Geophysical Research Abstracts
Vol. 19, EGU2017-8525, 2017
EGU General Assembly 2017

Reproducible Earth observation analytics: challenges, ideas, and a study case on containerized land use change detection

Marius Appel, Daniel Nüst, and Edzer Pebesma
University of Münster, Institute for Geoinformatics, Münster, Germany

Geoscientific analyses of Earth observation data typically involve a long path from data acquisition to scientific results and conclusions. Before starting the actual processing, scenes must be downloaded from the providers' platforms and the computing infrastructure needs to be prepared. The computing environment often requires specialized software, which in turn might have lots of dependencies. The software is often highly customized and provided without commercial support, which leads to rather ad-hoc systems and irreproducible results. To let other scientists reproduce the analyses, the full workspace including data, code, the computing environment, and documentation must be bundled and shared. Technologies such as virtualization or containerization allow for the creation of identical computing environments with relatively little effort. Challenges, however, arise when the volume of the data is too large, when computations are done in a cluster environment, or when complex software components such as databases are used.

We discuss these challenges for the example of scalable Land use change detection on Landsat imagery. We present a reproducible implementation that runs R and the scalable data management and analytical system SciDB within a Docker container. Thanks to an explicit container recipe (the Dockerfile), this enables the all-in-one reproduction including the installation of software components, the ingestion of the data, and the execution of the analysis in a well-defined environment. We furthermore discuss possibilities how the implementation could be transferred to multi-container environments in order to support reproducibility on large cluster environments.