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

George Marchment, Marie Schmit, Clémence Sebe, Frédéric Lemoine, Hervé Ménager, Sarah Cohen-Boulakia

Abstract

ShareFAIR partners use diverse kinds of workflow systems (e.g., Snakemake [2], Nextflow [1], Galaxy [4]). A crucial challenge involves creating workflows in a **clear and standardised manner**, enabling partners to easily **comprehend**, **exchange**, and **utilise them**.

Our aim is to evaluate the capacity of current standards, particularly **RO-Crate** [3], to describe **Nextflow workflows** (to begin with).

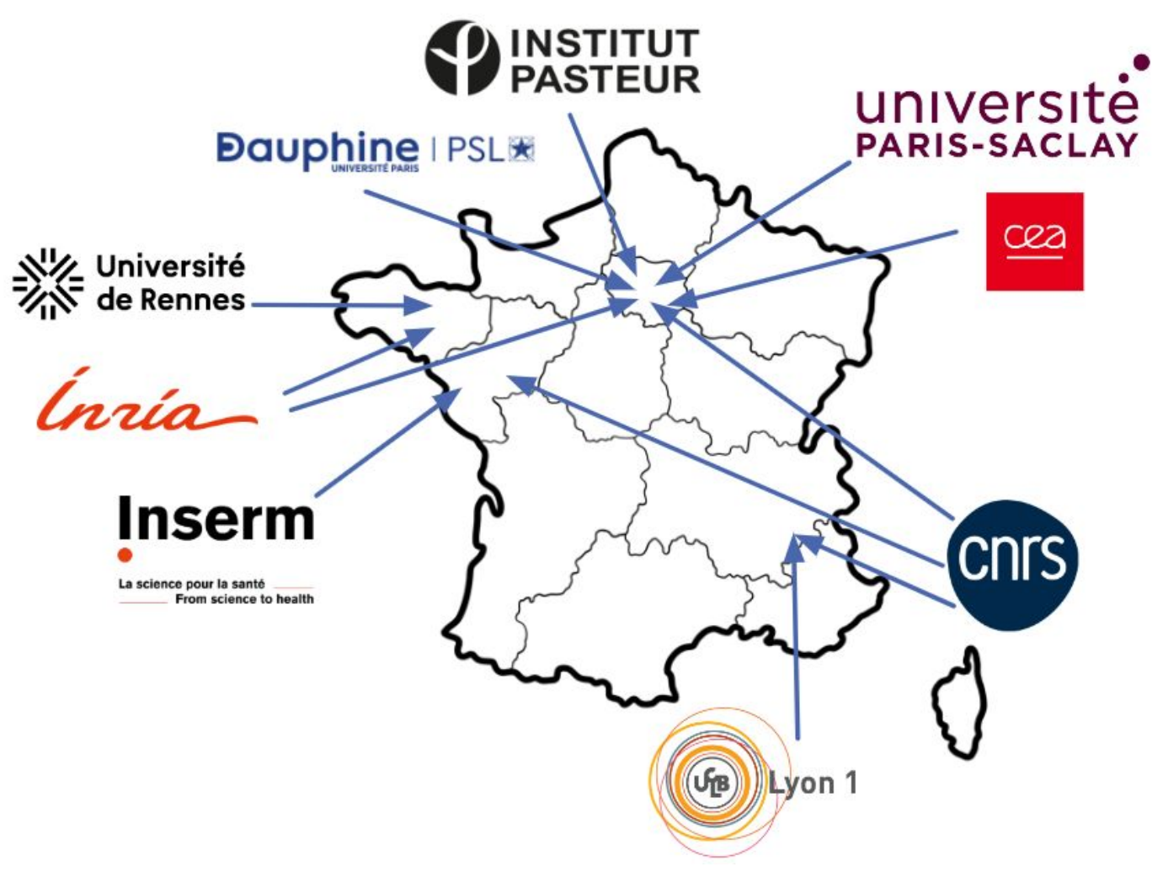
To do so, we have collected over 1,500 Nextflow workflows using a dedicated crawler specifically designed to extract public Nextflow workflows from GitHub repositories. Figure 1 shows the number of workflows found on GitHub by creation date, as extracted by our crawler. We then parsed and analysed our workflow collection, in order to extract several pieces of information. This includes the workflow's metadata, such as its **authors**, **publication date**, etc; its **subworkflows**; and its constituting **processes** alongside their **inputs and outputs**, thus forming a comprehensive dataset (Figure 2 depicts a simplified representation of a workflow and its steps).

