FAIR Enabling Re-Use of Data-Intensive Workflows and Scientific Reproducibility

Line C. Pouchard
Brookhaven National Laboratory, Upton, NY 11973, USA
pouchard@bnl.gov
ORCID: 0000-0002-2120-6521

ABSTRACT

Scientific computing communities often run their experiments using complex data- and compute-intensive workflows that utilize high performance computing (HPC), distributed clusters and specialized architectures targeting machine learning and artificial intelligence. FAIR principles for data and software can be useful enablers for the reproducibility of performance (a key HPC metric) and that of scientific results (a crucial tenet of the scientific method) that are based in part on re-use, the R of FAIR principles. FAIR principles are under-used by HPC and data-intensive communities who have been slow to adopt them. This is due in part to the complexity of workflow life cycles, the numerous workflow management systems, the lack of integration of FAIR within existing technologies, and the specificity of managed systems that include rapidly evolving architectures and software stacks, and execution models that require resource managers and batch schedulers. Numerous challenges emerge for scientists attempting to publish FAIR datasets and software for the purpose of re-use and reproducibility, e.g. what data to publish and where due to sizes, how to "FAIRify" data subsetting, at what level of granularity to attribute persistent identifiers to software, what is the minimal amount of metadata needed to guarantee a certain level of reproducibility, what does reproducible AI actually mean? This talk will focus on such challenges and illustrate the negative impact of not applying FAIR on the reproducibility of experiments. We will introduce the notion of FAIR Digital Objects and present RECUP, a framework for data and metadata services for high performance workflows that proposes micro-solutions for adapting FAIR principles to HPC.

CCS Concepts

Information Systems -> Data Management Systems; Information Systems -> Information Systems Applications

Keywords: FAIR; HPC; high performance computing; data intensive; reproducibility; FDO; RECUP

ACM Reference format:
Line C. Pouchard. 2023. FAIR enabling re-use of data-intensive workflows and scientific reproducibility. In *the Companion of the 2023 ACM/SPEC International Conference on Performance Engineering (ICPE'23 Companion), April 15-19, 2023, Coimbra, Portugal.* ACM, New York, NY, USA, 1 pages. https://doi.org/10.1145/3545945.3586012

ICPE'23 Companion, April 15-19, 2023, Coimbra, Portugal.

ACM ISBN 979-8-4007-0072-9/23/04. https://doi.org/10.1145/3578245.3586012

BIOGRAPHY

Line C. Pouchard is an internationally recognized expert with more than two decades of experience in computational science with over 100 publications. She leads multi-disciplinary technical projects to create innovative approaches for scientific data discovery, high performance and data-intensive workflows, and FAIR data management and curation. Her present research focuses on provenance for workflows at scale, computational reproducibility, and text mining for big data. Prior to her current position at Brookhaven National Laboratory, she was Staff Scientist at Oak Ridge National Laboratory, and Assistant Professor at Purdue University. She has a PhD from the Graduate Center of the City University of New York, and an MS from the University of Tennessee, Knoxville.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the funding support from the U.S. Department of Energy Office of Science. This paper has been authored by employees of Brookhaven Science Associates, LLC under Contract No. DESC0012704.

REFERENCES

Line Pouchard, Tanzima Islam, Bogdan Nicolae. 2022. Challenges for Implementing FAIR Digital Objects with High Performance Workflows. Research Ideas and Outcomes 8. https://doi.org/10.3897/rio.8.e94835