rTempo – an R package to access the TEMPO-Online database

Ana-Maria ȚÎRU (ana.tiru@insse.ro)
National Institute of Statistics, Romania

Iulia Elena TOMA, PhD. Candidate (iulia.toma92@gmail.com)
Bucharest University of Economic Studies, Romania

Marian NECULA (marian.necula@insse.ro)
National Institute of Statistics, Romania

ABSTRACT

In this paper we describe rTempo which is an R package that provides functions to access statistical data from TEMPO-Online database. Our package provides only the bulk download method since the TEMPO-Online database, Romanian National Institute of Statistics public database, does not have yet an API or a web service to access the data. Accessing data directly from R improves the efficiency of statistical production since it eliminates the tedious GUI. The communication between our package and TEMPO-Online database is achieved using the HTTP. Querying and searching for the data is done using rvest package. Besides the actual data, our package also provides facilities for downloading the metadata.

Keywords: *R, Tempo online, statistical database access, R packages.*
JEL Classification: *C18, C63, C88*

1. INTRODUCTION

According to Principle 15, part of the European Statistics Code of Practice, European Statistics must be "presented in a clear and understandable form, released in a suitable and convenient manner, available and accessible on an impartial basis with supporting metadata and guidance" (Eurostat, 2011). Moreover, according to Principle 11 of the European Statistics Code of Practice, European Statistics must "meet the needs of users" (Eurostat, 2011), needs who encompass fast and reliable information in areas such as housing markets (Vlag et al, 2009), financial markets (Jianu et al, 2014) or macroeconomic indicators such as GDP (UNSD, 2016). Consequently, rTempo is a package that can be used for ease and functionality without the need to know the underlying computations under the hood.

Many european countries have developed R packages in order to facilitate access to a wide variety of data. For example, R packages such as ”openPoland” and ”SmarterPoland” allow users to easily access and process various datasets from Central Statistical Office of Poland, Eurostat, API of Google Maps and other sources (Biecek, 2016). Other examples include R packages for accessing population health: ”fingertipsR” from England (Fox et al., 2017), hydrological data: ”hddtools” (Vitolo, 2017), International Monetary Fund data: ”IMFData” (Lee, 2016) and so on.

R is a completely free software environment for statistical analyses and graphics. It is mostly used in universities for academic research and in other areas, such as econometrics, bioinformatics, medicine, biology, business, banking etc. The R programming language provides functions that support linear and nonlinear modelling, classical tests used in statistics, classifications, time-series analysis, clustering, forecasting and makes data visualization easy through a wide variety of graphs (R Core Team, 2017a).

The functionality of R can be improved by various extensions called packages. These extensions can be created by users according to their needs (Heiss, 2016). R includes base packages which are activated by default, such as “base”, “compiler”, “graphics”, “stats” and more, and numerous packages that can be downloaded from CRAN (“Comprehensive R Archive Network”) and then installed and activated. Some of these packages are: “Matrix”, “class”, “foreign”, “survival”, “ggplot2”, “spatial” etc. (Dușa et al., 2015). Moreover, R allows users to develop new packages with functions used for solving various problems.

In order to perform statistical analysis in R, users must have data access. One of the biggest and most accessed database in Romania is TEMPO-Online from National Institute of Statistics available at the following address: http://statistici.insse.ro/shop/. TEMPO-Online is a statistical database provided by the National Institute of Statistics that gives access to a wide range of information. The content of this database consists of statistical indicators, metadata of statistical indicators (definition, statistical methodology) and time series from 1990 to the present (with monthly, quarterly and annual periodicity). TEMPO-Online offers free access to tables with detailed statistical information that can be export in format .csv or .xls. The access is based on site registration with a valid email address and a password. Once registered, users can download tables (the first step) and import them in a statistical analysis software for future examination (the second step). This is a time consuming activity that can be quickly done in one step, only by importing tables from TEMPO-Online into R. Therefore, downloading data on users’computers is no longer required.

2. METHODOLOGY

The underlying idea behind rTempo is that by using the rich features of extensibility in the R programming environment, e.g. packages, and web scraping techniques we can provide a versatile tool for accesing and downloading tables from online databases, e.g. TEMPO-Online, very easily.

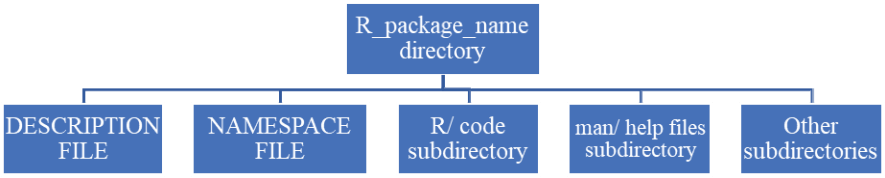
Web scraping is composed from a series of techniques and tools, i.e. programming libraries, for interacting with Web sites in an automated fashion. It is mainly used for parsing and downloading unstructured data from Web. An Web scraping architecture it is composed primary from two a tool for sending and receiving request from Web servers and a web document parser.

