

Data Archiving

Data Preservation Practices

Motivation

Scientific research must be reproducible. If it is not reproducible, it is not science. When you publish a paper, you must include enough information so that others could, in principle, reproduce the results of your work. You also want to be sure that in several months or a few years, that you can re-produce your own results. You would be surprised at how many times we need to go back and re-do some calculations we did some time ago – usually to double check results or to look in more detail at some aspects of the problem that we did not address initially. Another reason is to help new students get started and learn how to do research. If they can reproduce the results of a previous student's work, they are probably ready to tackle a problem of their own. Finally, scientific misconduct does occur. If you are ever accused of research misconduct, you want to be able to go back and show the original data and programs that produced the results. Maybe you will discover that you made a mistake. It's OK to make a mistake, that is not scientific fraud, and your data will show that you did not make things up.

So there are many reasons to be careful about preserving data. This document explains how members of the Lundstrom group should preserve research results and data.

Procedure

Our process is a paper-driven one. For each paper that is submitted, you should create a file of the relevant data. This file should be submitted to me at the time that you submit the paper for publication. If the paper undergoes review and revision, the data file will be updated. The lead author has the responsibility for preserving the data that went into the paper. The name of the folder containing the data should be:

Last_Name_Title_of_Paper_Date

Contents of the Data File

- README (a file containing a brief (one-line) description of each of the other files in the folder)
- The last two lines of the README file should be:
 - 1) Name and date of the person who created the file and collected the data
 - 2) Name and date of the person who checked the Data File to be sure that another person, not familiar with the work, could understand the contents of the file.
- The original draft manuscript exactly as it was submitted for publication
- A copy of the cover letter used for paper submission

- Supplemental Information that was too detailed to include in the paper but which makes the paper easier to understand and/or describes the methods in more detail (some journals like Science, Nature, and the APS journals make a practice of supplying supplemental information. We should do so too.)
- The source code used to generate the results (or a pointer to the specific version of the code on SVN).
- Any MATLAB scripts used to generate plots along with the raw data for the plots
- JPEG files for each figure
- Data files for each figure
- Any other information or data that you believe is important
- After the paper is reviewed, revised, and re-submitted, additional data should be included
- The detailed point-by-point response to the reviewers
- New source code, MATLAB scripts, JPEG figure files, data, etc. that reflect revisions to the manuscript
- The revised manuscript exactly as it was re-submitted to the editor
- A copy of the cover letter used when re-submitting the paper
- A copy of the page proofs and the corrections that we made to the proofs
- After the paper is published, an electronic copy of the final paper should be added