Studying 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.

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., detailed subworkflow or 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 enrich it for capturing more fine-grained workflow informations**, and that can be extended to **Snakemake** and **Galaxy** workflows.

ShareFAIR



This research was conducted as part of the **ShareFAIR** project, 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.

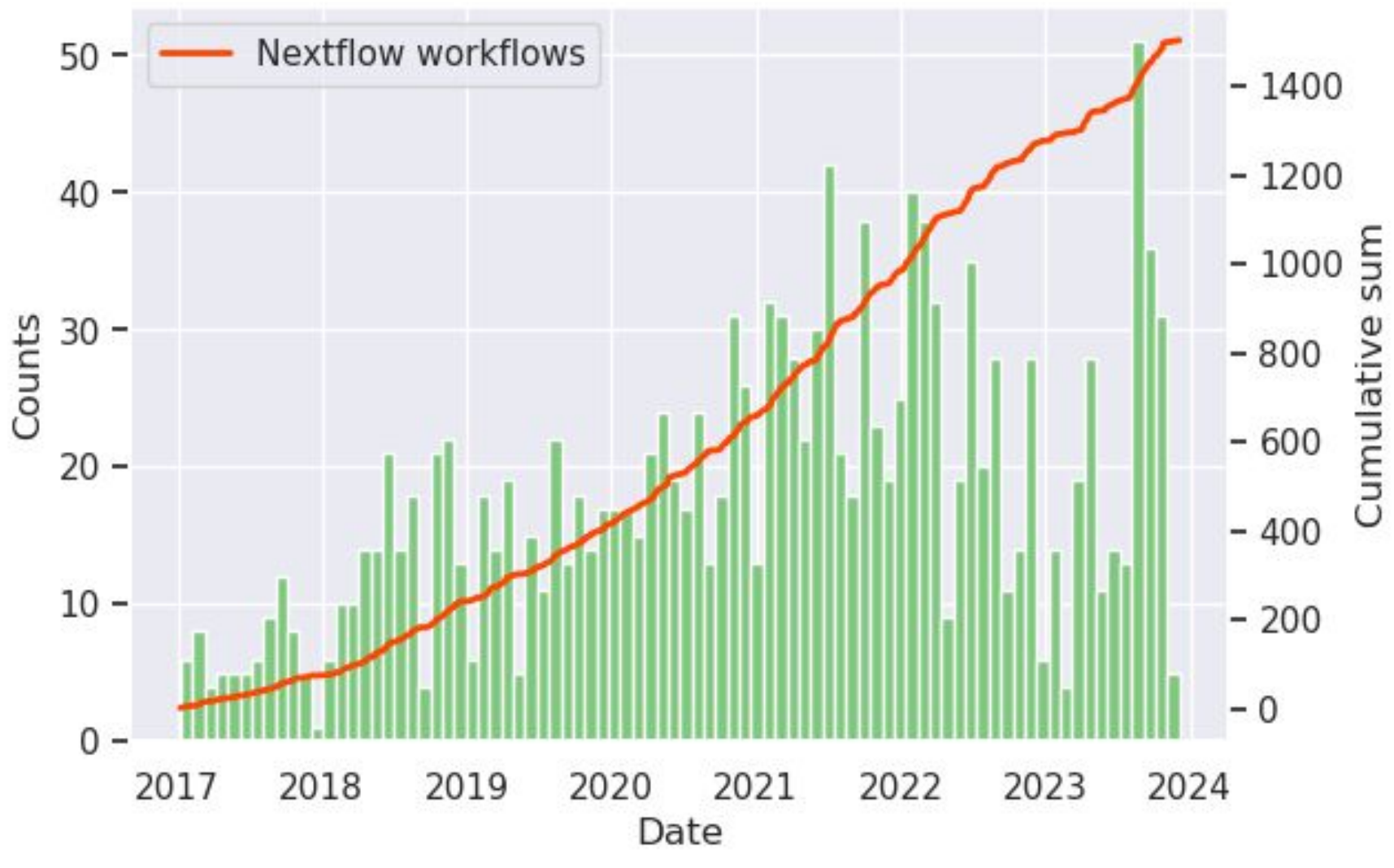


Figure 1: Evolution of the monthly and cumulative number of Nextflow workflows available on GitHub since 2017

