Shaylyn Adams

Mount Holyoke College
Mentor: Emery Boose and Barbara Lerner

Capturing Data Provenance From R Script Executions

To ensure the reproducibility and validity of scientific work, it is essential to keep a complete record of data analysis. Recording data's history can be tedious, especially without any explicit guidelines. Consequently, the records of data transformation are generally vague with insufficient details. Our research focused on using computer software to capture and display the data analysis history, or data provenance. We worked on increasing the usability of a formal metadata structure that captures data provenance, a Data Derivation Graph (DDG), to make it more accessible to scientists. The DDG describes a post computation data trace but its creation was previously limited to software unsuited for scientific data analysis. Our work this summer involved extending the DDG representations to work with the statistical language, R. A textual syntax describing DDGs was constructed and used with this software. We added more interactive features to the visualization tool allowing scientists to examine the visual DDG and view data values, input and output data files, plots, URLs and the R functions used in the analysis. The resulting DDG is stored in a database and available to be viewed at anytime. We tested the new features on data sets from meteorological and hydrological collections and from other summer student projects at the Harvard Forest. By supporting R, we hope to use these preliminary results to determine what features of the DDGs are most useful for scientists to understand the provenance of their work and how to further extend this technology.