REGION
The Journal of ERSA
Powered by WU

Volume 2, Number 2, 2015, R1-R2
DOI: 10.18335/region.v2i2.85

journal homepage: region.ersa.org

WooW-II: Workshop on open workflows*

Daniel Arribas-Bel[1], Thomas de Graaff[2]

[1] School of Geography, Earth and Environmental Sciences, University of Birmingham (UK) (email: D.Arribas-Bel@bham.ac.uk).
[2] Department of Spatial Economics, VU University, Amsterdam (Netherlands) (email: t.de.graaff@vu.nl).

Received: 29 June 2015/Accepted: 29 June 2015

Abstract. This resource describes WooW-II, a two-day workshop on open workflows for quantitative social scientists. The workshop is broken down in five main parts, where each of them typically consists of an introductory tutorial and a hands-on assignment. The specific tools discussed in this workshop are *Markdown*, *Pandoc*, *Git*, *Github*, *R*, and *Rstudio*, but the theoretical approach applies to a wider range of tools (e.g., *LaTeX* and *Python*). By the end of the workshop, participants should be able to reproduce a paper of their own and make it available in an open form applying the concepts and tools introduced.

1 Background

As in most social sciences, virtually no training is provided in regional science on workflow design and choice of appropriate tools, especially not from the viewpoint of open science (Healy 2011, Arribas-Bel 2014). Students and young researchers typically receive no guidance as to why or how they should adopt habits that favor the open science principles in their research activity. This is unfortunate, because learning and adopting new tools and workflows require a large time investment, which will only pay-off in the long run. The best time to get started is early in the career when one still has (some) time available to invest. Therefore, this workshop is specifically aimed at young researchers and covers the main ideas behind a well-designed workflow with openness, transparency and reproducibility in mind. At the same time, the content provides an introductory, hands-on overview of a set of free tools that have been designed with such values in mind.

We do not get into every detail of each tool. Instead, we aim to give a gentle introduction, to provide further material, and to place these in the appropriate context. Specific emphasis is set on how certain tools contribute to building a coherent open workflow and how they relate to each other. The main areas reviewed are: mark-up languages such as *Markdown*; reference managers – particularly those open and free such as *Bibtex*, which are compatible with LaTeX; conversion tools such as *Pandoc*; open environments for statistical computing such as *R* or *Python*; version control systems such as *Git*; and online hosting on open repositories such as *GitHub*. At the end of the workshop, participants should be able to reproduce a paper of their own and make it available in an open form applying the concepts and tools introduced. Materials are organized on a website that is openly hosted on *GitHub* and licensed using Creative Commons meaning that access, remix and redistribution are permitted.

*The creation of this workshop is generously sponsored by the European Union's Seventh Framework Programme "Foster" (see as well https://www.fosteropenscience.eu/event/workshop-open-workflows).

2 Description of the resource

The structure of the workshop is organized in two main blocks. The first session introduces basic concepts such as open science, transparency and reproducibility. Here, we stress the relevance of paying attention to the way science is carried out and connect it to the choice of tools that allow such values to be seamlessly embraced in the day-to-day practice of quantitative research in social science. The second, longer, part of the workshop includes four sessions with hands-on overviews of specific tools that have been designed with open science principles in mind and that hence provide the ingredients of a well-thought-out open workflow. The delivery alternates presentation time with hands-on practice, allowing participants to get a real taste of what using the tools implies and therefore experience their advantages.

The five sessions are presented as follows:

1. In this three-hour session, we introduce the concepts of workflow, openness and reproducibility. In the first part, we argue why these concepts are important and what as social scientists we can learn from data scientists. Our main argument is that, although reproducibility is often infeasible in the social sciences, we should strive for research to become as reproducible as possible.

2. In this two-hour session, we introduce the concepts of version control and task automation. The first hour relates to keeping track of changes as they occur throughout the process, while the second hour allows us to break up the different components of an analysis and have them automatically run, when needed, in the correct sequence. The two tools we use to explore these ideas practically are *git* and *make*.

3. In this two-hour session, we introduce the concept of markup languages and working with the terminal. In particular, we focus on *Markdown*, a very lightweight markup language (and probably the fastest way to create slides), and *RStudio*. This enables writing part of a paper in *Markdown* using *RStudio* for document elements such as headers, links, formulas, tables, and references. Using *RStudio* also allows for exporting to better-known formats, such as *docx*, *HTML* and *pdf*.

4. In this three-hour session, we provide an overview of the main ideas behind making data analysis reproducible and transparent. We use the *R* statistical platform in combination with *RStudio* for two main reasons: (i) it works the best out of the box for our purposes and (ii) currently most researchers probably work with this combination for reproducibility.

5. In this final 90-minute session, we introduce how one could make their reproducible research open. This essentially means making use of repositories such as *Github*, which not only serves as a backup repository, but as a method of collaboration with known and unknown authors. Further, we show that making slides in *RStudio* is simple and why authors might prefer to publish a document in *HTML* instead on paper.

3 Resource links

- Website: http://darribas.org/WooWii/
- Materials: https://github.com/darribas/WooWii

References

Arribas-Bel D (2014) Open workflow for open regional science. NARSC newsletter 2(1):4

Healy K (2011) Choosing your workflow application. *The Political Methodologist*: 9–18