Annotating Data Sets

Nicholas J. Gotelli

February 2, 2017

Annotation of data sets with proper meta-data is essential. The best way to do this is to embed the meta-data at the start of the data file itself. In R, this is easy to do by using the `#` comment and then whatever follows in that line will be skipped when R reads in the data. Here is a template for meta-data and the data that follows:

```
#//////////////////////////////////
# -----START OF METADATA -----
# ------------------------------------------------
# TITLE: Brief title for entire data set
# DATE: date of creation of data set
# AUTHOR: author/owner of data
# ------------------------------------------------
# AUTHOR EMAIL:
# AUTHOR ADDRESS:
# AUTHOR WEBSITE:
# ------------------------------------------------
# OWNERSHIP: Information on who owns/controls the data
# COLLABORATORS: One or more lines identifying others who collected/own data
# FUNDING SOURCES: Grant numbers or funding sources for acknowlegment
# REPOSITORY: One or more lines for GitHub, Dryad, or permanent web repositories whe
re data set can be accessed
# CITATIONS: One or more lines for publications that cite or use these data
# ------------------------------------------------
# SAMPLING LOCATIONS: One or more lines about where data were collected; GPS for ind
ividual sites should be data columns
# SAMPLING TIMES: One more lines on when data were collected
# VARIABLE DESCRIPTION: One line for each column in the data set stating what it is
and what the units of measurement are
# MISSING DATA: One or more lines for each variable describing the source of NA valu
es throughout
# ------------------------------------------------
# DATA TRACK CHANGES LOG (use this section to record any changes to the data set aft
er it is created)
# DATE:              # CHANGES:
# DATE:              # CHANGES:
# DATE:              # CHANGES:
#DATE:             #CHANGES:
# ------------------------------------------------
# ----- END OF METADATA -----
#//////////////////////////////////
#
#
# ----- START OF DATA -----
#
ID VarName1 VarName2 VarName3



# Can insert comments throughout text to add notes for individual data rows


# ----- END OF DATA -----
```

If you save this as a `.csv` file from Excel, you should first be aware that Excel will mangle your text, introducing quotation marks and optional extra commas in each comment row. The file will need to be

cleaned up in a text editor, and then you should avoid opening it again in Excel.

When the file is cleaned up, R should be able to read it, but if you try the following, the file will not be read properly:

```
read.csv("Input_File.csv",header=TRUE,row.names=1)
```

Perversely, the `read.csv` command has the comment character disabled. You could do this by skipping the proper number of lines:

```
read.csv("Input_Data.csv",header=TRUE,row.names=1,skip=10)
```

This will indeed skip the first 10 lines of this file and begin reading where it should. However, each data set would need to be hand-wired in this way, which is problematic for batch processing.

A better solution is to use the `read.table` command, and insert the needed delimiter:

```
read.table("Input_File.csv",header=TRUE,row.names=1,sep=",",stringsAsFactors = FALSE
)
```

In this way, all of the comments are skipped, and you don't have to worry about how many lines of meta-data are contained in the file. Always keep `stringsAsFactors=FALSE` so that strings come in as characters at first.