WooW-II: Workshop on open Workflows

Daniel Arribas-Bel, Thomas de Graaff*

June 24, 2015

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 part typically consists of an introductionary 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 (including LaTeX and `Python`). 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.

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*. 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 do this is early in the career when one still has (some) time available. 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, and provides an introductory, hands-on, overview of a set of free tools that have been designed with such values in mind.[1]

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 them in the appropriate context. Specific emphasis will be put on how certain tools contribute to building a coherent open workflow and how they relate to other tools. The main areas we will review 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 back-up solutions on open repositories such as `GitHhub`. 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.

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 open workflow. This is delivered alternating presentation time with demo time, allowing participants to get a real taste of what using the tools implies and experience their advantages.

The five sessions are presented as follows:

1. In this 3h. session we introduce the concepts of workflow, openness and reproducability. In the first part, We argue why they are important and what as social scientists we can learn from data

*d.arribas-bel@bham.ac.uk,t.de.graaff@vu.nl

[1]The creation workshop is generously sponsored by the European Union's Seventh Framework Programme "Foster".

scientists. Our main argument is that, even though in the social sciences complete reproducibility is often infeasible, we should strive for research to become *as reproducable* as possible.

2. In this 2h. session we introduce the concepts of version control and automation of tasks. The first relates to keeping track of changes as they occur throughout the proccess, while the second one 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 with which we will play to explore these ideas practically are `git` and `make`.

3. In this 2h. 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). In particular we deal with `RStudio` and `Markdown`. This enables writing a (part of a) paper in `Markdown` in `RStudio` including headers, links, formula's, tables and references. Using `RStudio` this allows as well for exporting to more well-known formats , such as `docx`, `html` and `pdf`.

4. In this 3h. session we take 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*) at the moment most researchers probably work with this combination for reproducability (at least it gets the biggest buzz...).

5. In this final 1.5h session we introduce how one could make your reproducable research open. This basically means making use of repositories such as `Github`, which not only serves as a backup repository, but as well as a way of collaboration with known and unknown authors. Further, we show that making slides in `RStudio` is a breeze and why you actually might want to publish a document in `HTML` instead on paper.

3 Resource links

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