Challenges for Implementing FAIR Digital Objects with High Performance Workflows

Line C. Pouchard‡, Tanzima Z. Islam§, Bogdan Nicolae|

‡ Brookhaven National Laboratory, Upton, NY, United States of America
§ Texas State University, San Marco, TX, United States of America
| Argonne National Laboratory, Argonne, IL, United States of America

Corresponding author: Line C. Pouchard (pouchard@bnl.gov)

Abstract

New types of workflows are being used in science that couple traditional distributed and high-performance computing (HPC) with data-intensive approaches, and orchestrate ensembles of numerical simulations and artificial intelligence (AI) models. Such workflows may use AI models to supplement computation where numerical simulations may be too computationally expensive, to automate trivial yet time consuming operations, to perform preliminary selections among intractable numbers of combinations in domains as diverse as protein binding, fine-grid climate simulations, and drug discovery.

They offer renewed opportunities for scientific research but exhibit high computational, storage and communications requirements [Goble et al. 2020, Al-Saadi et al. 2021, da Silva et al. 2021]. These workflows can be orchestrated by workflow management systems (WMS) and built upon composable blocks that facilitate task placement and resource allocation for parallel executions on high performance systems [Lee et al. 2021, Merzky et al. 2021].

The scientific computing communities running these kinds of workflows have been slow to adopt Findable, Accessible, Interpretable, and Re-usable (FAIR) principles, in part due to the complexity of workflow life cycles, the numerous WMS, and the specificity of HPC systems with rapidly evolving architectures and software stacks, and execution modes that require resource managers and batch schedulers [Plale et al. 2021]. FAIR Digital Objects (FDO) that encapsulate bit sequences of data, metadata, types and persistent identifiers (PID) can help promote the adoption of FAIR, enable knowledge extraction and dissemination, and contribute to re-use [De Smedt et al. 2020]. As workflows typically use data and software during planning and execution, FDOs are particularly adapted to enable re-use [Wittenburg et al. 2020]. But the benefits of FDOs such as automating data processing and actionable DO collections cannot be realized without the main components of FAIR, rich metadata and clear identifiers, being universally adopted in the community. These components are still elusive for HPC digital objects. Some metadata are added after results have been produced, are not described by controlled vocabularies, and

typically left unconstrained, resulting in inefficient processes and loss of knowledge. Persistent identifiers are added at the time of publication to data supporting conclusions, so only a very small amount of data are being shared outside a small community of researchers “in the know”.

In this conceptual work, one can distinguish several kinds of FDOs for HPC workflows that present both common and specific challenges to the development of canonical DO infrastructure and the implementation of FDO workflows that we discuss below:

- *result FDOs* represent computational results obtained when program execution complete,
- *performance FDOs* that contain performance measures and results from code optimization on parallel, heterogeneous architectures,
- *intermediate FDOs* from intermediate states of workflow execution, obtained from HPC checkpointing.

All these FDOs for HPC workflows should include the computing environment and system specifications on which code was executed for metadata rich enough to enable re-usability [Pouchard et al. 2019]. Containers are often being used to capture dependencies between underlying libraries and versions in the execution environment for the installation and re-use of software code [Lofstead et al. 2015, Olaya et al. 2020]. But containers published in code repositories are made available without identifiers registered with resolvers. For instance, to attribute a Digital Object Identifier to software shared in github, one must perform the additional step of registering the code into Zenodo. FDOs extracted and built in the context of a canonical workflow framework including collections will help with the attribution of persistent identifiers and the linking of execution environment with data and workflow.

Computational results may include machine learning predictions resulting form stochastic training of non-deterministic models. Neural networks and deep learning models present specific challenges to *result FDOs* related to provenance and the selection of quantities needed to include in an FDO for the re-use of results. What information needs to be included in a FAIR Digital Object encapsulating deep learning results to make it persistent and re-usable? The description of method, data and experiment recommended in [Gundersen and Kjensmo 2018] can be instantiated in a FDO collection. To make it re-usable, it should include the model architecture, the machine learning platform and its version, a submission script that contains hyperparameters, the loss function, batch size and number of epochs [Pouchard et al. 2020].

Challenges specific to digital objects containing performance measures for HPC workflows are those related to size, selection and reduction. Performance data at scale tends to be very large, thus a principled approach to selection is needed to determine which execution counters must be included in FDOs for performance reproducibility of an application [Patki et al. 2019]. *Performance FDOs* should include the variables selected to show their impact on performance and the methods used for selection: do such variables represent outliers in

performance metrics? What methods and thresholds are used to qualify as outliers, what impact do these outliers have on overall performance of an execution?

A key contributor to the failure to capture important information in HPC workflows is that metadata and provenance capture is often “bolted on” after the fact and in a piecemeal, cumbersome, inefficient manner that impedes further analysis. An FDO approach including DO collections at the appropriate level of abstraction and rich metadata is needed. Capturing metadata automatically must take into account the appropriate granularity level for re-use across system layers and abstraction levels. *Intermediate FDOs* capture and fuse metadata across multiple sources during the planning and execution stages [Nicolae 2022]. Some tools already exist. Darshan is a scalable tool summarizing Input/Output file characteristics [Dai et al. 2019], Radical Cybertools [Merzky et al. 2021] can produce the provenance task graph of an execution. Such tools could be included in a canonical workflow framework as they present a path forward for composable services for HPC and would guarantee a level of encapsulation into DOs favorable to re-use.

