Male          FALSE
## 8  data                  sex                  Sex   FALSE        1       Female          FALSE
## 9  data                  sex                  Sex   FALSE      666    Ambiguous           TRUE
## 10 data                  sex                  Sex   FALSE      999      Missing          FALSE
## 11 data                  age                  Age   FALSE     <NA>         <NA>           TRUE
## 12 data social_connectedness Social Connectedness    TRUE        0            0          FALSE
## 13 data social_connectedness Social Connectedness    TRUE        1            1          FALSE
## 14 data social_connectedness Social Connectedness    TRUE        2            2          FALSE
## 15 data social_connectedness Social Connectedness    TRUE        3            3          FALSE
## 16 data social_connectedness Social Connectedness    TRUE        4            4          FALSE
## 17 data social_connectedness Social Connectedness    TRUE        5            5          FALSE
## 18 data social_connectedness Social Connectedness    TRUE        6            6          FALSE
## 19 data social_connectedness Social Connectedness    TRUE        7            7          FALSE
## 20 data social_connectedness Social Connectedness    TRUE      666    Ambiguous          FALSE
## 21 data social_connectedness Social Connectedness    TRUE      999      Missing          FALSE
## 22 data             comments             Comments   FALSE     <NA>         <NA>           TRUE
```

The final output will be a new R script to replace the original downloaded from REDCap. It will reflect any changes to the data dictionary. Use the `dd2script` function to create the R script. Like `dd2yaml` it has a "file" argument that can be set to a location or left blank. It also has an argument "factorHandle" that can be used to change factor variable behavior. This can be set to one of three values:

- duplicate - The original character string variable will be copied to a new factor variable that includes ".factor" at the end of the variable name. This is the default.
- unchanged - The original character string variable will not be turned into a factor variable.
- changed - The original character string variable will be turned into a factor variable.

```r
dd2script(dd)
```

```
## label(data$record_id) = "Record ID"
## label(data$visit_date) = "Visit Date"
## label(data$randomization) = "Randomization"
## label(data$sex) = "Sex"
## label(data$age) = "Age"
## label(data$social_connectedness) = "Social Connectedness"
## label(data$comments) = "Comments"
##
## data$redcap_event_name = NULL
##
## data$randomization.factor = factor(data$randomization, levels = c("0","1"))
## data$sex.factor = factor(data$sex, levels = c("0","1","999"))
## data$social_connectedness.factor = factor(data$social_connectedness, levels = c("0","1","2","3","4",
##
## levels(data$randomization.factor) = c("Control","Treatment")
## levels(data$sex.factor) = c("Male","Female","Missing")
## levels(data$social_connectedness.factor) = c("0","1","2","3","4","5","6","7","Ambiguous","Missing")
```

```r
dd2script(dd, factorHandle = 'unchanged')
```

```
## label(data$record_id) = "Record ID"
## label(data$visit_date) = "Visit Date"
## label(data$randomization) = "Randomization"
## label(data$sex) = "Sex"
## label(data$age) = "Age"
## label(data$social_connectedness) = "Social Connectedness"
## label(data$comments) = "Comments"
##
## data$redcap_event_name = NULL
```

```r
dd2script(dd, factorHandle = 'changed')
```

```
## label(data$record_id) = "Record ID"
## label(data$visit_date) = "Visit Date"
## label(data$randomization) = "Randomization"
## label(data$sex) = "Sex"
## label(data$age) = "Age"
## label(data$social_connectedness) = "Social Connectedness"
## label(data$comments) = "Comments"
##
## data$redcap_event_name = NULL
##
## data$randomization = factor(data$randomization, levels = c("0","1"))
## data$sex = factor(data$sex, levels = c("0","1","999"))
## data$social_connectedness = factor(data$social_connectedness, levels = c("0","1","2","3","4","5","6"
##
## levels(data$randomization) = c("Control","Treatment")
## levels(data$sex) = c("Male","Female","Missing")
## levels(data$social_connectedness) = c("0","1","2","3","4","5","6","7","Ambiguous","Missing")
```