Contribution ID: **72** Type: **Talk (15min + 5min)**

Fine-grained exploration of the reproducibility of research-related Jupyter notebooks at scale

Wednesday 26 February 2025 16:00 (20 minutes)

Jupyter notebooks have revolutionized the way researchers share code, results, and documentation, all within an interactive environment, promising to make science more transparent and reproducible. In research contexts, Jupyter notebooks often coexist with other software and various resources such as data, instruments, and mathematical models, all of which may affect scientific reproducibility. Here, we present a study that analyzed the computational reproducibility of 27,271 Jupyter notebooks from 2,660 GitHub repositories associated with 3,467 biomedical publications (https://doi.org/10.1093/gigascience/giad113). The resulting reproducibility data were loaded into a knowledge graph –FAIR Jupyter– that allows for a highly granular exploration and interrogation.

The FAIR Jupyter graph is accessible via https://w3id.org/fairjupyter and described in a preprint available at https://doi.org/10.48550/arXiv.2404.12935 . It contains rich metadata about the publications, associated GitHub repositories and Jupyter notebooks, and the notebooks' dependencies and reproducibility. Through a public SPARQL endpoint, it enables detailed data exploration and analysis by way of queries that can be tailored to specific use cases. Such queries may provide details about any of the variables from the original dataset, highlight relationships between them or combine some of the graph's content with materials from corresponding external resources.

We provide a collection of example queries addressing a range of use cases in research software engineering and education. We also outline how sets of such queries can be used to profile specific content types, either individually or by class. We conclude by discussing how such a semantically enhanced sharing of complex datasets can both enhance their FAIRness i.e., their findability, accessibility, interoperability, and reusability, and help identify and communicate best practices, particularly with regards to the quality, standardization and reproducibility of research-related software and scripts.

I want to participate in the youngRSE prize

Primary author: Dr MIETCHEN, Daniel (FIZ Karlsruhe —Leibniz Institute for Information Infrastructure, Germany)

Co-author: SAMUEL, Sheeba

Presenter: Dr MIETCHEN, Daniel (FIZ Karlsruhe —Leibniz Institute for Information Infrastructure, Germany)

Session Classification: Reproducibility and Discovery of Research Software

Track Classification: Data and Software Management: computational reproducibility