Data Management Plan

Importance of proper data practices

Modern science revolves around computer code: it is used to run simulations, analyze data, and create publishable figures and papers. I will follow best coding practices and ensure that my code is open-source, published, and freely accessible. Not only will this improve the reproduceability of my results, but openly releasing code has been found to increase citation counts in research (Piwowar & Vision 2013, PeerJ, 10.7717/peerj.175). Thus, the data management practices that I will follow, outlined below, will serve to simultaneously increase the trustworthiness, visability, and impact of the work I carry out during my fellowship studies.

Description & handling of expected data products

The research I have proposed is computational, and the products of these studes will be Python code, raw data, and processed data; the handling of these products is laid out below.

Python code Each proposed task will generate Python code that accomplishes three tasks: running computer simulations, post-processing raw data, and generating publication-quality figures. Code used to run simulations and post-process data will be stored in version-controlled Git repositories and made open to the public upon publication of any related work. These Git repositories will be linked to Zenodo repositories which automatically generate referencable DOIs upon the release of official code versions. By releasing official code versions in this way, I will ensure that interested code users can reproduce my published results by using precisely the same code that I used. Any code that I use to generate publication-quality figures will be packaged into separate Zenodo repositories, along with data (see below), and released upon submission of any work. These distribution plans will ensure that even novice users of Python or the Dedalus code will have direct access to my full research pipeline, and will be able to reproduce my simulations and produced figures with ease.

Raw & processed simulation data The simulations that I will run will create raw data in an HDF5 output format, which allows for extreme output versatility during simulation runs and simple location of data products afterwards. Unfortunately, it is not straightforward to publish complete datasets of the full time evolution of these simulations. Individual simulations easily produce over a terabyte of data each, and such large quantities of data are cumbersome to make publically available, or to interact with as a consumer.

Rather than making the full set of simulation data available for all simulation times, I will instead publish select "checkpoint" data at interesting simulation times. These checkpoints will contain the full simulation state and can be partnered with my released code in order to re-run key times in my published simulations.

In addition to these raw checkpoint files, often times published figures feature a large collection of volume-averaged scalar data points, or partially averaged 1D profiles. All such reduced data which appears in published figures, as well as some additional potentially useful measurements, will be released in full alongside figure generation code in Zenodo repositories.