Geounit Hub: A Platform for Sharing and Reproducing GeoScience Applications

Tanu Malik§, Asti Bhatt#, Eunseo Choi*, Jonathan Goodall†, Ian Foster§§, Scott Peckham!, David Tarboton§%, Pabitra Dash%, Bakinam Essawy+, Dai Hai Ton That§, Gan Tian%, Todd Valentic#.

§DePaul University, †University of Virginia, %Utah State University, *Stanford Research Institute, #University of Memphis, §§University of Chicago, !University of Colorado, Boulder

"If I have seen further, it is by standing on the shoulders of giants."
Sir Isaac Newton, 167

Scientific questions of today are more global than ever before. Answers to scientific questions are often buried within multiple disciplines and across a diverse range of scientists and institutions. The expanse of data required and the complexity of software managed by researchers often exceeds the means of a single scientist. Data and software sharing in the form of a distributed collection and analysis is inevitable. Creating effective artifacts, which enable scientists to collaborate on data analyses and reproduce them independently, continues to be a significant challenge for today’s science activities. It is rare that providing a file system abstraction on distributed data enables acceleration of scientific discoveries.

We will present Geounit Hub, a platform for sharing and reproducing scientific applications. Geounit software eliminates “does not work on my machine” problems when collaborating or reproducing code. Geoscientists share science applications by building ‘geounits’—self-contained, annotated, and versioned units that describe and package computational experiments in an efficient and light-weight manner. Other scientists reproduce the science application by re-running the lightweight, self-contained geounit. Geounit software guarantees that application software will always run the same, regardless of where it’s deployed. The Geounit Hub conforms to modern research data practices. Each geounit is versioned appropriately at the server, and maintains version information about your application software. Based on Globus, the national cyberinfrastructure for research data management services, the platform provides authentication, sharing, publication, and transfer of geounits.

Through use cases, we will demonstrate how Geounit Hub is advancing the state of reproducible research in the geosciences. In particular, we will show three use cases of computational reproducibility, in which a scientific experiment produced at time \( t \), must be exactly reproduced at time \( t' \). For each use case, we will show a different computational reproducibility requirement that must be addressed. Our first use case will be an example of setup reproducibility in which a scientist is unable to reproduce an output from a model due to the complexity of setting up and configuring the required development environments. Our second use case will be an example of algorithm reproducibility, where in a scientist wants to easily substitute a shared data science model with an alternate one to verify model output results, and finally an example of data reproducibility, in which an experiment is dependent on specific versions of data to produce the result.