deRSE CONFERENCE WÜRZBURG 2024

Contribution ID: **114** Type: **Talk (15min + 5min)**

What do GitHub repositories of Potsdam researchers tell us about quality and reproducibility of their scientific software?

Tuesday, March 5, 2024 4:50 PM (20 minutes)

To enable and ensure the reproducibility and traceability of scientific claims, it is essential that scientific publications, the corresponding datasets, and the data analysies code are made publicly available.The adoption of the FAIR4RS principles and software engineering best practices could significantly enhance the success of delivering a codebase that produces consistent, reproducible results. Yet, not much is known about how practices like version control, continuous integration, scientific data management, automated testing and software documentation and citation are adopted within scientific community.
To gain a better understanding of the standards and practices followed by research software developers in Potsdam, we identified GitHub users who are PhD candidates, PostDocs, researchers or other academics affiliated with the University of Potsdam or one of the local research institutes. This focus is motivated by our goal of fostering collaboration and engagement with the research software development community in Potsdam. We then collected 13000+ open-source repositories of those users for thorough study to access coding practices and standards followed by them. The repositories collection has significant number of projects that are data analysis and web development with frequent mentions of keywords like data, web, machine learning along with use of programming languages like python and javascript. There is also a notable presence of non-technical projects, including educational repositories identified by terms like course, teaching, and workshop, along with other repositories with keywords survey, study, and assignments, thesis. To focus only on research repositories which are actually research software we classified them according to the DLRs application classes . This helped us to eliminate lot of repositories which do not contain relevant artefact and which are not meant for generating reproducible research results. We found 2100+ projects which falls under DLR application class 0 repositories which does not have relevant artefact that aligns with research software. In our study, which is presented as a short talk, we want continue to classify all open-source projects from the data collected into the DLR application classes with the goal of carefully assessing the use of git, git work flows, automated testing, and how they organize their code, test files, and documentation. Additionally, our goal is to assess the level of accessibility and quality of scientific documentation, code commenting, that ensure software reproducibility. With this short talk we hope to spark a discussion and further collaboration on this topic.

Slot length

Primary authors: DEVKATE, Akshay; Prof. LAMPRECHT, Anna-Lena

Presenter: DEVKATE, Akshay

Session Classification: RSE Research

Track Classification: Research Software: FAIRification and Its Implications for Research Software