R is a modern versatile programming language which provides powerful tools for adressing extensibility, the property of a language to add some new functionality or higher abstraction layers, by hidding the complexity of the implementation from the common user. Writing R extensions guide (R Core Team, 2017b) provides a detailed description which covers all the aspects of deploying packages in a stable, professional manner. Also a very friendly, non-programmers orientated tutorial for writing R packages, is provided by Leisch (2009) in which all the importants aspects, ranging from computer programming paradigms available in R, like the S3 and S4 systems, the legacy S formulas, the file structure of the package and the help documentation for the package, are provided by a hands-on approach by implementating a linear regression function as an R package.

A R package "is a directory of files which extend R" (R Core Team, 2017), providing the users, mainly data scientist, with fast and reliable tools beyond the ones present in R distribution, called R Base Package. The structure of an R package is presented in figure 1.

The file structure of an R package.

Figure 1



Source: adapted by the authors after R Core Team, 2017

R packages are built to run on different computational architectures, i.e. computing machines or operating systems, in a consistent and predictable manner, avoiding a lot of problems generated by filesystems, memory

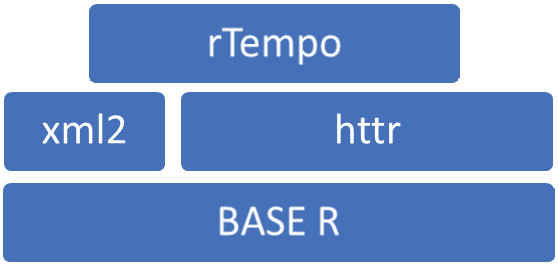
management and inconsistent error handling, to name only a very few sources. Dispatching packages is a very easy job, done through online repositories, e.g. CRAN or Github. But the main strength of R comes from the data-centric community that is expanding at an increasing rate. Researchers from very different fields of science, from particle physics to humanities, seek to exploit the current boom in data and share their experience and enthusiasm, like never before seen, through R packages. Much of the inspiration for this package comes from them.

rTempo package is built, mainly, upon two packages, used to interact with Web servers, developed by the prolific R author Hadley Wickham, rvest (2016) and httr (2017). Rvest is a package consisting of wrapper functions for httr and xml2 (Wickham, 2017) built as a lightweight implementation of a Web scrapping engine. Contains functions for downloading and parsing HTML or XML documents, some of which have capabilities for addressing scripts present in the Document Object Model. Objects like javascript forms are automatically detected by the rvest parser and are transformed in R objects, i.e. lists. The user can then modify the values present in the HTML document and send valid, simple GET or POST requests to the Web server. Xml2 package is a package designed with one idea in mind, and that is transforming HTML or XML documents in R objects. Httr package is also a wrapper for the curl (Stenberg, 1997) library, as a versatile R interface for interacting with HTTP. Built aroung HTTP verbs, like POST or GET, httr lets the user to create requests to a web server by providing functions with arguments like headers, cache control, cookies for simulating a valid HTTP session in the R environment. The server reponses are stored as lists, which in turn can be modified as the user pleases.

A simple description for the interaction with the aforementioned packages consists with the user providing a valid URL for the function which wraps a GET request, receiving a GET responses from the server, parsing the reponses as a list, modify the list and resend another GET or POST response for asking the server for something else, like another URL or document. As a first conclusion, its easy to manipulate HTTP objects in R, otherwise a very complex task. The figure 2. presents the hierarchical structure of the dependencies for rTempo package.

The hierarchical dependencies for rTempo.

Figure 2



Source: designed by the authors.

3. RTEMPO IMPLEMENTATION

rTempo consists of 11 functions which target the integration of the TEMPO-Online interface, through the National Institute of Statistics site in the R environment, adding extra functionality like downloading a bulk of tables who share a common statistical domain or by providing simple keywords for querying the underlying database. It also provides a function for creating a working directory to store all the tables in .csv format and the definitions, along with some metadata, for the statistical indices provided as part of the HTML document. The definitions can be accesed in the R Studio environment in the Viewer pane by simply typing the name of the R object who stores the table as an argument to the respective function. Consistency across the HTTP requests is maintained in by using a login function to be able to acess the downloading functionalities available on the site. Descriptions for the functions are provided below.