Keywords

FAIR Digital Object, FDO,High Performance Computing, HPC, FAIR4HPC

Presenting author

Line C. Pouchard

Presented at

First International Conference on FAIR Digital Objects, presentation

Acknowledgements

The submitted manuscript has been created in part by 1) Brookhaven Science Associates, LLC operator of Brookhaven National Laboratory, a U.S Department of Energy Office of Science laboratory operated under Contract No. DESC0012704, 2) by UChicago Argonne, LLC, Operator of Argonne National Laboratory, a U.S. Department of Energy Office of Science laboratory, operated under Contract No. DE-AC02-06CH11357.

Ethics and security

This is a concept paper. No ethics and/or security concerns.

Author contributions

Line Pouchard conceptualized the presentation and wrote the manuscript, Tanzima Islam and Bogdan Nicolae provided feedback and inspiration during work development

Conflicts of interest

N/A

References

- Al-Saadi A, Ahn D, Babuji Y, Chard K, Corbett J, Hategan M, Herbein S, Jha S, Laney D, Merzky A, Munson T, Salim M, Titov M, Turilli M, Uram T, Wozniak J (2021) ExaWorks: Workflows for Exascale. 2021 IEEE Workshop on Workflows in Support of Large-Scale Science (WORKS). https://doi.org/10.1109/WORKS54523.2021.00012
- Dai D, Chen Y, Carns P, Jenkins J, Zhang W, Ross R (2019) Managing Rich Metadata in High-Performance Computing Systems Using a Graph Model. IEEE Transactions on Parallel and Distributed Systems 30 (7): 1613-1627. https://doi.org/10.1109/TPDS.2018.2887380
- da Silva RF, Casanova H, Chard K, Altintas I, Badia RM, Balis B, Coleman T, Coppens F, Di Natale F, Enders B, Fahringer T, Filgueira R, Fursin G, Garijo D, Goble C, Howell D, Jha S, Katz D, Laney D, Leser U, Malawski M, Mehta K, Pottier L, Ozik J, Peterson JL, Ramakrishnan L, Soiland-Reyes S, Thain D, Wolf M (2021) A Community Roadmap for Scientific Workflows Research and Development. 2021 IEEE Workshop on Workflows in Support of Large-Scale Science (WORKS). https://doi.org/10.1109/WORKS54523.2021.00016
- De Smedt K, Koureas D, Wittenburg P (2020) FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units. Publications 8 (2). https://doi.org/10.3390/publications8020021
- Goble C, Cohen-Boulakia S, Soiland-Reyes S, Garijo D, Gil Y, Crusoe M, Peters K, Schober D (2020) FAIR Computational Workflows. Data Intelligence 2 (1-2): 108-121. https://doi.org/10.1162/dint_a_00033
- Gundersen OE, Kjensmo S (2018) State of the Art: Reproducibility in Artificial Intelligence. Proceedings of the AAAI Conference on Artificial Intelligence 32 (1). URL: https://ojs.aaai.org/index.php/AAAI/article/view/11503
- Lee H, Merzky A, Tan L, Titov M, Turilli M, Alfe D, Bhati A, Brace A, Clyde A, Coveney P, Ma H, Ramanathan A, Stevens R, Trifan A, Dam HV, Wan S, Wilkinson S, Jha S (2021) Scalable HPC and AI infrastructure for COVID-19 therapeutics. Platform for Advanced Scientific Computing Conference (PASC '21), July 5–9, 2021, Geneva, Switzerland. ACM, New York, NY, USA https://doi.org/https://doi.org/10.1145/3468267.3470573.
- Lofstead GF, Jimenez I, Maltzahn C, Moody A, Mohror K, Arpaci-Dusseau R (2015) The Role of Container Technology in Reproducible Computer Systems Research. Sandia National Laboratories (SNL-NM), Albuquerque, NM (United States).

- Merzky A, Turilli M, Titov M, Al-Saadi A, Jha S (2021) Design and Performance Characterization of RADICAL-Pilot on Leadership-class Platforms. arXiv. DOI: 10.48550/arXiv.2103.00091 arXiv:2103.00091 [cs] type: article. URL: http://arxiv.org/abs/2103.00091
- Nicolae B (2022) Scalable Multi-Versioning Ordered Key-Value Stores with Persistent Memory Support. URL: https://hal.archives-ouvertes.fr/hal-03598396
- Olaya P, Lofstead J, Taufer M (2020) Building Containerized Environments for Reproducibility and Traceability of Scientific Workflows. arXiv:2009.08495 [cs] URL: http://arxiv.org/abs/2009.08495
- Patki T, Thiagarajan J, Ayala A, Islam T (2019) Performance optimality or reproducibility: that is the question. Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. [ISBN 978-1-4503-6229-0]. https://doi.org/10.1145/3295500.3356217
- Plale B, Malik T, Pouchard L (2021) Reproducibility Practice in High-Performance Computing: Community Survey Results. Computing in Science Engineering 23 (5): 55-60. https://doi.org/10.1109/MCSE.2021.3096678
- Pouchard L, Baldwin S, Elsethagen T, Jha S, Raju B, Stephan E, Tang L, Dam KKV (2019) Computational reproducibility of scientific workflows at extreme scales. The International Journal of High Performance Computing Applications https://doi.org/10.1177/1094342019839124
- Pouchard L(, Lin Y, van Dam H((2020) Replicating Machine Learning Experiments in Materials Science. IOS Press. DOI: 10.3233/APC200105. URL: https://www.osti.gov/biblio/1635098
- Wittenburg P, czmiel, Betz D, Wieder P, Zünkeler M, Gülzow V, Jianhui L, Öster P, Spinuso A, Stotzka R (2020) CWFR Position Paper. n/ URL: https://osf.io/3rekv/