Marios Dardas
College of the Holy Cross
Mentors: Emery Boose and Barbara Learner
Group Project: Data Provenance in R

Searching data provenance created in the R scripting environment

Ensuring that data analysis leads to accurate and reproducible results is critical to scientific progress. As scientists explore increasingly sophisticated issues, they often utilize large data sets and complex models for handling the data. The data provenance, the processes performed to achieve reported results, has become increasingly complex, presenting a challenge to establish the authenticity behind scientific results. Scientists can use RDataTracker, an R library to collect provenance from executing R scripts, to create a Data Derivation Graph (DDG), a structured record of the provenance. DDG Explorer, a program that improves accessibility to DDG, assists with assessing the data provenance. Although the DDG displays this information in an organized fashion, it can be difficult to find specific information within the DDG efficiently – especially within DDG containing hundreds or thousands of nodes. To ensure a quick and intuitive experience with finding data, I focused on creating search capabilities for the DDG. The current search extracts information from the DDG and structures it so that information can be queried by the type of data and associated text information. The interface allows users to enter a search which returns a list of search results. Once the user selects a result the node is displayed. Future updates will incorporate searching within the DDG Explorer's database (implemented with APACHE JENA) by incorporating the SPARQL querying language in addition to an advanced search design to assist with discovering relationships between nodes. Search will assist scientists with debugging programs in addition to quick traversal of the DDG.