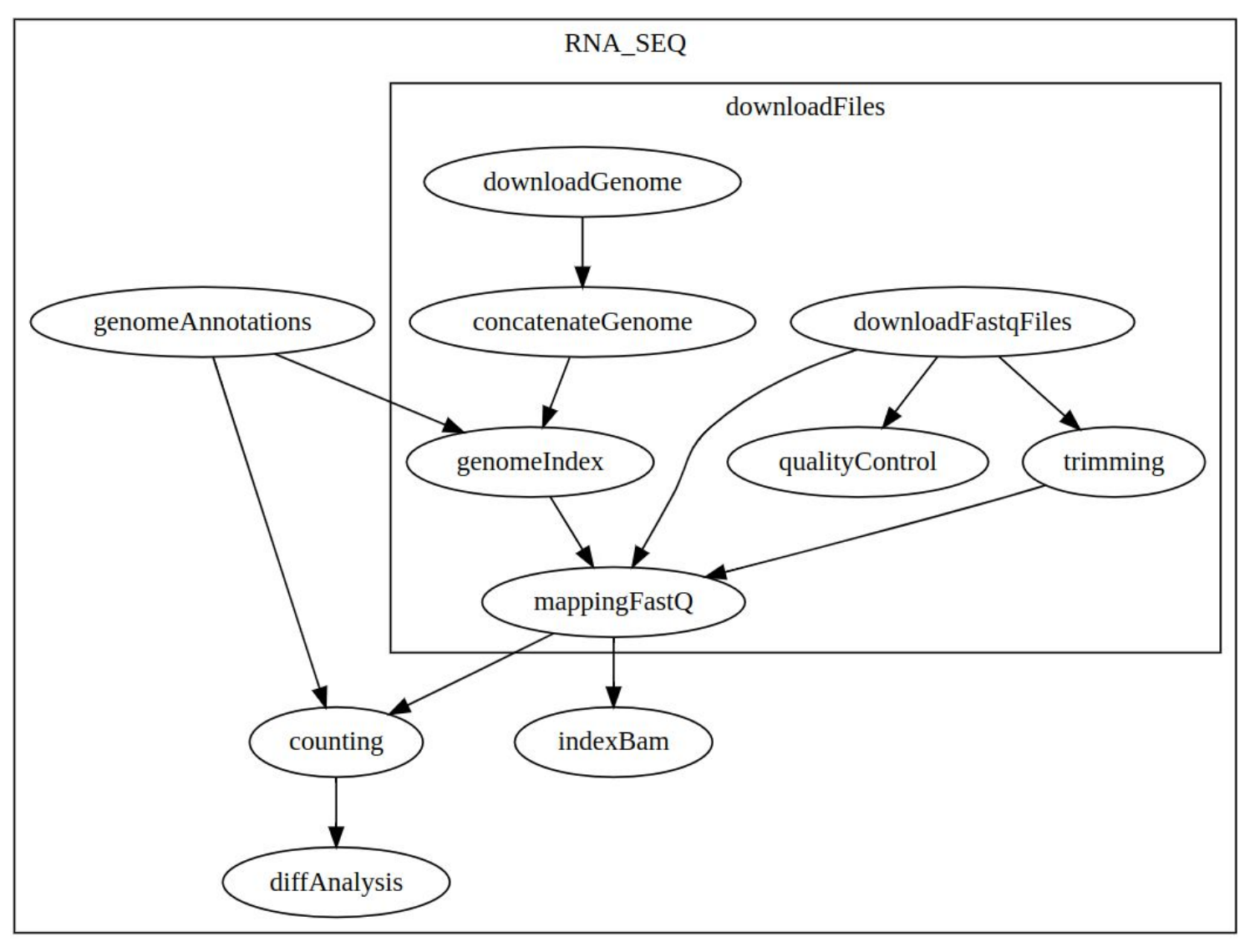
