

Effective Programming Practices for Economists

Reproducible Research

Definition of reproducibility

Janoš Gabler and Hans-Martin von Gaudecker

Availability

- Source data:
 - Always start from the data in the way you obtained it
 - Add a detailed description how you got it
 - If possible, include all datasets in a common format
- Source code:
 - Include any code that is needed to produce your results
- Programmes:
 - Document all programmes that need to be installed to run your code
 - Automate the installation as much as possible with environments
 - When the project reaches a milestone (submission?), pin the versions
- Essentially a version of <https://datacodestandard.org/>

Version control

- Raw data and source code are under version control
- Published results are created from the main branch with no uncommitted changes
- Use tags / releases to mark submissions, revisions, etc.

Separation of source files and output

- All generated files are in a separate folder that can be safely deleted
- Generated files are not under version control!
 - Can easily become outdated
 - GitHub repository size would explode
 - Does not help with reproducibility
- Intermediate data could be added upon final publication

Theory / Automation

- The workflow can be described by a directed acyclic graph
 - Files are nodes
 - Tasks operating on these files are nodes
 - Edges describe the dependencies
- There is one command that converts your source data into the paper with figures and tables

Documentation and readability

- You strive for readability in your source code
- There is a README file that documents
 - Your directory structure
 - How to install packages
 - How to run your code
- Docstrings and comments explain the code where necessary