Best Practices for Reproducible Research

Author: Meer Muhammad Khan

Reproducible Research

The reproducible research is gaining importance now a day and there are two basic reasons for making our research reproducible. The first is to show evidence of the correctness of our results. Descriptions contained in scholarly publications are rarely sufficient to convince readers of the reliability of our work and the second reason for reproducibility is to enable others to make use of our methods and results. Equipped with only our published article, anyone from our field/ research area might struggle to reconstruct our method in enough detail to apply it to their own data.

Key Elements for Reproducible Research

There are some key elements in reproducible research which should be considered and followed by researcher during their research. By adopting these key elements a researcher can end up with a fully reproducible research. These key elements are as follows:

- Workflow
- Literate Computing and Documentation
- Version Control
- Code
- Distribution

Best Practices

The main purpose of this document is to give some good guidelines about a reproducible research. We will try to give instruction/ tips that what should be consider (DO's) and what should be avoid (DON'Ts) during your reproducible research in further description of this documents. This document is consisted on previous experiences during the reproducible research. We will try to describe each and every things which should be consider and which should be avoid in light of above mentioned key elements.

1: Work Flow

The workflow is a major element of a reproducible research and we can say it is overall sketch of your reproducible research. The main benefit of a workflow is that it gives a main idea about your work at first glance. Always try to start your research with a workflow and by doing this you will be clear about your reproducible. According to us the workflow should be first step of a reproducible research. Some researcher avoids this step in their reproducible research which is not a good practice. The basic reproducible research workflow can be divided into three main stages: data acquisition, data processing, and data analysis. There are different tools for designing a work flow like Draw.io or Microsoft Visio.

DO's: You should consider these things during Workflow:

- Always try to start your reproducible research with a workflow.
- Use an appropriate tool for designing a workflow.
- According to our experience Draw.IO is best tool for designing a workflow
- Try to export your workflow in JPEG or PDF format because the quality is always good in these formats.
- Always try to save your file in editable format for future use.

DON'Ts: You should avoid these things during Workflow:

- Don't think to skip this step of designing workflow during reproducible research.
- Don't use only white backgrounds in each block of workflow make it colorful.
- Don't use so much text in your workflow.

2: Literate Computing and Documentation

The documentation is also important element of reproducible research. We always have documentation (research article/ paper) at the front of our reproducible research. The researcher first start from our documentation and then they further move to other steps for reproducing our research. There are different tools for documentation like Jupyter Notebook, Ipython Notebook, overleaf or Latex. You should be precise and clear cut which tool is best for your documentation. The Jupyter Notebook is a good tool for documentation. It has good features like you write a document in Markdown format, you can run code and draw your graph by using your datasets.

DO's: You should consider these things during documentation:

- Try to use Jupyter Notebook with Linux operating system because it is easy and light weight.
- Divide your text in different cells instead of in single cell in Jupyter notebook.
- At beginning search the way for exporting your documentation from Jupyter Notebook in IEEE or other research paper format.
- Also try to learn about overleaf or Latex for documentation.
- Make your research paper executable with help of Jupyter Notebook.
- Export your document in Latex format if you don't have option to make in directly IEEE format from Jupyter Notebook.

DON'Ts: You should avoid these things during Documentation:

- Don't use Jupyter Notebook with windows using anaconda because it sometimes create problem.
- Don't put your reference manually. Use Bib Tex in overleaf or Latex
- Don't show your code in Jupyter Notebook. Make a csv file and link it with your documentation in Jupyter Notebook.

3: Version Control

Reproducible research is a gradual process. From a starting point to an end there are several processes which involve in reproducible research. There are different version of your code, data sets in your reproducible research because reproducible is a constructive process. There is a need to manage different version of your source code, code or data sets. Different tools are available for version control like Github, Data version control (DVC), Figshare, Kaggle, Zenodo and Git large file storage. You can use any afore mentioned tool for version control.

DO's: You should consider these things during version control:

- Github is a best tool for managing your code, data set etc. you should try to use Github for managing and for version control.
- If you are working on some machine learning projects then go for Data version control (DVC) because Data version control is a best tool for machine learning projects.
- If you are dealing with large data set you can use Zenodo for version control because it offers upto 50 Gb space.

- If you want to store large files on Github then us Git large file storage. It is best option to use large files on your Github.
- Try to make Github repository at first day of your reproducible research.

DON'Ts: You should avoid these things during version control:

- Don't think to make Github repository on last day of your reproducible research and put all of your data at Github at last day.
- If you are dealing with small data set then don't think for other tools just use Github.
- Don't save your data with same version number. Always give new version number to your data.
- Don't skip this version control step in your reproducible research because it is good step to keep track your reproducible research.

4: Code

There is a main role of code in our computational reproducible research. Mostly we have to write a code to implement our idea. There are different tools and platform for coding like Jupyter Notebook. The Jupyte notebook is a good tool for an executable research paper. There are option for coding like you can perform coding in Python language. Always try a good tool for coding which shouldn't make trouble for other people during reproducibility of your work for dependencies or other issues.

DO's: You should consider these things during Code:

- Always try to give detailed information about code, tools, infrastructure and software you used for coding.
- Try to use VM if possible for your coding tools you can make a image of same environment which you are using.
- Share your source code without bugs and when you are confident that it is working.
- Write code that is a bit more general than your specific data.

DON'Ts: You should avoid these things during code:

- Avoid having repeated blocks of code.
- Don't forget to give the detail about environment you used for coding.

- Don’t forget to update your code on Github repository if you are using it for reproducibility.

5: Distribution

The distribution is main part of our reproducible research. We give the chance to other researcher to reproduce our work with the help of distribution. A substantial challenge in reproducing analyses is installing and configuring the web of dependencies of specific versions of various analytical tools. Virtual machines and Docker Images enable you to efficiently share your entire computational environment with all the dependencies intact. We can use VM images or Docker images for distribution of our environment or infrastructure. Always be clear about which is best way of distribution.

DO’s: You should consider these things during distribution:

- Always try to clone your virtual machine which you are using your research.
- Either you can use Docker for distribution.
- Share a link of your VM or Docker in your github repository.
- Also mentioned system requirements for your VM or Docker in github.
- If there are steps to follow for using VM or Docker then clearly mention them in your github.

DON’Ts: You should avoid these things during distribution:

- Don’t give manual instructions for installments of different tools for reproducibility of your research.
- Try to make a small size of VM if you are planning to give your own VM for reproducibility otherwise it makes issue for downloading.
- Don’t miss anything in your VM or Docker like dependencies.
- Make sure your VM or Docker is fully prepared for work and don’t share without surety.

Conclusion and Recommendations

In this documents we tried to give you detail information about what should be consider and what should be avoid during reproducible research. We hope you will get useful information which helps you in your reproducible research. One more things if you are using any repository then don’t forget to add a readme file which has all the detail information about the repository like what is purpose of each folder and which folder contains the following files etc. This readme

file should be clear and should contain sufficient information for users. Because sometimes people lost in your folders of your repository and they don’t know how to start reproducible research.