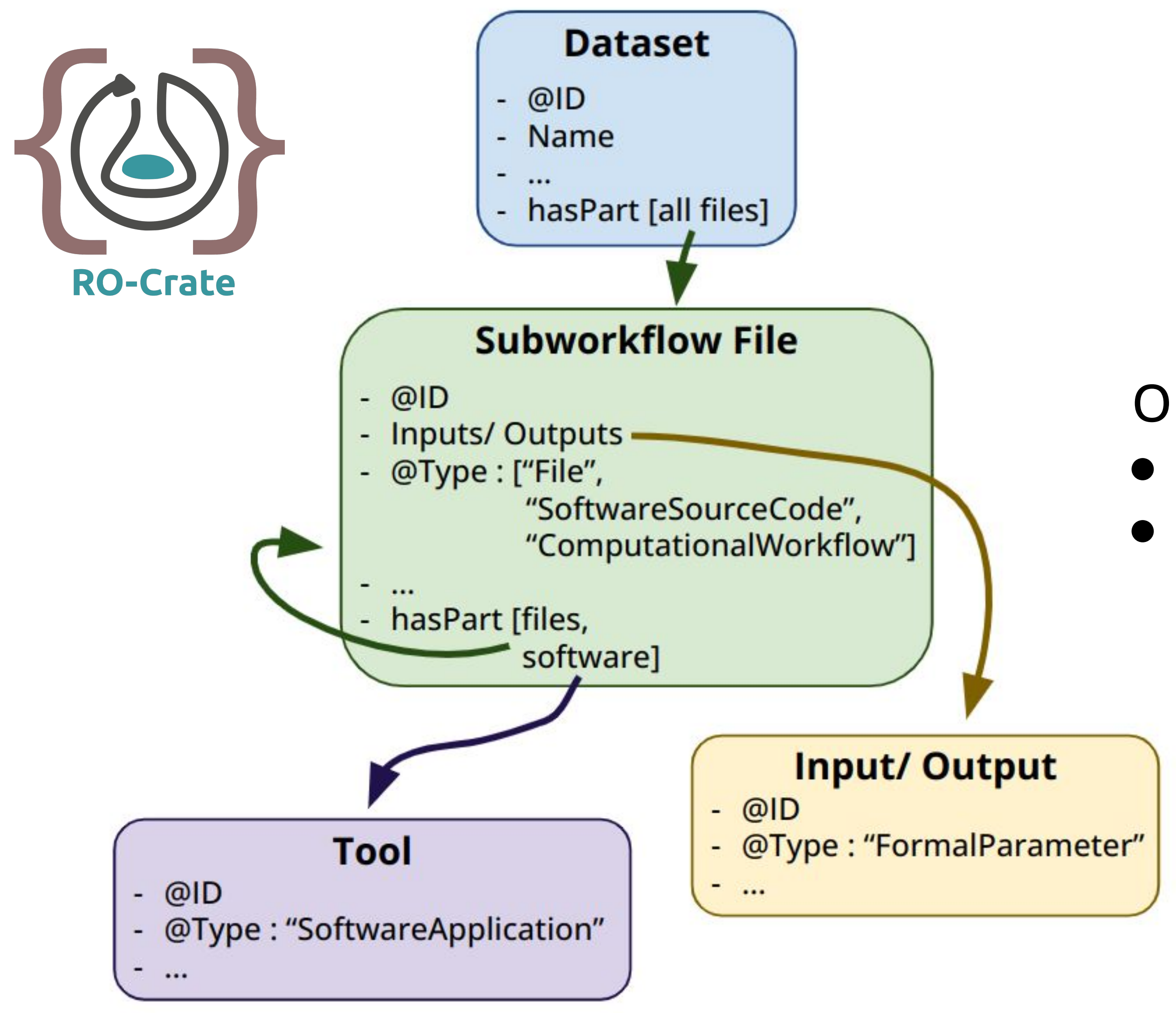


