The Reproducible Computing package

Patrick Wessa*, Ed van Stee

1. K.U.Leuven Association, Lessius Dept. of Business Studies, Belgium

* Contact author: patrick@wessa.net

Keywords: Reproducible Computing, Research, Compendium

The problem of irreproducible research received a great deal of attention within the academic community [1], [2], [3], [4], [5], [6], [7]. Several solutions have been proposed ([5], [7], [8]) - most prominently there is an R package [9] called “Sweave” ([8]) which allows us to create a so-called Compendium: an integrated collection of text, code, and data (that allows the presented science to be reproduced). All the necessary documents that are needed to create the article (with embedded R code) and the data are contained in an archive file (preferably in tar.gz or zip format). A few examples of such Compendia can be downloaded from [10].

This paper discusses a new approach towards reproducible computing by re-defining the concept of a Compendium as a document where each computation is referenced by a unique URL that points to an object which contains all the information that is necessary to recompute it [11], [12]. The key difference with the original definition of a Compendium is the fact that in our proposal there is a complete separation between text and computing. In other words, each computation (and associated meta data) is stored in a central, web-based repository that can be referenced from any text. Over the last two years, an easy-to-use Compendium Platform was developed (based on this new definition) and implemented in statistics education [11], [13].

The novelty about this paper is the introduction of a newly developed Reproducible Computing package which communicates with the Compendium Platform and allows the R user to do the following:

- store/retrieve image files
- archive/reproduce code snippets
- search archived objects in the repository
- data mining about archived objects
- convert code snippets into R modules (= web applications)
- protect data frames while computations are still reproducible,
- etc...

The bottom line about this package is that it allows the R user to quickly produce reproducible and reusable computations for the purpose of research and publishing.

References