

ingestr

reading environmental data from raw formats into dataframes

Rachael Blake, Joe Shannon, Tim Whiteaker

AGU Meeting, San Francisco
December 10, 2019

Package webpage <https://jpshanno.github.io/ingestr/>



Motivation



Environmental Data Initiative IMCR Hackathon 2018

Motivation



Environmental Data Initiative IMCR Hackathon 2018

Collaboratively design/develop software supporting information managers and scientists working with environmental data

ingestr guiding principles

ingestr guiding principles

- All sources of environmental-related data should be easy to read directly
- Reading in data should provide a standard output

ingestr guiding principles

- All sources of environmental-related data should be easy to read directly
- Reading in data should provide a standard output
- Header information should be stored in standard, easily readable format
- Data provenance records should be included throughout for reproducibility

Using `ingestr:verb` pattern

`ingest_*`

* is standard sensor or non-sensor data format ex: campbell

`ingest_header()`

`ingest_directory()`

Using `ingestr`: single file

```
library(ingestr)

campbell_data <- ingest_campbell(input.source = my_data_file, export.header = TRUE,
                                add.units = TRUE, add.measurements = TRUE)
```

```
## Header info for C:/Users/rblake/Documents/R/win-
library/3.6/ingestr/example_data/campbell_scientific_tao5.dat has been saved to a temporary file.
Run ingest_header('C:/Users/rblake/Documents/R/win-
library/3.6/ingestr/example_data/campbell_scientific_tao5.dat') to load the header data.
```


Using `ingestr`: single file

```
library(ingestr)
```

```
campbell_data <- ingest_campbell(input.source = my_data_file, export.header = TRUE,  
                                add.units = TRUE, add.measurements = TRUE)
```

```
## Header info for C:/Users/rblake/Documents/R/win-  
library/3.6/ingestr/example_data/campbell_scientific_tao5.dat has been saved to a temporary file.  
Run ingest_header('C:/Users/rblake/Documents/R/win-  
library/3.6/ingestr/example_data/campbell_scientific_tao5.dat') to load the header data.
```

```
str(campbell_data)
```

```
## 'data.frame':    1272 obs. of  24 variables:  
## $ TIMESTAMP      : POSIXct, format: "2014-11-02" "2014-11-03" ...  
## $ RECORD         : int  0 1 2 3 4 5 6 7 8 9 ...  
## $ Min_BattV_Min_Volts : num  13.5 13.2 13.2 13.3 13.4 ...  
## $ Max_PTemp_C_Max_Deg.C : num  0.989 11.02 12.88 8.61 4.064 ...  
## $ Min_PTemp_C_Min_Deg.C : num  -3.32 -1.14 1.01 1.84 1.22 ...  
## ...
```

Using `ingestr`: single file

```
campbell_header <- ingest_header(input.source = my_data_file)
```

```
## Header data was loaded from cached results created when C:/Users/rblake/Documents/R/win-  
library/3.6/ingestr/example_data/campbell_scientific_tao5.dat was ingested previously in this R  
session.
```

Using `ingestr`: single file

```
campbell_header <- ingest_header(input.source = my_data_file)
```

```
## Header data was loaded from cached results created when C:/Users/rblake/Documents/R/win-  
library/3.6/ingestr/example_data/campbell_scientific_tao5.dat was ingested previously in this R  
session.
```

```
str(campbell_header)
```

```
## 'data.frame':    1 obs. of  9 variables:  
## $ file_type      : chr "TOA5"  
## $ logger_name    : chr "FRF_Village"  
## $ logger_model   : chr "CR1000"  
## $ logger_serial_number : int 63162  
## $ logger_os_version : chr "CR1000.Std.27"  
## ...
```

Using `ingestr`: many files

```
temperature_data <- ingest_directory(directory = my_directory,  
                                     ingest.function = ingest_campbell,  
                                     pattern = ".dat")
```

```
## Header info for C:/Users/rblake/Documents/R/win-  
library/3.6/ingestr/example_data/campbell_directory has been saved to a temporary file. Run  
ingest_header('C:/Users/rblake/Documents/R/win-  
library/3.6/ingestr/example_data/campbell_directory') to load the header data.  
  
## Unable to restore lost column attributes when the datasets were combined.
```

Using `ingestr`: many files

```
temperature_data <- ingest_directory(directory = my_directory,  
                                     ingest.function = ingest_campbell,  
                                     pattern = ".dat")
```

```
## Header info for C:/Users/rblake/Documents/R/win-  
library/3.6/ingestr/example_data/campbell_directory has been saved to a temporary file. Run  
ingest_header('C:/Users/rblake/Documents/R/win-  
library/3.6/ingestr/example_data/campbell_directory') to load the header data.  
  
## Unable to restore lost column attributes when the datasets were combined.
```

```
str(temperature_data)
```

```
## 'data.frame':   3816 obs. of  24 variables:  
## $ TIMESTAMP      : POSIXct, format: "2014-11-02" "2014-11-03" ...  
## $ RECORD         : int  0 1 2 3 4 5 6 7 8 9 ...  
## $ Min_BattV_Min_Volts : num  13.5 13.2 13.2 13.3 13.4 ...  
## $ Max_PTemp_C_Max_Deg.C : num  0.989 11.02 12.88 8.61 4.064 ...  
## $ Min_PTemp_C_Min_Deg.C : num  -3.32 -1.14 1.01 1.84 1.22 ...  
...  
...  
...
```

Currently ingested

Sensors

- Campbell Scientific
- PP Systems EGM-4
- Solinst Levellogger

Non-sensors

- El Nino Southern Oscillation
- Pacific Decadal Oscillation
- North Pacific Gyre Oscillation

Future plans

- `ingest_*` function for Tecan i-control plate reader (Spring 2020)
- add more `ingest_*` functions: please request an ingest function and contribute example data in the issues; you can also put in a pull request.
- peer-reviewed via rOpenSci, and published on CRAN

Future plans

- `ingest_*` function for Tecan i-control plate reader (Spring 2020)
- add more `ingest_*` functions: please request an ingest function and contribute example data in the issues; you can also put in a pull request.
- peer-reviewed via rOpenSci, and published on CRAN

Thanks!

Thanks to EDI IMCR Hackathon 2018 for bringing the authors together, and for supporting initial development of this package.

Thanks to all our contributors who have helped improve this package via pull requests, issues, and discussion.

Slides made with the R package `xaringan` created by yihui.

Please contribute!

<https://jpshanno.github.io/ingestr/CONTRIBUTING>



Package webpage <https://jpshanno.github.io/ingestr/>

Please contribute!

<https://jpshanno.github.io/ingestr/CONTRIBUTING>



Also ask me about rslurm!

<http://cyberhelp.sesync.org/rslurm/>



Package webpage <https://jpshanno.github.io/ingestr/>