Miruna Oprescu

Harvard College

Mentors: Emery Boose and Barbara Lerner

Visualization Tools for Digital Dataset Derivation Graphs

The ability to provide proper documentation on how research data was collected and processed is essential for ensuring reproducible results in any research project. We use the term data provenance to define the practice of recording all the processes the data has undergone from its collection to its output as a final result. While it is common to record data provenance in the form of narrative description, the increasing complexity of processes applied to large datasets requires software tools for automatically capturing and storing provenance data in a digital format. To make data provenance easier to understand, we developed a way to represent it graphically through powerful interactive visualizations.

We applied these concepts to the needs expressed by the community of scientists at Harvard Forest. The motivation comes from the necessity of these scientists to constantly process 15 min data from a meteorological station and six stream and wetland gauges on the Prospect Hill Tract for the study of the ecology of forest watersheds. Checking for equipment malfunction, sensor drift and modeling or replacing corrupted data are some of the problems scientists at Harvard Forest are faced with on a day-to-day basis. To respond to these issues, we used Little-JIL, a graphical programming language for defining processes developed at the University of Massachusetts, Amherst, and results from previous research to produce a visual implementation of an abstract mathematical object, a Dataset Derivation Graph (DDG), which documents how every piece of data in a dataset was turned into information. Extending Prefuse, a graphical platform written in Java which supports visualizations of data structures such as graphs and trees, we turned the information provided by the DDG into easy-to-read, easy-to-follow visual graphs built and displayed in real time as the process runs on the computer. Furthermore, this interactive visual program collapses or expands parts of the graph, allowing the user to focus on certain parts of the process and manage large DDGs. To address other concerns expressed by the scientific community, future research will seek to implement this tool to make queries on data processes available to scientists in a visual form.