Figure 2 : Structure depicting the different steps of a Nextflow workflow that performs a RNA-Seq analysis.

Results

Current : Workflow RO-Crate profile 1.0



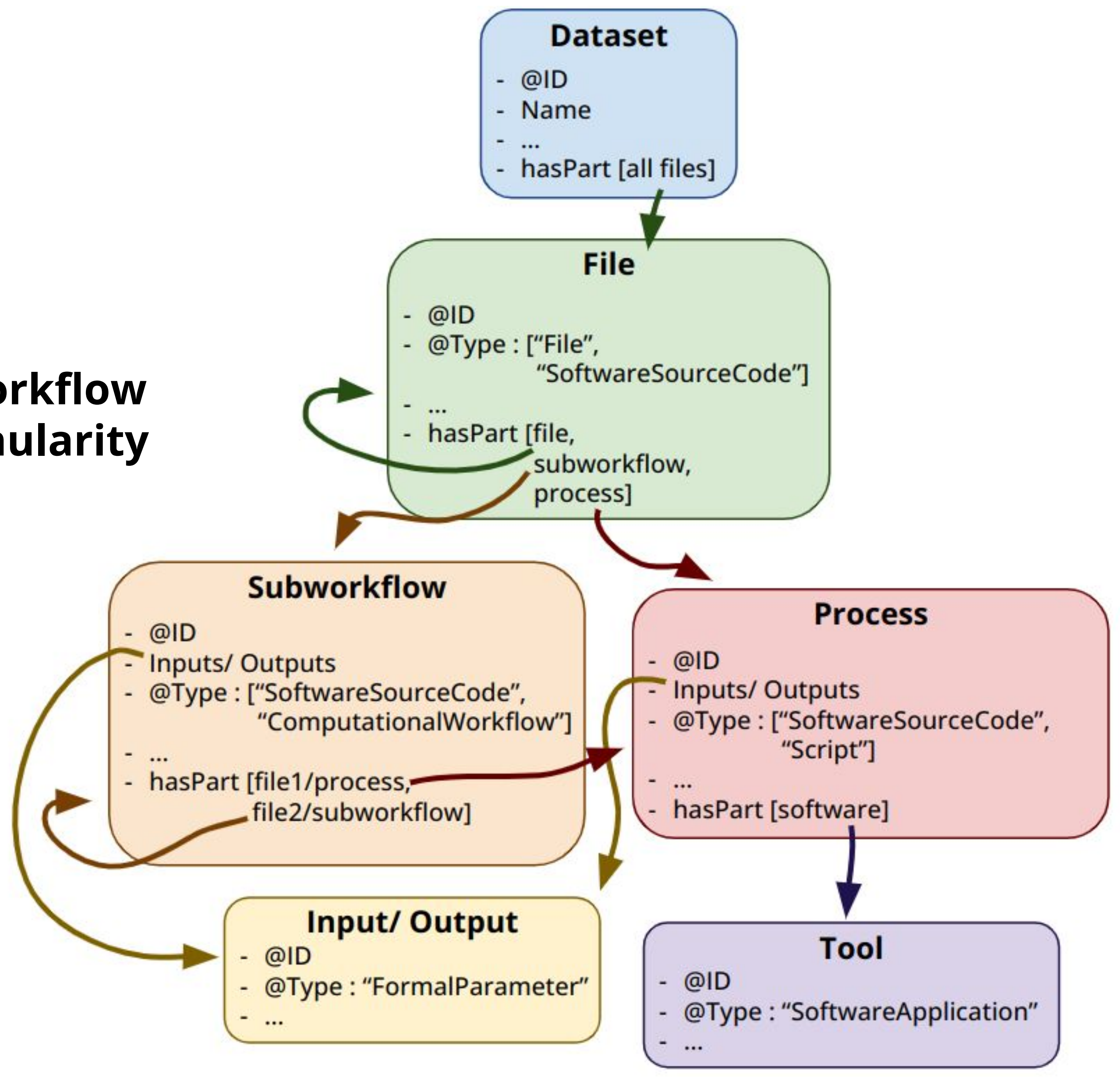
Limitations :

- **1 file is linked to 1 subworkflow**. There are no capabilities for defining multiple subworkflows within a single file.
- There is no **Process/Module granularity**.

Our proposition : Workflow RO-Crate profile 2.0

Our proposition :

- **Decouples files and subworkflow**
- Adds **Process/Module granularity**



Conclusion

In this work, we propose an extended profile for describing workflows using RO-Crate, with the new proposition:

- If a workflow can be described **using the original RO-Crate standard**, it can **also be described using our proposed extension**.
- Workflows implemented **Nextflow** and **Snakemake** can be fully described, at a **higher level of granularity**.

For the full **description and documentation** of the extended workflow RO-Crate profile, please refer to our GitHub repository (https://gitlab.liris.cnrs.fr/sharefair/posters/swat4hcls-2024 or scan the QR code).

[1] 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
[2] 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
[3] 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
[4] 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