Criteria for the Evaluation of Workflow Management Systems for Scientific Data Analysis

Aleyna Dilan Kiran kirandilan@gmail.com Turkey Ege University

Mehmet Can Ay mehmettt.can.ay@gmail.com Turkey Izmir Institute of Technology

Jens Alllmer jens@allmer.de Germany Hochschule Ruhr West

Many scientific endeavors, such as molecular biology, have become dependent on large-scale data and its analysis. For example, precision medicine depends on molecular measurements and data analysis on a per-patient basis. Data analysis, supporting medical decisions, has to be standardized and performed in a consistent manner across patients. While perhaps not life-threatening, data analyses in basic research have become increasingly complex. RNA-seq data, for example, entails a multi-step analysis ranging from quality assessment of the measurements to statistical analyses.

Workflow management systems (WFMS) enable the development of data analysis workflows (WF), their reproduction, and their application to datasets of the same type. However, there are far more than a hundred WFMS available to choose from and no way to convert data analysis WFs among WFMS. Therefore, the initial choice of a WFMS is important as it entails a lock-in to the system. Perhaps the reach in the particular field (number of citations) can be used as a proxy for the selection of a WFMS, but of the about 25 WFMS we mention in this work, at least 5 have a large reach in scientific data analysis.

Hence other criteria are needed to delineate among WFMS. By extracting such criteria from selected studies concerning WFMS and adding additional criteria, we arrived at five critical (reproducibility, reusability, FAIRness, versioning support, and security) and five important criteria (providing a graphical user interface, WF flexibility, WF scalability, WF shareability, and computational transparency) for the assessment of WFMS. We applied the criteria to the most cited WFMS in Pubmed and found that none of them support all criteria. We hope that suggesting these criteria will spark a discussion on what features are important for WFMS in scientific data analysis and perhaps will lead to the development of WFMS that fulfill such criteria.