1. login_tempo(usr, passwd)

login_tempo() has two arguments provided by the user when is called. The first argument is a valid user name which was supplied by the user when he first created a valid TEMPO-Online account. The second arguments required is a valid password for the account. A valid call will return the following response:

```
> login_tempo(usr="marian.necula@insse.ro", passwd = "**********")
Please select Tempo Online language:
1. Romanian 2. English
1
Submitting with 'NULL'
<session> http://statistici.insse.ro/shop/?lang=ro
 Status: 200
Type:   text/html;charset=UTF-8
Size:   18711
```

The user is required to provide an option for the language, by selecting from the Romanian or the English language provided on the NIS site. If no username or password is provided the function returns an error message and asks the user to provide a valid username and password.

2. nav_tempo()

nav_tempo() gives the possibility to navigate across the TEMPO-Online tree by returning the a list of the statistical domains and statistical tables available for download. A call will look like this:

```
> nav_tempo()
Do you wish to dowload
 1. A single table?
 2. Bulk download a statistical sub-domanin of tables?
 3. Logout?
```

Here the user is prompted to supply an option from the three available, downloading a single table, a bulk of tables from a statistical domain or logout from the current session. Pressing 1 returns the following response:

```
Please type the number of the statistical domain:
4
Navigating to index.jsp?page=tempo2&lang=ro&context=15
[1] "POPULATION AND ITS DEMOGRAPHIC STRUCTURE"
 [2] "VITAL STATISTICS"
 [3] "INTERNAL AND INTERNATIONAL MIGRATION"
 [4] "LABOUR FORCE"
...
[31] "ENVIRONMENT"
[32] "TERRITORY ADMINISTRATION"
[33] "PUBLIC UTILITIES OF LOCAL INTEREST"
Please type the number of the table:
5
```

After the statistical domain is selected, an internal package function for creating and sending a POST request to the Web server is called. The response is then parsed as a list of tables from which we may select the target table for downloading:

```
[1] "FOM101A - Labour resources by gender, macroregions, development region and
counties"
 [2] "FOM116A - Employment rate of labour resources by gender, macroregions, development
regions and counties"
 [3] "FOM102A - Civil economically active population by sex, macroregions, development
regions and counties"
```

[4] "FOM114A - Gross activity rate"
...
[225] "FPC107B - Enterprises that offered initial vocational training, by size classes and categories of units"
[226] "FPC107C - Enterprises that offered initial vocational training, by ownership type and categories of units"

The internal function is called, which in turn sends another POST request, parses the response and saves the table as a .csv file, a R object, i.e. data frame and aditionally downloads the definitions and metadata available on the NIS site as HTML, so the user can start to use as fast as possible the new data.

3. search_tempo(search_word = c(…))

search_tempo() provides the interface for searching the tables by requesting the user a keyword to do an approximate matching, based on the javascript code implemented by the website, with the character strings from the tables names. If more than one keyword is provided the function enters bulk download mode, sending iterative requests to the internal function responsible for creating and sending request to start downloading all the available table which match the keyword. Some inherent drawbacks from approximate matching can lead to unwanted results, as the keyword provided is matched no matter where is found inside the table's name. A call to this function return results similar to nav_tempo() function, only this time the list of table names are drawn from the entire database and depend on the language selected at login. Also an efficient use of this function will require from the user to have some knowledge about the table names in TEMPO-Online.

> search_tempo(search_word = "inflatie")
Navigating to ./?page=search&lang=ro
[1] "IPC101A - Rata medie lunara a inflatiei pe categorii de marfuri si servicii cumparate

4. def_tempo("*.html")

def_tempo() provides for the R Studio users a fast way to see the definitions and metadata associated with the tables inside R Studio's Viewer pane as html documents, without the need to consult them in a traditional browser on the NIS site.

> def_tempo("IPC101A.html")

5. dir_tempo()

dir_tempo() can be used to before the download start to create a dedicated working directory called "~/R_current_working_directory/INS", to save all the files, either .csv or the .html definition and metadata files.

> dir_tempo()

6. logout_tempo()

logout_tempo() provides the means to terminate the current http session from R and reset the working directory. The server has under the belt an Oracle stateless web architecture which means only the current variables in the working environment need to removed for the session to be closed.

> dir_tempo()

Additional functions are present in the rTempo package such as a decision function which determines if the parsing and the download should be done for a single or multiple tables based on the user's input and the real workhorse of this package, the function which creates and sends the POST request to the server, reads the response and saves it's content as .csv files and R data frame objects in the current working directory and environment, respectively. Additional functionality may be provided as the package tries to evolve to be as intuitive and user-friendly as possible.

rTempo is currently in a very early, but functional, stage of development. The main functions are there, but some refinement is much needed. Some come from development decision taken on the fly and in the near future we expect to re-write part of the code, to improve efficiency and stability, and choose different approaches to tackle this.

