Open Computational Ecosystems and Reproducible Research

Markus Kainu, Joona Lehtomäki, Juuso Parkkinen, Juha Yrjölä, Måns Magnusson, Mikko Tolonen, Niko Ilomäki, Leo Lahti

Contact: http://ropengov.github.io

Open data analytics The recent explosion in open data availability has created novel opportunities for research. Efficient data analytical tools are crucial for taking full advantage of digital data streams. Custom software libraries are now rapidly emerging and have a huge potential to contribute to transforming computational social sciences, digital humanities, and related fields.

Advantages of the open development model Efficient data analysis relies on customized workflows that are best developed jointly by the user community, as already is the standard practice in bioinformatics (Bioconductor), particle physics, and other fields (rOpenSci). Similar communities are now shaping up in social sciences and humanities. The open analytics has many benefits (Ioannidis 2014, Morin et al. 2012):

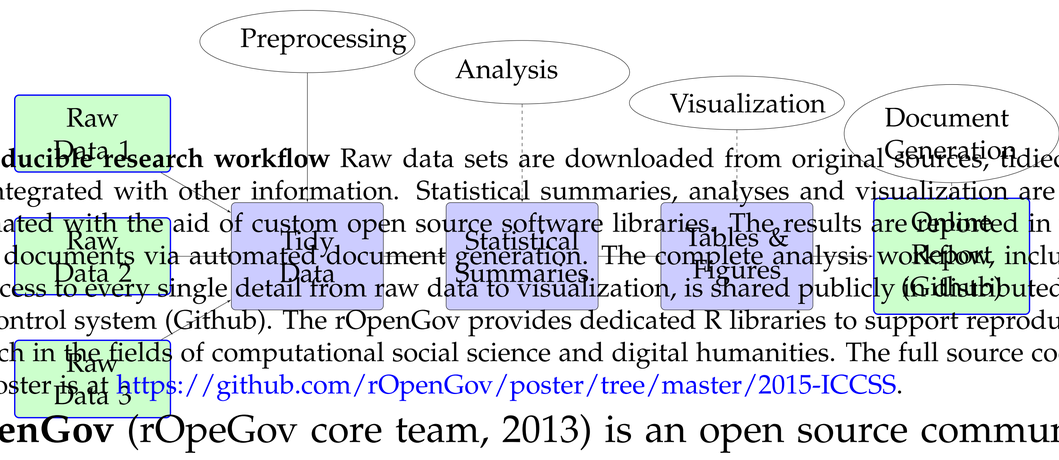
```
## Warning in download.file(url, tfile): unable to resolve 'ec.europa.eu'
## Error in download.file(url, tfile): cannot open URL 'http://ec.europa.eu/eurostat/estat-navtree-portlet-prod/BulkDownloadListing?
```

searchers more time to focus on the specific problems.

```
## Error in df$time: object of type 'closure' is not subsettable
## Error in df$time: object of type 'closure' is not subsettable
## Error in as.data.frame.default(y): cannot coerce class ""function"" to a data.frame
## Error in eval(expr, envir, enclos): object 'dat2' not found
## Error in eval(expr, envir, enclos): object 'dat2' not found
## Error in dat2$NUTS_ID <- NULL: object 'dat2' not found
## Error in spCbind(map_nuts2, dat2): error in evaluating the argument 'x' in selecting a
method for function 'spCbind': Error: object 'dat2' not found
## Error in rownames(shape@data): object 'shape' not found
## Error in fortify(shape, region = "id"): object 'shape' not found
## Error in merge(map.points, shape, by = "id"): error in evaluating the argument 'x' in
selecting a method for function 'merge': Error: object 'map.p
## Error in eval(expr, envir, enclos): object 'map.df' not fou
## Error in ggplot(data = map.df, aes(long, lat, group = group)
found
## Error in eval(expr, envir, enclos): object 'p' not found
## Error in eval(expr, envir, enclos): object 'p' not found
## Error in eval(expr, envir, enclos): object 'p' not found
## Error in eval(expr, envir, enclos): object 'p' not found
## Error in eval(expr, envir, enclos): object 'p' not found
```

```
## Error in print(p): object 'p' not found
```

This poster, including the Eurostat analysis example, is fully reproducible. Download the full source code at https://github.com/rOpenGov/poster/blob/master/eurostat_map.R



Reproducible research workflow Raw data sets are downloaded from original sources, tidied up, and integrated with other information. Statistical summaries, analyses and visualization are then automated with the aid of custom open source software libraries. The results are reported in web-based documents via automated document generation. The complete analysis workflow, including full access to every single detail from raw data to visualization, is shared publicly (Github) using version control system (Github). The rOpenGov provides dedicated R libraries to support reproducible research in the fields of computational social science and digital humanities. The full source code of this poster is at https://github.com/rOpenGov/poster/tree/master/2015-ICCSS.

rOpenGov (rOpeGov core team, 2013) is an open source community and a statistical ecosystem based on the **R statistical programming language** which has rich data analytical capabilities. We develop data analysis methods for computational social science (Lazer et al. 2009) and digital humanities. Main components include:

- **Reproducible research blog** (http://ropengov.github.io) highlights the opportunities of open data analytics.
- **Online tutorials** demonstrate how to access and analyse open data streams.
- **R packages** provide the means to share computational algorithms to support reproducible data analysis. Our collection includes tools for open data in various countries (Finland, Poland, Russia, USA), cities (Helsinki), statistics authorities (Eurostat, PX-Web, QOG), data anonymization, geographic information (OpenStreetMap, WFS), weather, demography, bibliograpies, media APIs, political science, elections and parliamentary monitoring. For a full list, see http://ropengov.github.io/projects

```
# Download Eurostat data using
library(eurostat)
df <- get_eurostat("tgs00026", time_format = "raw")
# Download geospatial data
download.file("http://ec.europa.eu/eurostat/cache/GISCO/geodatafiles
              destfile="NUTS_2010_60M_SH.zip")
# Manipulate data using
library(tidyr)
library(maptools)
library(rgdal)
library(sp)
library(rgeos)
# Plot data using
library(ggplot2)
library(scales)
library(grid)
```

References

1. J. Ioannidis (2014). How to Make More Published Research True? PLoS Medicine 11(10): e1001747.
2. D. Lazer et al. (2009). Computational Social Science 323, 721–723
3. A. Morin et al. (2012). Research priorities. Shining light into black boxes. Science 336, 159-160.
4. rOpenGov core team (2013). R ecosystem for open government data and computational social science. NIPS Machine Learning Open Source Software workshop (MLOSS). December 2013, Lake Tahoe, Nevada, US

Political Science Review, 107(02), 326–343

We are thankful for a number of developers. For a full list, see http://ropengov.github.io