8th World Conference on Research Integrity (Hybrid)
2-5 June 2024
Megaron Athens International Conference Centre (MAICC)
Athens - Greece

Proposing a New RO-Crate Profile for Enhanced Reproducibility in Computational Experiments

Eleni Adamidi[1], Panagiotis Deligiannis[1], Aikaterina Mastoraki[1], Thanasis Vergoulis[1]
[1]IMSI, "Athena" RC

Computational Reproducibility

Computational reproducibility refers to the ability to obtain consistent results using the same input data, computational methods, and conditions of analysis as were used in the original study [1]. It emphasizes on the importance of being able to exactly reproduce the outcomes of a computational study using the original raw data and code which is fundamental to the integrity and advancement of scientific research, fostering trust, verification, and collaboration among the scientific community.

Existing technologies, such as Git [2], Docker [3], and Jupyter Notebooks [4], enhance reproducibility by allowing consistent execution and understandable research logs. Scientific workflow technologies like Galaxy [5], Nextflow [6], and CWL [7] automate and manage complex processes, essential in fields like bioinformatics and environmental science, to ensure reproducibility and efficiency. These tools help standardize practices but challenges in standardized packaging and metadata documentation persist.

Research Object Crate

To address the challenges around standardized packaging of experiments and the detailed metadata they require, RO-Crates (Research Object Crates) [8] offer a promising solution. This framework is a lightweight approach to package research data with their associated metadata in a structured, machine-readable format. Utilizing JSON-LD web standards, RO-Crates encapsulate a variety of digital objects, from datasets to computational workflows, in portable containers. The core of each crate is the ro-crate-metadata.json file, which details the dataset and associated digital objects.

RO-Crate Profiles

RO-Crate profiles [9] specify the structure for these crates according to the needs of different research outputs or disciplines, ensuring consistency and tailoring to specific data requirements. These profiles help standardize data packaging and improve transparency and reproducibility in research. The profiles are a set of conventions, types and properties that one minimally can require and expect to be present in that subset of RO-Crates. The Workflow Run RO-Crate profile collection extends this system to capture detailed provenance of computational workflows, including execution specifics and environmental conditions, facilitating precise replication and verification of scientific experiments. This enhances the reproducibility of research and addresses common issues related to workflow portability and environment discrepancies.

Examples of RO-Crate profiles, each tailored to specific research needs, are:

- Workflow RO-Crate Profile: Documents computational workflows with related diagrams and abstract descriptions.
- Workflow Testing RO-Crate Profile: Extends Workflow RO-Crate for defining and documenting test suites for workflows.
- Workflow Run Crate Profile: Captures provenance and execution details of computational workflows and scripts.
- Common Provenance Model RO-Crate Profile: Aligns with W3C PROV to document distributed provenance chains.

The Need for Enhanced Metadata in Computational Research

While RO-Crates have proven effective in life sciences, their current profiles primarily capture the end results of research, often overlooking the computational specifics crucial in computer science disciplines. These disciplines frequently require detailed metadata about computational models, algorithms, and performance metrics, aspects critical for reproducibility but currently underrepresented in RO-Crate profiles. Metrics such as execution time, memory usage, and computational complexity are often crucial for evaluating the practical applicability of algorithms and software solutions. These details allow researchers to not only replicate results but also to understand the decision-making process behind model selection and optimization, which is often as critical as the results themselves.

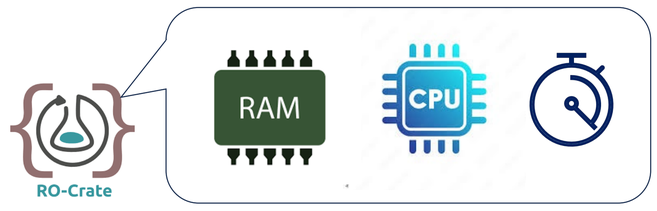
The Concept of a Workflow Run Performance Profile

To better serve the computer science community we propose extending the Workflow Run profile to include metadata fields specifically designed to capture information about computational efficiency details. This new profile is called Workflow Run Performance Profile.

