Workflow for the analysis of secondary quantitative data sources

Introduction

The goal of social science research is to produce and make publicly available the results to advance knowledge, deepen understanding and, ultimately, improve well-being. That broad goal is advanced when scientists keep up with recent developments of their colleagues by attending conferences and reading the literature, when they make well-reasoned arguments, and when they write and publish research papers to communicate their own empirical findings and interpretations. Good workflow contributes to this goal by making us more efficient, providing us and our colleagues greater confidence in our work, reducing error, and creating stronger foundations for future research. By workflow I mean practices implemented on a day-to-day basis while conducting research that organize and document our analysis of empirical data. Any empirical research, whether quantitative or qualitative, involves workflow, but the focus of this document is on workflow for the analysis of secondary quantitative data sources.

Workflow Goals

Your workflow should...

- Organize and document your research results
- Link results with process that produce them
- Help you to find what you were doing last time you worked on the project.
- Document known errors and inconsistencies in data.
- Allow collaboration. Your workflow needs to accommodate different work styles and computing systems.
- Provide opportunities to find errors.
- Allow you to build on past results in future studies, but also archive materials to replicate past results.

Elements of a good workflow (click on links to see tips on how to implement each element to good effect)

File structure – directory and sub-directory structure, file naming conventions

Documentation – research notebook, project document, code documentation, data documentation (e.g. date and method of access).

Code Structure – organization of scripts to process data that makes clear the function of each and its role in the project.

Archive – records of code and data that produced published results, of supplementary analyses, and of earlier drafts of papers written for publication. For archiving code, you might use GitHub and GitHub Desktop.

Automation – the code produces the empirical results as they appear in the publications and supplements with minimal human intervention.

Collaboration – workflow facilitates work on teams, but teamwork also encourages good workflow, You'll see in the page on File Structure, my projects have a workflow that requires that each collaborator replicate the analysis independently. This allows us to check to make sure that the code is archived and works the same in different environments. Sharing code through a GitHub repository allows each one of us to review and edit the code. Often I find myself adding documentation as I read code written by someone else to make it clearer to me. My collaborators do the same for my code.

Tools

GitHub – GitHub is a location for archiving, sharing, and keeping a version history of your files. There are many introductions to Git and GitHub, but t his one is especially good.

GitHub Desktop – graphical interface for interacting with github repositories.

Zotero – software to manage and format your references. Share libraries with you collaborators.

Cloud Storage – storing your files where they will be backed up and your collaborators can also access them.

Coding Tools

Macros

Assert

Markdown

See also Automation