Findings from the DIL Interviews: Data Quality and Documentation

Skills in this competency may include:
● Recognizes, documents and resolves any apparent artifacts, incompletion, or corruption of data.
● Utilizes metadata to facilitate an understanding of potential problems with data sets.
● Documents data sufficiently to enable reproduction of research results and data by others.
● Tracks data provenance and clearly delineates and denotes versions of a data set.

Average Ranking of Importance (5=essential): Faculty = 4.63, Students = 4.12

Faculty responses:
Many faculty feel that their students are trained to check for any discrepancies in their data to resolve issues before analysis, however, faculty don’t have a lot of confidence in the students’ abilities to do the job well, and to document the steps taken.

One of the faculty commented that it is “very hard to motivate students to write documentation,” mostly because the focus is on getting the work done and graduating. Self-documentation (a log of commands used as well as all of the parameters) is also described as being important so that students can reproduce results. Another faculty cited the lack of tools for automating the process as a real challenge. He also noted that students consistently find themselves more concerned with the outputs of an experiment rather than the steps taken to get to the outputs. Still another faculty interviewee is confident that students learn skills on how to write a methods section of a paper, but not enough documentation is done concerning the research process itself. He feels that students are overconfident when it comes to artifacts and corruptions, generally thinking that their data is in good shape. Error-checking procedures are utilized in one of the labs to ensure that the measurements taken fall within known boundaries. The students in this lab are also involved in basic data quality checks, including ensuring that measurements are not out-of-bounds or deemed to be too drastic.

Most faculty report using some kind of version control practices in the lab, whether it is a specific system like SVN or SharePoint, or file-naming practices indicating the version.

Student responses:
Students are aware of and/or participate in quality control steps. Students tended to be less critical of documentation than the faculty we interviewed, consistent with one faculty’s assertion that they are overconfident. Students in the computer engineering program are aware that this is an area that could benefit from “drastic improvement,” (in the words of one student), but they also report that documentation of the steps taken during research is stressed by faculty, so logging of calculations, thoughts, and the entire research process begins early. These students were also more likely to use versioning software; students in ecology and natural resources were more likely to use file naming strategies for versioning. Skills in this competency were generally learned through trial-and-error, from peers, and from the PI.