The enhanced metadata will include the following fields:

1. "**memoryUsedPerTask**": A field that records the memory consumption for each individual task within the workflow during its execution. This metric should specify the amount of RAM utilized, measured in MB units.
2. "**cpuCoresUsedPerTask**": This field would document the number of CPU cores utilized by each task during its execution. It provides insight into the computational power required by each task, which is crucial for optimizing processing time and parallelization strategies.
3. "**executionTimePerTask**": A field to capture the duration of each task within the workflow. This will be recorded in milliseconds, to facilitate easy comparison and aggregation.

Documenting memory usage, CPU cores, and execution times for each task within a workflow enables researchers to pinpoint resource bottlenecks and optimize configurations, leading to more efficient scaling of workflows, especially in cloud environments where resource usage directly impacts costs.



Future RO-Crate profile enhancements

- **Incorporation of Detailed Metadata Fields**
  - Focus on tracking computational performance and environment across workflow stages.
  - Include detailed metrics on GPU usage
- **Integration of Machine Learning-Specific Metadata**
  - Document model explainability techniques, such as LIME [10] and SHAP [11], to illustrate decision-making processes in AI.
  - Aim to enhance transparency and understanding in machine learning workflows.
- **Collaboration for Wider Adoption**
  - Engage with researchers and institutions to refine the metadata extensions.
  - Seek feedback from scientific community stakeholders to ensure practical and widespread use.
  - Test the new extension in the context of the TIER2 Horizon European project [12].

References

[1] "Reproducibility and Replicability in Science," Reproducibility and Replicability in Science, pp. 1–234, Jan. 2019, doi: 10.17226/25303.
[2] J. D. Blischak, E. R. Davenport, and G. Wilson, "A Quick Introduction to Version Control with Git and GitHub," PLoS Comput Biol, vol. 12, no. 1, p. e1004668, 2016, doi: 10.1371/JOURNAL.PCBI.1004668.
[3] C. Boettiger, "An introduction to Docker for reproducible research," ACM SIGOPS Operating Systems Review, vol. 49, no. 1, pp. 71–79, Jan. 2015, doi: 10.1145/2723872.2723882.
[4] "Project Jupyter | Home." Accessed: Apr. 15, 2024. [Online]. Available: https://jupyter.org/
[5] "Creating Workflows and Advanced Workflow Options - Galaxy Community Hub." Accessed: Apr. 15, 2024. [Online].
[6] P. DI Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame, "Nextflow enables reproducible computational workflows," Nature Biotechnology 2017 35:4, vol. 35, no. 4, pp. 316–319, Apr. 2017, doi: 10.1038/nbt.3820.
[7] "Home | Common Workflow Language (CWL)." Accessed: Apr. 12, 2024. [Online]. Available: https://www.commonwl.org/
[8] P. Sefton et al., "RO-Crate Metadata Specification 1.1," Oct. 2020, doi: 10.5281/ZENODO.4031327.
[9] "Profiles | Research Object Crate (RO-Crate)." Accessed: Apr. 15, 2024. [Online]. Available: https://www.researchobject.org/ro-crate/profiles.html
[10] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' Explaining the predictions of any classifier," Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol. 13-17-August-2016, pp. 1135–1144, Aug. 2016, doi: 10.1145/2939672.2939778.
[11] S. M. Lundberg, P. G. Allen, and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions", doi: 10.5555/3295222.3295230.
[12] "TIER2." https://tier2-project.eu/ (accessed May 7, 2024).

Contact
Thanasis Vergoulis
Researcher, IMSI, ARC
vergoulis@athenarc.gr

Eleni Adamidi
Researcher, IMSI, ARC
eleni.adamidi@athenarc.gr

ATHENA Research & Innovation Information Technologies

INSTITUTE for the MANAGEMENT of INFORMATION SYSTEMS

@SchemaServices

TIER2 receives funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101094817.
Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the EU nor the EC can be held responsible for them.