bio³

Biomedicine, Bioinformatics & Biotechnology Forum:
Fostering Collaboration in Industry & Academia

Friday 15 - Sunday 17 September 2023

Technopolis, Gazi - Athens, Greece

Containerization solutions for biomedical informatics

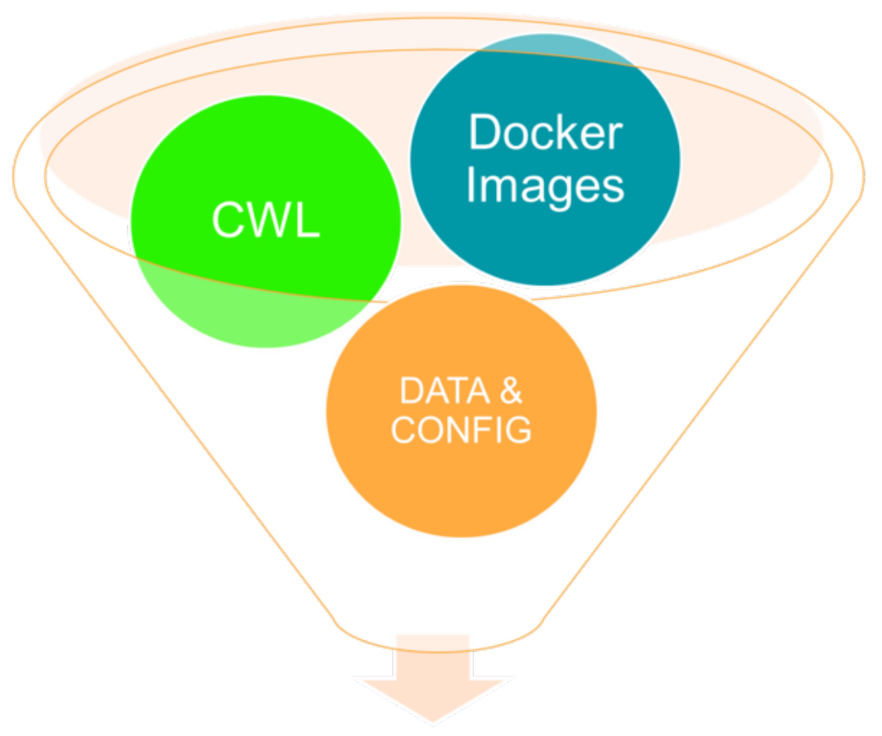
Eleni Adamidi[1], Panagiotis Deligiannis[1], Aikaterina Mastoraki[1], Thanasis Vergoulis[1]
[1]IMSI, "Athena" RC

SCHEMA Platform

Scheduling Scientific Containers on a Cluster of Heterogeneous Machines

In the rapidly evolving field of biomedical informatics the demand for **robust containerization technologies** has become increasingly evident, representing a crucial tool to address the challenges faced by **life scientists** and make tangible strides in the **biomedical industry**. This poster delves into the containerization solutions and their significance in the life science research and the biomedical industry. Specifically, we present SCHEMA [1], an open-source platform developed by members of the ELIXIR-GR community [2] which harnesses the power of containerization technologies to provide a comprehensive solution that advances scientific endeavors in precision medicine, medical imaging, bioinformatics, and more. SCHEMA empowers life scientists with containerization tools, offering them the means to create, execute, and manage computational data-analysis workflows efficiently using the Common Workflow Language (CWL) [3]. These workflows can be later easily executed on diverse platforms, promoting accessibility and reproducibility of computational experiments.

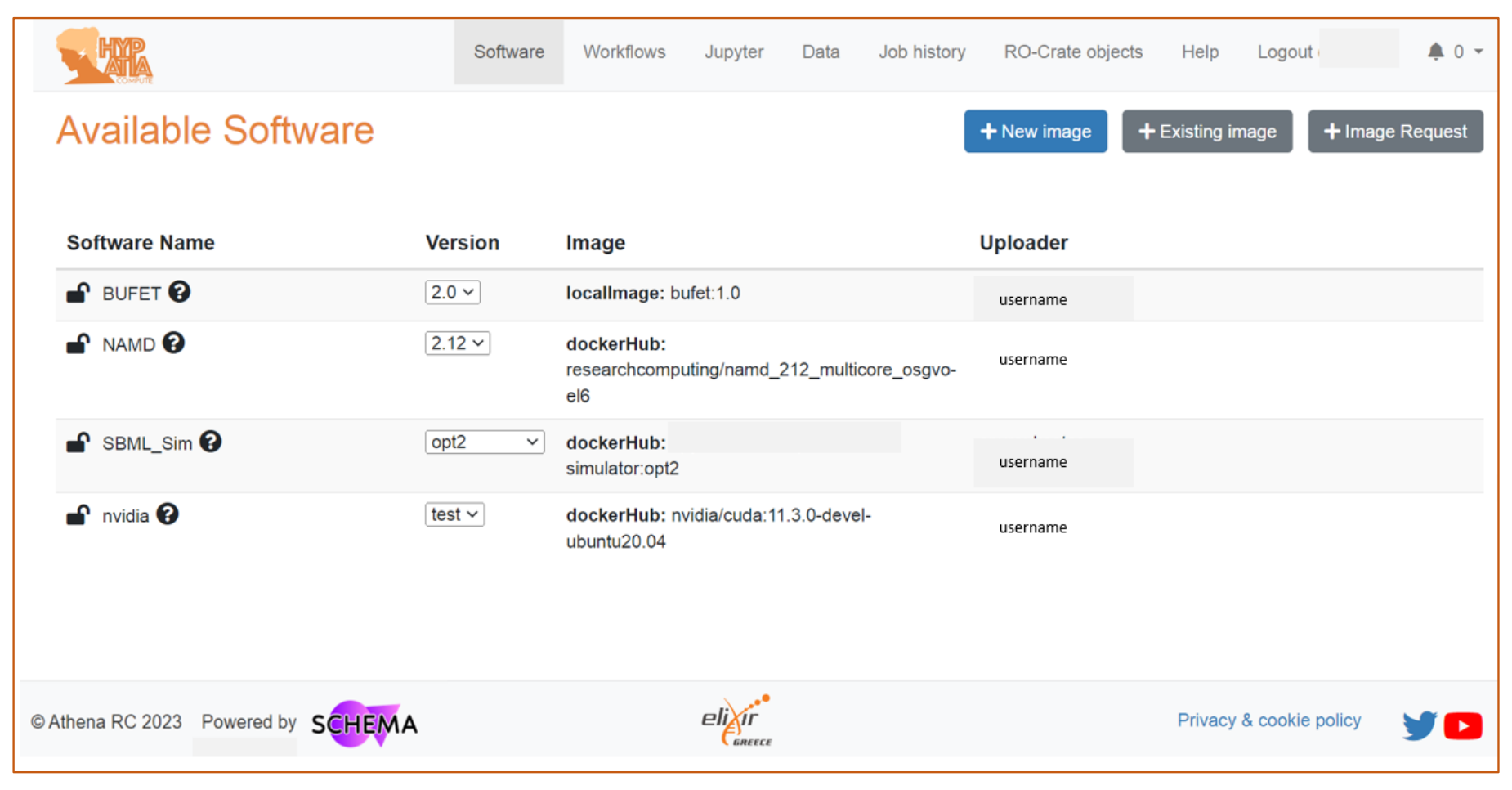
More specifically, each workflow execution can be easily transformed into an **RO-crate object** [4] that incorporates all the metadata that are required for it to be re-executed (i.e., the location of the container images involved, the software configuration used, the respective input and output data, etc.), (see Fig. 1).



RO-Crate Object
Fig. 1. Packaging research artefacts

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].



Available Software

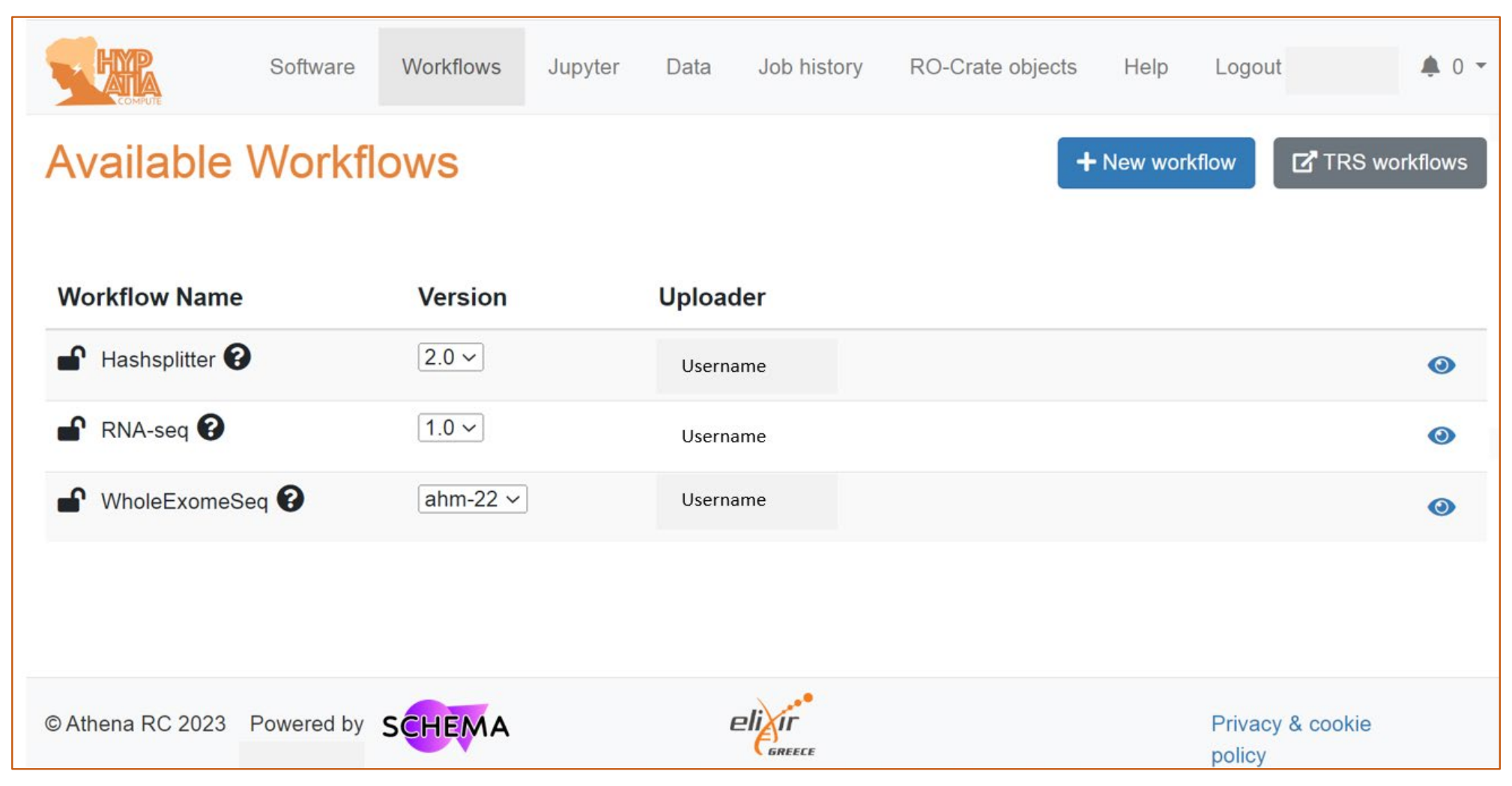
Software Name	Version	Image	Uploader
BUFET	2.0	**localImage:** bufet:1.0	username
NAMD	2.12	**dockerHub:** researchcomputing/namd_212_multicore_osgvo-el6	username
SBML_Sim	opt2	**dockerHub:** simulator:opt2	username
nvidia	test	**dockerHub:** nvidia/cuda:11.3.0-devel-ubuntu20.04	username

