EGU2020-19648, updated on 13 Aug 2022
https://doi.org/10.5194/egusphere-egu2020-19648
EGU General Assembly 2020

From source to sink - Sustainable and reproducible data pipelines with SaQC

David Schäfer, Bert Palm, Lennart Schmidt, Peter Lünenschloß, and Jan Bumberger
Helmholtz Centre for Environmental Research - UFZ

The number of sensors used in the environmental system sciences is increasing rapidly, and while this trend undoubtedly provides a great potential to broaden the understanding of complex spatio-temporal processes, it comes with its own set of new challenges. The flow of data from a source to its sink, from sensors to databases, involves many, usually error prone intermediate steps. From the data acquisition with its specific scientific and technical challenges, over the data transfer from often remote locations to the final data processing, all carry great potential to introduce errors and disturbances into the actual environmental signal.

Quantifying these errors becomes a crucial part of the later evaluation of all measured data. While many large environmental observatories are moving from manual to more automated ways of data processing and quality assurance, these systems are usually highly customized and hand written. This approach is non-ideal in several ways: First, it is a waste of resources as the same algorithms are implemented over and over again and second, it imposes great challenges to reproducibility. If the relevant programs are made available at all, they expose all problems of software reuse: correctness of the implementation, readability and comprehensibility for future users, as well as transferability between different computing environments. Beside these problems, related to software development in general, another crucial factor comes into play: the end product, a processed and quality controlled data set, is closely tied to the current version of the programs in use. Even small changes to the source code can lead to vastly differing results. If this is not approached responsibly, data and programs will inevitably fall out of sync.

The presented software, the 'System for automated Quality Control (SaQC)' (www.ufz.git.de/rdm-software/saqc), helps to either solve, or massively simplify the solution to the presented challenges. As a mainly no-code platform with a large set of implemented functionality, SaQC lowers the entry barrier for the non-programming scientific practitioner, without sacrificing the possibilities to fine-grained adaptation to project specific needs. The text based configuration allows the easy integration into version control systems and thus opens the opportunity to use well established software for data lineage. We will give a short overview of the program's unique features and showcase possibilities to build reliable and reproducible processing and quality assurance pipelines for real-world data from a spatially distributed, heterogeneous sensor network.