Towards a dataset of executions of computational experiments

SAMUEL GRAYSON, University of Illinois Urbana Champaign, USA
DANIEL S. KATZ, University of Illinois Urbana Champaign, USA
REED MILEWICZ, Sandia National Laboratories, USA
DARKO MARINOV, University of Illinois Urbana Champaign, USA

We call on the researcher community to collaborate to create a FAIR dataset of executions of computational experiments.

There are registries of workflows, which are often computational experiments, such as WorkflowHub [1], but these registries do not store any execution data, not even a command for executing them.

While some execution artifacts are large, some useful execution details are small, including: the command used to execute the experiment, process success, date/time, platform (OS and architecture), compute resources used (CPU time, wall time, memory, disk space, network bandwidth). For each file read or written: path or URL, type (returned by `file`), 'summary' (hash for binary files, summary statistics for numerical data, etc.), truncated contents. These can automatically be collected from an execution without domain knowledge; if a domain specialist is available, they could add a machine-readable description of what the input variables are and how the outputs are to be compared.

For an individual project, this eases answering the following questions: Is the experiment replicable? How big of a machine is needed to replicate it? What output should one expect? What files need to exist? If the end output is different, which is the earliest intermediate file that differs significantly? Can tolerably similar output be produced with fewer computational resources (by changing the fidelity parameters)?

A collection of projects with this data would enable the following research questions: What libraries do these computational experiments use directly or transitively? Are these computational experiments repeatable in the short-term? What about the long-term (months or years)? Or repeatable within some margin of error? For those that crash, what are the common causes of crashes (could compare to Zhao et al. [2])?

Large-scale studies are the best candidates to collect this dataset. Most large-scale studies, like Zhao et al. [2] , do not share execution details of individual computational experiments, but only aggregate results. Collberg and Proebsting [3] report the success or failure and how to run individual computational experiments but no other data. Trisovic et al. [4] report the success or failure of individual R scripts in Harvard Dataverse but no other data.

We have taken steps towards collecting this dataset for nf-core, Snakemake Workflow Catalog, and R scripts in Harvard Dataverse. Our data is publicly available.

References

[1] Ferreira da Silva, Pottier, Coleman, Deelman, and Casanova. WorkflowHub: Community Framework for Enabling Scientific Workflow Research and Development. In 2020 WORKS.

[2] Zhao, Gomez-Perez, Belhajjame, Klyne, Garcia-cuesta, Garrido, Hettne, Roos, De Roure, and Goble. Why workflows break — understanding and combating decay in Taverna workflows. In 2012 E-Science

[3] Collberg and Proebsting. Repeatability in computer systems research. Commun. ACM Feb 2016

[4] Trisovic, Lau, Pasquier, and Crosas. A large-scale study on research code quality and execution. Scientific Data Feb 2022

United States Research Software Engineer Association Conference 2023, Oct 15–17, 2023, Chicago, IL
2023. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . . $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn