Introducing SimpleDataManager
A simple data management workflow for R

Finn Sandø¹,*; Camilla Trab Damsgaard¹; Stine-Mathilde Dalskov¹; Christian Ritz¹

1. Department of Nutrition, Exercise and Sports, Faculty of Science, University of Copenhagen, Denmark
*Contact author: fsa@life.ku.dk

Keywords: data management workflow, data formats, data hub structure, data storage

The aim of SimpleDataManager is to facilitate a data management workflow in a simple, easy to use fashion. Therefore SimpleDataManager does not intend to support analytical features except for some very simple validation functionality. The motivation behind SimpleDataManager is that many projects have a specific needs for data management that are better dealt with as separate tasks not tightly integrated into the subsequent analytical process. There have over the years been a few attempts at making such integrated workflow platforms into R packages. Some of these come with a graphical user interface, most notably John Fox’ Rcmdr. A few other packages are currently available in the CRAN repository: R2 (a "GUI Tool for Data Management like SPSS or Stata") and memisc ("Tools for Management of Survey Data, Graphics, Programming, Statistics, and Simulation"). These packages support data management, but are more focused on analytics and graphs. Some other data management initiatives have been developed within very specific contexts like eiPack ("Ecological Inference and Higher-Dimension Data Management").

Regarding ‘raw’ data management, a typical workflow could be as follows: Data are generated or obtained by one or more individuals from one or more sources. These data are collected centrally, some kind of clean-up, validation, and standardization steps are performed, and the data are merged into one or a few related datasets. An official standard for long-term storage may apply. Finally, there may be restrictions limiting who has access to which parts of the data, and usually there is a need to deliver in specific formats (spss, sas, stata, csv).

SimpleDataManager supports such a workflow. On the input side it operates with a repository of incoming ‘raw’ data, it promotes a unified and repeatable workflow for cleaning, validating, transforming, and merging data into a consistent format. On the output side it uses a request mechanism where recipients select required variables on an auto-generated list. Variables can be grouped to simplify the selection process. It supports missing values, multi-lingual variable and value labels defined by the user. Data can be exported to common formats (sas, spss, stata, csv), new formats can relatively easily be implemented. We are currently working on an extension that will make it possible to deliver data as a single RData file which will come with built-in functionality for export to all supported formats.

SimpleDataManager follows the configuration by convention pattern where, if the user adheres to a specific structure, most things will work ‘automagically’, it is even possible to configure the standard behavior with configuration files. SimpleDataManager will be made available through one of the public R repositories.

SimpleDataManager is developed as part of the large-scale research project OPUS (Optimal well-being, development and health for Danish children through a healthy New Nordic Diet), supported by a grant from the Nordea Foundation’.