Jul 11th, 9:50 AM - 10:10 AM

Challenges with Maintaining Legacy Software to Achieve Reproducible Computational Analyses: An Example for Hydrologic Modeling Data Processing Pipelines

Bakinam T. Essawy
University of Virginia, bte2rn@virginia.edu

Jonathan L. Goodall
University of Virginia, goodall@virginia.edu

Tanu Malik
University of Chicago, tanum@uchicago.edu

Hao Xu
University of North Carolina, xuh@cs.unc.edu

Michael Conway
University of North Carolina, michael_conway@unc.edu

Presenter/Author Information

Bakinam T. Essawy, Jonathan L. Goodall, Tanu Malik, Hao Xu, Michael Conway, and Yolanda Gil

International Environmental Modelling and Software Society (iEMSs)
8th International Congress on Environmental Modelling and Software
Toulouse, France, Sabine Sauvage, José-Miguel Sánchez-Pérez, Andrea Rizzoli (Eds.)
http://www.iemss.org/society/index.php/iemss-2016-proceedings

Challenges with Maintaining Legacy Software to Achieve Reproducible Computational Analyses: An Example for Hydrologic Modeling Data Processing Pipelines

Bakinam T. Essawy[a], Jonathan L. Goodall[a], Tanu Malik[b], Hao Xu[c], Michael Conway[c] and Yolanda Gil[d]

[a] *University of Virginia (bte2rn@virginia.edu,goodall@virginia.edu)*
[b] *University of Chicago (tanum@uchicago.edu)*
[c] *University of North Carolina (xuh@cs.unc.edu, michael_conway@unc.edu)*
[d] *University of Southern California (gil@isi.edu)*

Abstract: In hydrology, like many other scientific disciplines with large computational demands, scientists have created a significant and growing collection of software tools for data manipulation, analysis, and simulation. While core computation model software are likely to be well maintained by the groups that develop these codes, other software such as data pre- and post-processing tools, used less often but still critical to scientists, may receive less attention. These codes will become "legacy" software, simply meaning that the software is out of date by modern standards. A challenge facing the scientific community is how to maintain this legacy software so that it achieves reproducible results now and in the future, with minimal investment of resources. This talk will present an example of this problem in hydrology with the pre-processing tools used to create a Variable Infiltration Capacity (VIC) model simulation. The data processing pipeline for creating the input files for VIC is complex requiring code written over the years by various student researchers and sometimes requiring out-of-date compilers (e.g., FORTRAN 77) to compile portions of the code. We are confident that the use of legacy software is not a unique problem for VIC, but rather a wider problem common with other hydrologic models and scientific modeling in general. Through prior work, we have automated a VIC data processing pipeline, but moving these pipelines to new machines remains a significant challenge due in large part to the need to install legacy software dependencies. This work takes the following steps to address these challenges. The first step is to create containers using Docker to more easily execute legacy software across machines. This is done using the NSF funded projects GeoDataspace and Data Net Federation (DFC) to create and execute the Docker container as a Web application. The second step is to capture metadata for the large number of processing tools within the VIC data processing pipeline so that provenance of the software can be more easily tracked in the future. This is done using metadata frameworks created through the NSF funded HydroShare and OntoSoft projects. This methodology could serve as a general approach for making data processing pipelines more transparent and reproducible.

***Keywords**:* Reproducibility; legacy software; hydrologic modeling; Docker containers; metadata