There is a list of limitations currently under the scope, from which the biggest one is that we had to choose if an implementation for the tables which are populated dinamically by a script and who requires 43 requests for downloading the data in bulk for a single table would be wise, as the database is quite old by current standards. Another limitation comes directly from the method employed to acces the database. Webscraping techniques are not very reliable in the medium to long term, as sites may change quickly and this will require some work from the maintainer in an unpredictable manner. A solution to this drawback may come in from addressing directly the TEMPO-Online database through an API, but this depends on factors outside our reach.

Another important limitation comes from the lack of thorough testing the package on multiple different systems.

CONCLUSIONS

Developing an R package for the TEMPO-Online is seen as a small part in an effort made by the National Institute of Statistics for the modernisation of the National Statistical System, with the user of reliable and timely statistical data in mind. This can be accomplished by adopting the new technologies as a mean for developing a robust and efficient way to tackle the big and small problems for our century.

R programming environment provides modern tools for data scientists to implement their ideas in a straigthforward and easy way as R package, by adding higher abstraction layers and extending R functions present in other packages.

rTempo provides an easy and fast method for integrating the data present in TEMPO-Online directly in the R environment without the need of using a Web browser. Currently it has some limitations which will be addressed in the near future by relying on the users feedback and much more testing.

REFERENCES

1. **Biecek, P.**, 2016. *SmarterPoland: Tools for Accessing Various Datasets Developed by the Foundation SmarterPoland.pl.,* [Online] Available at: https://cran.r-project.org/web/packages/SmarterPoland/index.html [Accessed 20 September 2017].
2. **Dușa, A., Oancea, B.**, **Caragea N., Alexandru, C., Jula, N.M., Dobre, A.M.,** 2015. *R cu aplicații în statistică.* Bucharest: The Bucharest University Press.
3. **Eurostat**, 2011. *European Statistics Code of Practice,* Luxembourg: European Statistical System.
4. **Flowers, J., Fox, S., Thelwall, S., Flint, D., Hain, D.,** 2017. *fingertipsR: an R package for accessing population health information in England,* bioRxiv, doi: https://doi.org/10.1101/189167
5. **Heiss, F.**, 2016. *Using R for Introductory Econometrics.* Dusseldorf: Germany.
6. **Jianu, I., Jianu, I., Ileanu, B. V., Nedelcu, M.V., Herteliu, C.**. *The Value Relevance Of Financial Reporting In Romania.* Economic Computation and Economic Cybernetics Studies and Research, Vol. 48(4), 167-182.
7. **Leisch, F**. 2009. *Creating R Packages: A Tutorial,* in Brito, P. (Ed.), Compstat 2008 – Proceedings in Computational Statistics. Physica Verlag, Heidelberg.
8. **Lee, M.J.**, 2016. *IMFData: R Interface for International Monetary Fund(IMF) Data API*, R package version 0.2.0. [Online] Available at: https://CRAN.R-project.org/package=IMFData [Accessed 20 September 2017].
9. **RCoreTeam**,2017a.*R:ALanguageandEnvironmentforStatistical Computing.*[Online] Available at: https://www.r-project.org/ [Accessed 20 September 2017].
10. **R Core Team,** 2017b. *Writing R extensions.* Pages: 3-17.[Online] Available at: https://cran.r-project.org/doc/manuals/R-exts.pdf [Accessed 20 September 2017].
11. **Stenberg, D.** 2017**,** *curl*, Version 7.55.1 [Online] Available at: https://curl.haxx.se/changes.html [Accessed 20 September 2017].
12. **United Nations Statistical Division**, 2016. *Handbook on Rapid Estimates.* Pages: 16-20. [Online] Available at: https://unstats.un.org/unsd/nationalaccount/consultationDocs/Handbook_RE.pdf [Accessed 20 September 2017].

13. **Vitolo, C.**, 2017. *hddtools: Hydrological Data Discovery Tools.* R package version 0.3.0, [Online] Available at: https://CRAN.R-project.org/package=hddtools, doi: 10.5281/zenodo.61570 [Accessed 20 September 2017].
14. **Wickham, H**., 2017. *httr: Tools for Working with URLs and HTTP*. R package version 1.3.1, [Online] Available at: https://cran.r-project.org/web/packages/httr/index.html [Accessed 20 September 2017]
15. **Wickham, H.**, 2016*, rvest: Easily Harvest (Scrape) Web Pages*. R package version 0.3.2. [Online] Available at: https://cran.r-project.org/web/packages/rvest/index.html [Accessed 20 September 2017]
16. **Wickham, H.,** 2017, *xml2: Parse XML,* R package version 1.1.1. [Online] Available at: https://cran.r-project.org/web/packages/xml2/index.html [Accessed 20 September 2017]