Representing bioinformatics Nextflow workflows in RO-Crate : challenges and opportunities

George Marchment[a], Marie Schmit[b], Clémence Sebe[a], Frédéric Lemoine[b], Hervé Ménager[b], Sarah Cohen-Boulakia[a]

[a] *Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, Orsay 91405, France.*
[b] *Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, Paris, France, 28, rue du Dr Roux, Paris 75015, France.*

Abstract

This research was conducted as part of the ShareFAIR project dedicated to assisting bioinformaticians in creating, comparing, and exchanging robust analysis workflows for multi-scale datasets related to neuro-vascular pathologies. ShareFAIR partners use workflows from diverse types of systems. One important challenge lies in uniformly and simply representing workflows such that partners can understand, share and reuse them. The aim of this research is to evaluate how the standards currently available, especially RO-Crate, allow to represent workflows.

Keywords
Workflows, RO-Crate, Nextflow

1. Body

This research was conducted as part of the ShareFAIR project (https://projet.liris.cnrs.fr/sharefair/), a collaborative initiative involving nine French research partners. ShareFAIR is dedicated to assisting bioinformaticians in creating, comparing, and exchanging robust analysis workflows for multi-scale datasets related to neuro-vascular pathologies, encompassing genomic, neuro-vascular imaging, and clinical data.

ShareFAIR partners use workflows from diverse types of systems (e.g., Snakemake [3], Nextflow [2], Galaxy [5]). One important challenge lies in uniformly and simply representing workflows such that partners can understand, share and reuse them.

The work presented here specifically focuses on workflows from the Nextflow workflow system where already two distinct types of workflows can be designed, namely, domain specific language DSL1 and DSL2. Available Nextflow workflows are equally distributed between DSL1 and DSL2.

Our aim is to evaluate how the standards currently available, especially RO-Crate [4], allow to describe Next -flow workflows, both from DSL1 and DSL2, at various levels of granularity.

To do so, we have collected over 1,500 Nextflow workflows using a dedicated crawler specifically designed to extract public Nextflow workflows from GitHub repositories. Figure 1 shows the number of workflows found on GitHub by creation date, as extracted by our crawler. We then parsed and analysed our workflow collection, in order to extract several pieces of information. This includes the workflow's metadata, its subworkflows and its constituting processes alongside their inputs and outputs, thus forming a comprehensive dataset.

Parsing and analysing this dataset allowed us to realise how heterogeneous the diverse implementations of Nextflow workflows are. Annotating and describing them in a homogeneous way would greatly facilitate their sharing, comparison and interrogation. RO-Crate emerges as a strong contender for this undertaking.

CEUR Workshop Proceedings (CEUR-WS.org)
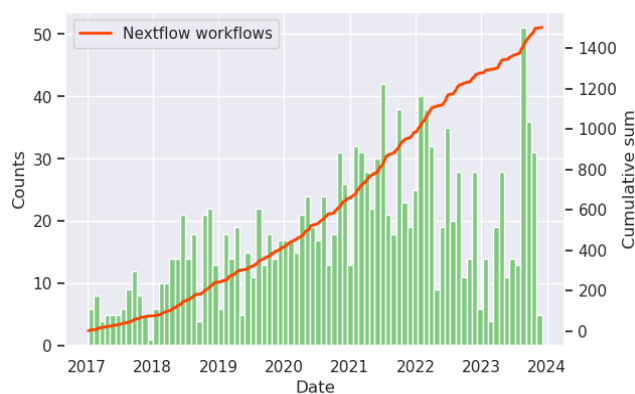
Figure 1: Evolution of the monthly and cumulative number of Nextflow workflows available on GitHub since 2017

RO-Crate is a standard for aggregating and describing research data along with associated metadata. It allows, among other things, to describe workflows and scripts. However, the framework provided by RO-Crate may not be fully suitable to describe workflows at a high level of detail (e.g., up to the data flow structure or detailed process description). To do so, it may be adapted to reach a higher level of granularity.

In this study, we investigate the possibilities offered by RO-Crate for describing Nextflow workflows and present solutions to enhance it for capturing a more advanced level of workflow information.

Ultimately, we aim to extend this work to Snakemake and Galaxy workflows, enabling cross-platform comparisons between workflows. Additionally, we intend to investigate how other solutions, such as the Common Workflow Language [1], could contribute to improving workflow descriptions.

2. References

[1] Crusoe, M. R.; Abeln, S.; Iosup, A.; Amstutz, P.; Chilton, J.; Tijanic, N.; Ménager, H.; Soiland-Reyes, S.; Gavrilovic, B.; Goble, C.; Community, T. C.: Methods included: standardizing computational reuse and portability with the Common Workflow Language. Communications of the ACM 65/6, S. 54–63, Mai 2022, I SSN: 0001-0782, URL: https://dl.acm.org/doi/10.1145/3486897

[2] Di Tommaso, P.; Chatzou, M.; Floden, E. W.; Barja, P. P.; Palumbo, E.; Notredame, C.: Nextflow enables reproducible computational workflows. en, Nature Biotechnology 35/4, Number: 4 Publisher: Nature Publishing Group, S. 316–319, Apr. 2017, ISSN: 1546-1696, URL: https://www.nature.com/ articles/nbt.3820

[3] Köster, J.; Rahmann, S.: Snakemake—a scalable bioinformatics workflow engine. Bioinformatics 28/19, S. 2520–2522, Okt. 2012, ISSN: 1367-4803, UR L: https://doi.org/10.1093/bioinformatics/bts480

[4] Soiland-Reyes, S.; Sefton, P.; Crosas, M.; Castro, L. J.; Coppens, F.; Fernández, J. M.; Garijo, D.; Grüning, B.; La Rosa, M.; Leo, S.; Ó Carragáin, E.; Portier, M.; Trisovic, A.; Community, R.-C.; Groth, P.; Goble, C.: Packaging research artefacts with RO-Crate. en, Data Science 5/2, Publisher: IOS Press, S. 97–138, Jan. 2022, I SSN: 2451-8484, URL: https://content.iospress.com/ articles/data-science/ds210053

[5] The Galaxy Community: The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. Nucleic Acids Research 50/W1, W345–W351, Juli 2022, ISSN: 0305-1048, URL: https://doi.org/ 10.1093/nar/gkac247