Read, write, format Excel 2007 (xlsx) files

Adrian A. Drăgulescu

last compiled July 14, 2010; last modified on 2009-12-28

Abstract

The `xlsx` package provides tools neccessary to interact with Excel 2007 files from R . The user can read and write xlsx files, and can control the appearance of the spreadsheet by setting data formats, fonts, colors, borders. Set the print area, the zoom control, create split and freeze panels, adding headers and footers. The package uses a java library from the Apache POI project.

Contents

1 Introduction

The package `xlsx` makes possible to interact with Excel 2007 files from R . While a power R user usually does not need to use Excel or even avoids it altogether, there are cases when being able to generate Excel output or read Excel files into R is a must. One such case is in an office environment when you need to collaborate with co-workers who use Excel as their primary tool. Another case is to use Excel for basic reporting of results. For moderate data sets, the formatting capabilities of Excel can prove useful. A flexible way to manipulate Excel 2007 xlsx files from R would then be a nice addition.

The `xlsx` package focuses on Excel 2007 because for Excel 97 there are already several solutions, `RODBC`[2], `xlsReadWrite`[3], `RExcelInstaller`[4], etc. While some of these packages work with Excel 2007 files too, the contribution of the `xlsx` package is different.

The xlsx Excel 2007 file format is essentially a zipped set of xml files. It is possible to interact directly with these files from R as shown by the package `RExcelXML`[5]. All the functionality of the `xlsx` package can be replicated with `RExcelXML` package or by extending it. Working directly with the zipped xml files, and using the xml schema to extract the useful information into R gives you ultimate control.

The approach taken in the `xlsx` package is to use a proven, existing API between Java and Excel 2007 and use the `rJava` [6] package to link Java and R . The advantage of this approach is that the code on the R side is very compact, easy to maintain, and easy to extend even for people with little Java experience. All the heavy lifting of parsing XML schemas is being done in Java. We also benefit from a mature software project with many developers, test suites, and users that report issues on the Java side. In principle, this should make the maintainance of the `xlsx` package easy. The Java API used by the `xlsx` is one project of the Apache Software Foundation, called Apache POI and can be found at `http://poi.apache.org/`.

The Apache POI Project's mission is to create and maintain Java APIs for manipulating various file formats based upon the Office Open XML standards (OOXML) and Microsoft's OLE 2 Compound Document format (OLE2). These include Excel, Word, and PowerPoint documents. While the focus of the `xlsx` package has been only on Excel files, if there is interest in the R community, it can be easily extended for Word and PowerPoint documents.

2 High level API

See `read.xlsx` for reading the sheet of an xlsx file into a data.frame. See `write.xlsx` writing a data.frame to an xlsx file.

3 Low level API

See `Workbook` for creating workbooks. See `Worksheet` for code to manipulate worksheets. See `Cell` for manipulating cells.

See `OtherEffects` for various spreadsheet effects, for example, merge cells, auto size columns, create freeze panels, create split panels, set print area, set the zoom, etc.

Use `PrintSetup` for customizing the settings for printing.

3.1 Cell formatting

See `CellStyle` for how to format a particular cell.

Use `Font` to set a font.

4 Conclusions

By adding a lightweigth R layer on top of the Apache project Java interface to Excel 2007 documents, we achieve a multi-platform solution for interacting with Excel 2007 file formats from R .

References

[1] R Development Core Team: *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

[2] Brian Ripley and from 1999 to Oct 2002 Michael Lapsley (2009). *RODBC: ODBC Database Access*. R package version 1.3-1. http://CRAN.R-project.org/package=RODBC

[3] Hans-Peter Suter, Treetron and Switzerland (2006). *xlsReadWrite: Natively read and write Excel files.* http://treetron.googlepages.com/

[4] Erich Neuwirth, with contributions by Richard Heiberger, Christian Ritter, Jan Karel Pieterse and and Jurgen Volkering (2009). *RExcelInstaller: Integration of R and Excel, (use R in Excel, read/write XLS files).* R package version 3.0-18. http://CRAN.R-project.org/package=RExcelInstaller

[5] Duncan Temple Lang (2009). *RExcelXML: Read and manipulate new-style (Office '07) Excel files.* http://www.omegahat.org/RExcelXML/

[6] Simon Urbanek (2009). *rJava: Low-level R to Java interface.* R package version 0.8-1. http://CRAN.R-project.org/package=rJava