Fig. 2. SCHEMA Software selection

On-demand batch computations projects can be created though the SCHEMA platform. This type of project is suitable for batches of computational tasks to be executed. Each computational task could involve the execution of a particular software product or of a workflow that combines many software products (see Fig. 2).

SCHEMA platform users have access to a comprehensive set of functionalities designed to streamline the workflow management process (see Fig. 3). These functionalities include the ability to add a new workflow, specifying its key attributes such as Workflow Type, name, version, and description. Additionally, users can provide important instructions, making it easier for collaborators to understand and execute the workflow effectively. There is also the option to control visibility, allowing users to choose whether the workflow is accessible to everyone or restricted to specific collaborators.

For researchers involved in COVID-19 studies, the platform offers the option to indicate the workflow's relevance to COVID-19 research, fostering the sharing of important data. Furthermore, users can provide links to bio.tools and add relevant DOIs (Digital Object Identifiers) for enhanced traceability. Lastly, the platform allows users to upload their CWL workflow definition file, facilitating the execution and reproducibility of computational analyses. These functionalities collectively empower researchers to manage their workflows efficiently and collaborate seamlessly within the SCHEMA platform.



Available Workflows

Workflow Name	Version	Uploader
Hashsplitter	2.0	Username
RNA-seq	1.0	Username
WholeExomeSeq	ahm-22	Username

Fig. 3. SCHEMA Workflow selection

Containerization strategies, consulting, and solutions for biomedical informatics

Presently, SCHEMA has been included as one of the services within the **European Digital Innovation Hub for Smart Health: Precision Medicine and Innovative E-health Services (smartHEALTH)** [5], aiming to support digital transformation in biomedical research, enabling SMEs, startups, mid-caps, and the public sector to leverage digital technology effectively. In this context, SCHEMA's team if offering containerization solutions and consulting to members of the biomedical industry paving the way for cutting-edge biomedical informatics.

The **consulting services** are focusing on providing support for the development of computational data-analysis workflows in CWL. The workflows that will be developed with our support can be used in SCHEMA platform to enable the easy software execution and the reproducibility of the computational experiments. **Training activities** will educate the audience about the notion of containers and containerization, the advantages of such technologies, the relevant existing best practices, any common pitfalls, and the latest technological advancements in the field. The aim is that the trainees will eventually be able to use containerization strategies to produce tools that can be easily executed on different platforms and that can be effortlessly incorporated in biomedical pipelines.

Future platform extensions

As the biomedical landscape continues to evolve, SCHEMA envisions a more impactful role. Plans are underway to extend the platform with **machine learning** technologies, to enhance how life scientists approach complex problems. This expansion will open new opportunities for innovation within the biomedical industry, allowing for more sophisticated data analysis, predictive modeling, and insights that can advance research and applications in practical and meaningful ways. Moreover, in the context of TIER2 Horizon European project [6], we aim to adapt & extend SCHEMA to further facilitate data/code reproducibility in life sciences, computer sciences and social sciences.

References

[1] "SCHEMA." https://schema.athenarc.gr/about/ (accessed July 5, 2023).
[2] "ELIXIR- Greece" https://www.elixir-greece.org/ (accessed July 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] "smartHEALTH" https://smarthealth-edih.eu/en/ (accessed July 31, 2023).
[6] "TIER2." https://tier2-project.eu/ (accessed May 6, 2023).

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

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

Co-funded by the European Union

ΕΣΠΑ 2021-2027
Partnership Agreement 2021-2027

Co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

smartHEALTH EDIH

@SchemaServices

ATHENA Research & Innovation Information Technologies

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

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.