Guideline for Improving Traceability of Research Products

optional subtitle below

Cornell Urban Environmental Fluid Mechanics Lab

April 2022

Version 0

1 Motivation

'Traceability is the procedure of tracking (and documenting) all your raw materials, parts, and finished goods throughout your manufacturing process.'

The motivation is to provide some guidelines and suggest a common protocol that lab members adopt in daily research practice, such that the following three objectives can be achieved: 1) mistakes are more easily tracked, 2) standards are more easily applied across projects, and 3) dealing with multiple rounds of reviews becomes more efficient.

In addition, beyond the aforementioned practical points, keeping your research products traceable is a basic requirement for **open research**.

Two things to keep in mind for the guidelines and protocol:

1. The backbone of a traceable research practice as outlined in this document needs to be naturally integrated into your research activities.
2. The implementation details of the protocol are by no means static: they are meant to be adapted to best aid the above-mentioned three objectives. The high-level goal is to make the whole research trackable and reproducible to yourself, your supervisor(s), collaborators, and future readers.

2 Protocol for research activities

Since the productivity of research activities is measured by journal publications, we take a reverse order to outline the essential steps in ensuring traceability.

2.1 Figures in a manuscript

The two key questions addressed here:

1. What is a figure of publishable quality?

2. What are the necessary auxiliary information of a figure to ensure that it is reproducible?

2.1.1 Figure of publishable quality

1. **Axes labels:** Correct texts, symbols, units must be present for axes labels and/or colorbar, plot title. Legends are appropriately indicated. Font size: 10-12. Consistent font is needed throughout the manuscript.

2. **Figure format:** Figures with no rendering needs to be 600 dpi, preferably vector format (.eps, .pdf). Figures with rendering 300 dpi (if size is too large) of (.pdf, .tiff) format.

3. **Figure size:** Raw figure size needs to be appropriate for an A4-sized page (8-1/4 $\times$ 11-3/4 in). This means that with approximately 1 inch margins from both sides, the figure width should be smaller than 6-1/2 in. **Please learn the proper way of exporting figures with the required dpi, size and format in the context of your programming language.**

Note:

1. If your manuscript draft does not meet the above-mentioned requirements, I will not read your manuscript.

2. In your weekly meeting and/or group meeting, a similar standard applies to figures shown on your powerpoint slides, especially regarding Axes labels.

2.1.2 Necessary auxiliary information of a figure

A figure needs to accompanied by auxiliary information to ensure that it can be reproduced during and after the duration of your project. There are multiple ways to ensure this and you should develop an approach that works for you. For example, I use an excel sheet (e.g.

record.xsl) to record the 1) directory/ies and file names of data; 2) directory/ies and file names of the scripts used for plotting; 3) directory and file name of the figures, for figures in any version of the manuscript. A new sheet is added to record.xsl to record any changes made to produce the figures for manuscript revisions. I will use record.xsl in the following points, but you may choose the most sensible way for you to record the necessary auxiliary information, such as using Markdown in GitHub, a simple text file, and even handwritten notes (it needs to be digitized eventually, of course).

1. **Directory of data** If figures are generated based on post-processed or raw data, indicate the directory of your data and the file names of your data in record.xsl. If data is later backed up on a permanent storage volume, remember to update record.xsl. A separate backup log is a good idea to keep. This will be discussed more in . 2.4.

2. **Directory of script** Indicate directory of your scripts and the name of the script (or line numbers/sections, if multiple figures are generated from one script). Version controls are important. This will be discussed more in Sec.2.2.

3. **Directory of figure** Clearly itemize each figure in a given version of your manuscript. For each figure, the file names and the directory of that figure needs to be clearly indicated.

4. **Any other supplementary information** There are times when figures generated are not included in the manuscript, such as figures for replying reviewers and figures for sensitivity analysis/sanity check/etc., the above three points (directories of figure, data and script) should be all checked.

2.2 Scripts for post-processing

The two aspects addressed here are:

1. Readability of scripts

2. Version control of scripts

2.2.1 Readability of scripts

Good practice of coding applies for writing scripts to analyze data. Commenting your code, such that you can go back to it in at least two or three months for paper revision is essential.

2.2.2 Version control of scripts

Some efforts in version control are worth investing time. I strongly suggest you to use GitHub for all your scripts to produce figures for your manuscripts. **No matter what version control practice you adopt, you must have a final GitHub repo after your paper is submitted, where all your scripts and auxiliary information (e.g., record. xsl) are uploaded for future reference.**

2.3 Theoretical derivation

If your research involves significant amounts of derivations, especially involving a lot of symbolic maths, I suggest you to learn to use Mathematica. Wolfram Engine, which is free, paired with VS Code's extension will give you a Mathematica-like notebook environment.

If your derivation can be mostly derived by hand, you need to have a hand-written version (paper or electronic) and then type-set all your derivations using LaTeX. Learning to use Overleaf to document any of your derivations and thoughts is useful for eventually writing your manuscript and thesis.

Note: I will not read a manuscript prepared in Word if it involves more than just governing equations.

2.4 Data

2.4.1 Data generation

If data are generated from your numerical experiments, you will need to have necessary auxiliary information of data (it can be a separate file or within the record.xsl), which records the following (if relevant):

1. **Data file names and information:** Contents of the data (e.g., experimental run number); key variables and their information (variable physical meanings, dimensions and spatial-temporal resolutions).
2. **Code used to generate the data:** For each data file, indicate the version of the code and directory of the code.
3. **Code used to post-process the raw data:** Very often the raw data need to be post-processed before scripts can be run for plotting figure. Therefore, any post-processing code's file names and directory needs to be indicated in record.xsl.
4. **Other files related to performing numerical experiments:** This includes submission scripts in the server and any other relevant batch scripts.

2.4.2 Data backup

No matter how data are obtained, data need to be backed up in a permanent volume after the project. It can either be a hard drive or the local server of our lab in BioHPC. Please learn to use Globus (https://www.globus.org/) file transfer for data backup.

In the duration of the ongoing projects, backing up essential files is also a good practice. e.g. For LES runs, backing up the fields for restarting purpose is recommended. A 'data backup log' needs to be kept, which should contain information about what, when and where the data have been backed up.

A data backup log needs to be kept and be included as part of the final GitHub repo.