

Towards a dataset of executions of computational experiments

SAMUEL GRAYSON, University of Illinois Urbana-Champaign, USA

REED MILEWICZ, Sandia National Laboratories, USA

DARKO MARINOV, University of Illinois Urbana-Champaign, USA

DANIEL S. KATZ, University of Illinois Urbana-Champaign, USA

The community of researchers should collaborate to create a FAIR dataset of executions of computational experiments.

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

While some of the execution artifacts are large, some useful execution details are small enough, 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.

The above are automatically collectible 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? 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 enables the following research questions:

- What libraries do these computational experiments use or transitively use?
- 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.)?

Large-scale studies are the best candidates to collect this dataset. Most large-scale studies, like Zhao et al., do not share any execution details of individual computational experiments (only aggregate results). Collberg and Proebsting report the success or failure and how to run individual computational experiments but no other data. Trisovic et al. 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.