Repeatable and reproducible workflows using the RENKU open science platform

Louis Krieger¹, Remko Nijzink¹, Gitanjali Thakur¹, Chandrasekhar Ramakrishnan², Rok Roskar², and Stan Schymanski¹
¹Luxembourg Institute of Science and Technology, Environmental Research and Innovation, Catchment and Eco-hydrology Research Group, Belvaux, Luxembourg (louis.krieger@list.lu)
²Swiss Data Science Center, Zurich, Switzerland

Good scientific practice requires good documentation and traceability of every research step in order to ensure reproducibility and repeatability of our research. However, with increasing data availability and ability to record big data, experiments and data analysis become more complex. This complexity often requires many pre- and post-processing steps that all need to be documented for reproducibility of final results. This poses very different challenges for numerical experiments, laboratory work and field-data analysis. The platform Renku (https://renkulab.io/), developed by the Swiss Data Science Center, aims at facilitating reproducibility and repeatability of all these scientific workflows. Renku stores all data, code and scripts in an online repository, and records in their history how these files are generated, interlinked and modified. The linkages between files (inputs, code and outputs) lead to the so-called knowledge graph, used to record the provenance of results and connecting those with all other relevant entities in the project.

We will discuss here several use examples, including mathematical analysis, laboratory experiments, data analysis and numerical experiments, all related to scientific projects presented separately. Reproducibility of mathematical analysis is facilitated by clear variable definitions and a computer algebra package that enables reproducible symbolic derivations. We will present the use of the Python package ESSM (https://essm.readthedocs.io) for this purpose, and how it can be integrated into a Renku workflow. Reproducibility of laboratory results is facilitated by tracking of experimental conditions for each data record and instrument re-calibration activities, mainly through Jupyter notebooks. Data analysis based on different data sources requires the preservation of links to external datasets and snapshots of the dataset versions imported into the project, that is facilitated by Renku. Renku also takes care of clear links between input, code and output of large numerical experiments, our last use example, and enables systematic updating if any of the input or code files are changed.

These different examples demonstrate how Renku can assist in documenting the scientific process from input to output and the final paper. All code and data are directly available online, and the recording of the workflows ensures reproducibility and repeatability.