Improving reproducibility of computational analyses performed on the ELIXIR-GR cloud

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

ELIXIR All Hands , 5-8 June 2022, Dublin, Ireland

HYPATIA is the Cloud infrastructure that has been developed to support the computational needs of the ELIXIR-GR community. It leverages **SCHEMA**, an open-source platform developed by members of the ELIXIR-GR community, to offer to its users a functionality to run on-demand computational analyses on the underlying, heterogeneous cluster. One key feature of SCHEMA is that it exploits containerization (e.g., Docker), experiment packaging (e.g., RO-Crates), and workflow management (e.g., CWL) technologies to facilitate reproducibility of the respective analyses. In this poster, we present this functionality and we elaborate on our plans to extend it in the context of TIER2, a new Horizon Europe project for research reproducibility into which we participate.

HYPATIA Cloud Infrastructure

ELIXIR-GR services and resources

HYPATIA [1] is the Cloud infrastructure that has been developed to support the computational needs of the ELIXIR-GR community, but also the broader community of life scientists in Greece and abroad. HYPATIA consists of a powerful computational cluster of heterogeneous physical machines. Currently, the cluster is comprised of:

- 32 basic nodes: (2 CPUs, 14 cores/CPU, 512GB DDR4 RAM).
- 2 hefty nodes: (2 CPUs, 24 cores/CPU, 1TB DDR4 RAM)
- 3 GPU nodes: (2 CPUs, 14 cores/CPU, 768GB DDR4 RAM, 2 GPUs)
- 8 I/O nodes: (2 CPUs, 14 cores/CPU, 512GB DDR4 RAM, 2x2TB SSD 6G)
- 9 infrastructure nodes:(2 CPUs, 14 cores/CPU, 192GB DDR4 RAM)

SCHEMA Platform

Scheduling Scientific Containers on a Cluster of Heterogeneous Machines

HYPATIA leverages SCHEMA [2], an open-source platform developed by members of the ELIXIR-GR community, that facilitates the execution and reproducibility of computational analysis on heterogeneous clusters, leveraging containerization, experiment packaging, workflow management, and machine learning technologies [3].

SCHEMA Functionalities

SCHEMA implements a wide range of functionalities to assist scientists in the data-driven and reproducible science era. Most notable are (a) the option to upload custom-made scientific containers or container-based workflows, (b) a wizard and an API that facilitate the execution of individual containers or workflows, (c) a monitor that informs the users about the consumption of computational resources, (d) a wizard to transform executed analyses into RO-crate-based "experiment packages", and (e) a wizard to facilitate interconnection with open data repository services [3].

Automatic creation of RO-crates from executions

A Research Object (RO) combines the ability to bundle multiple types of artefacts together, such as spreadsheets, code, examples, and figures [4]. RO-Crate (Research Object Crate) is a method of organizing file-based data with associated metadata, using linked data principles, in both human and machine-readable formats, with the ability to include additional domain-specific metadata.

The core of RO-Crate is a JSON-LD (JSON for Linked Data) file that contains structured metadata about the dataset as a whole (the Root Data Entity) and, optionally, about some or all of its files. This provides a simple way to assert the authors (e.g. people, organizations) of the RO-Crate, or to capture more complex provenance for files, such as how they were created using software and equipment [4].

One key feature of SCHEMA is that it exploits containerization (e.g., Docker), experiment packaging (e.g., RO-Crates), and workflow management (e.g., CWL) technologies to facilitate reproducibility of the respective analyses.

More specifically, each workflow execution can be easily transformed into an RO-crate object that incorporates all the metadata that are required for it to be re-executed (i.e., the location of the container images involved, the software configuration used, the respective input and output data, etc.), (see Fig. 1).

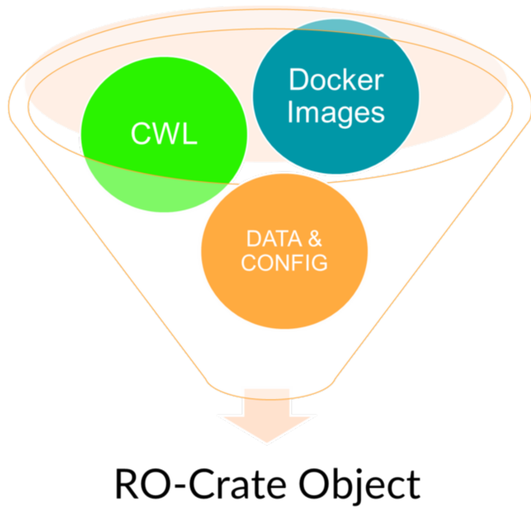

Fig. 1. SCHEMA packaging research artefacts

Future Research

TIER2 project

Reproducibility is often claimed as a central principle of the scientific method referring to the possibility for the scientific community to obtain the same results as the originators of a specific finding [5]. Poor reproducibility is identified in many research areas, particularly in computationally intensive domains where results rely on a series of complex methodological decisions that are not well captured by traditional publication approaches [6].

TIER2 is a new Horizon European project aiming to increase reproducibility of scientific research results that will bring **t**rust, **i**ntegrity, and **e**fficiency to the European Research Area (ERA) and the global Research and Innovation (R&I) system [7].

In the context of TIER2, we aim to adapt & extend SCHEMA to further facilitate data/code reproducibility in life sciences, computer sciences and social sciences. More specifically, we plan to (a) examine alternative representations of reproducible objects (various RO-crate templates, other choices like ReproZip, Packrat, etc.), (b) improve the accessibility and searchability of these objects, (c) simplify the execution of these objects and (d) investigate domain-specific aspects.

References

[1] "HYPATIA." https://hypatia.athenarc.gr/ (accessed May 5, 2023).
[2] "SCHEMA." https://schema.athenarc.gr/about/ (accessed May 5, 2023).
[3] Thanasis Vergoulis, Konstantinos Zagganas, Loukas Kavouras, Martin Reczko, Stelios Sartzetakis, and Theodore Dalamagas. "SCHeMa: Scheduling Scientific Containers on a Cluster of Heterogeneous Machines." arXiv preprint arXiv:2103.13138 (2021).
[4] S. Peroni et al., "Packaging research artefacts with RO-Crate," Data Sci., vol. 5, no. 2, pp. 97–138, 2022, doi: 10.3233/DS-210053.
[5] K. Popper, The Logic of Scientific Discovery. Routledge, 2005.
[6] B. Grüning et al., "Practical Computational Reproducibility in the Life Sciences," Cell Syst., vol. 6, no. 6, pp. 631–635, Jun. 2018, doi: 10.1016/j.cels.2018.03.014.
[7] "TIER2." https://tier2-project.eu/ (accessed May 6, 2023).

Contact

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

HYPATIA has been funded by the "ELIXIR-GR: Managing and Analysing Life Sciences Data (MIS: 5002780)" project (co-funded by Greece and the EU - European Regional Development Fund)

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.

These projects have received funding from the European Union's Horizon 2020 research and innovation programme

@ELIXIREurope /company/elixir-europe