EOSC-Life Workflow Collaboratory for the Life Sciences

Carole Goble, Finn Bacall, Stian Soiland-Reyes, Stuart Owen, Ignacio Eguinoa, Bert Droesbeke, Hervé Ménager, Laura Rodriguez-Navas, José Fernández, Björn Grüning, et al.

1st Conference on Research Data Infrastructure

Life Sciences

https://doi.org/10.52825/CoRDI.v1i.352

Published: 07 Sept. 2023

The EOSC-Life WORKFLOW COLLABORATORY for the LIFE SCIENCES

Carole Goble[1][https://orcid.org/0000-0003-1219-2137], Finn Bacall[1][https://orcid.org/0000-0002-0048-3300], Stian Soiland-Reyes[1][https://orcid.org/0000-0001-9842-9718], Stuart Owen[1][https://orcid.org/0000-0003-2130-0865], Ignacio Eguinoa[2][https://orcid.org/0000-0002-6190-122X], Bert Droesbeke[2][https://orcid.org/0000-0003-0522-5674], Hervé Ménager[3][https://orcid.org/0000-0002-7552-1009], Laura Rodriguez-Navas[4][https://orcid.org/0000-0003-4929-1219], José M. Fernández[4][https://orcid.org/0000-0002-4806-5140], Björn Grüning[5][https://orcid.org/0000-0002-3079-6586], Simone Leo[6][https://orcid.org/0000-0001-8271-5429], Luca Pireddu[6][https://orcid.org/0000-0002-4663-5613], Michael R. Crusoe[7][https://orcid.org/0000-0002-2961-9670], Johan Gustafsson[8][https://orcid.org/0000-0002-2977-5032], Salvador Capella-Gutierrez[4][https://orcid.org/0000-0002-0309-604X], and Frederik Coppens[2] [https://orcid.org/0000-0001-6565-5145]

[1] The University of Manchester, UK

[2] VIB-UGent Center for Plant Systems Biology, Belgium

[3] Institut Pasteur, Paris

[4] Barcelona Supercomputing Center, Spain

[5] Albert-Ludwigs-University Freiburg, Germany

[6] Center for Advanced Studies, Research and Development in Sardinia, Italy

[7] Common Workflow Language & VU Amsterdam, Netherlands

[8] Australian BioCommons, Australia

Abstract. Workflows have become a major tool for the processing of Research Data, for example, data collection and data cleaning pipelines, data analytics, and data update feeds populating public archives. The EOSC-Life Research Infrastructure Cluster project brought together Europe's Life Science Research Infrastructures to create an Open, Digital and Collaborative space for biological and medical research to develop a cloud-based Workflow Collaboratory. As adopting FAIR practices extends beyond data, the Workflow Collaboratory drives the implementation of FAIR computational workflows and tools. It fosters tool-focused collaborations and reuse via the sharing of data analysis workflows and offers an ecosystem of services for researchers and workflow specialists to find, use and reuse workflows. It's web-friendly Digital Object Metadata Framework, based on RO-Crate and Bioschemas, supports the description and exchange of workflows across the services.

Keywords: Fair Data; Computational Workflows; Digital Objects; Data Intensive Bio-Science; Fair Workflows; Fair Software

1. Background

Performing computational data processing using workflows has taken hold in the biosciences as the discipline becomes increasingly computational. Adopting FAIR practices extends beyond data to include workflows and the tools they use. The COVID-19 pandemic highlighted the importance of systematic and shared analysis of SARS-CoV-2 data processing and surveillance pipelines, data analytics at scale, and the reproducibility of computational processes [1].

The EOSC-Life Research Infrastructure Cluster project brought together Europe's Life Science Research Infrastructures to create an open, digital and collaborative space for biological and medical research. The Research Infrastructures, range from biobanking and clinical trials to plant phenotyping and bioimaging. A major development by EOSC-Life has been a cloud-based Workflow Collaboratory to drive the implementation of FAIR computational workflows [2] and foster tool-focused collaborations and reuse via the sharing of data analysis workflows.

1.1 The EOSC-Life Workflow Collaboratory Services

The Workflow Collaboratory offers an ecosystem of services for researchers and workflow specialists to find, use and reuse workflows, and deploy them using European Open Science Cloud (EOSC) infrastructure (Figure 1). The heterogeneity of the Research Infrastructures is reflected in the diversity of their data analysis practices and the variety of workflow management systems they use, including specialist platforms.

Workflow Managers and Execution systems include pre-existing and emerging workflow management systems. Currently 14 different workflow platforms are represented including Jupyter Notebooks [3] and Python scripts, general systems (e.g. Nextflow [4], Snakemake[5], Galaxy[6], CWL[7]) and specialist systems (e.g. SCIPION). Workflow execution platforms include Galaxy Europe and back-end services Sapporo and WfExS that execute workflows in different languages (e.g. CWL, Nextflow, snakemake) through a common interface. WfExS handles sensitive data securely.

Community Workflow Repositories both pre-existing and emerging, typically use Git, GitLab or GitHub. Curated collections include Galaxy's Intergalactic Workflow Commission and Nextflow's nf-core; others include project focused repositories.

Registries provide one stop to find and share containers (BioContainers [8]), tools (bio.tools [9]) and workflows (WorkflowHub [10]), and support FAIRness through rich metadata and inter-registration integration. WorkflowHub is system agnostic, with dedicated support for popular management systems. Galaxy and CWL workflows, and entries on WorkflowHub, are annotated with tool identifiers to link through to bio.tools entries, and bio.tools links to workflows that use a given tool.

Testing and Benchmarking services support the usability as well as reusability of tools and workflows as software [11]. LifeMonitor [10] monitors and triggers automated workflow tests and automated checks on metadata, and adherence to best practices on the workflow's source code Git repository. OpenEBench benchmarks tools, and monitors software quality as well as scientific benchmarking to help determine the precision, recall and other metrics of bioinformatics resources in unbiased scenarios.

Research Information System services plug into the wider services of the EOSC and scholarly communication to support publication, citation, and knowledge discovery. WorkflowHub's integration DataCite mints DOIs for workflows publication, and its integration with ORCID and citation.cff file format supports workflow author credit and citation. The WorkflowHub contributes to the DataCite PID Graph and OpenAIRE Research Graph.

Infrastructure services range from AAI (LS-Login[1]) to cloud and cluster compute systems such as Galaxy's PULSAR network[2].

[1] https://lifescience-ri.eu/ls-login/

[2] https://pulsar-network.readthedocs.io/en/latest/

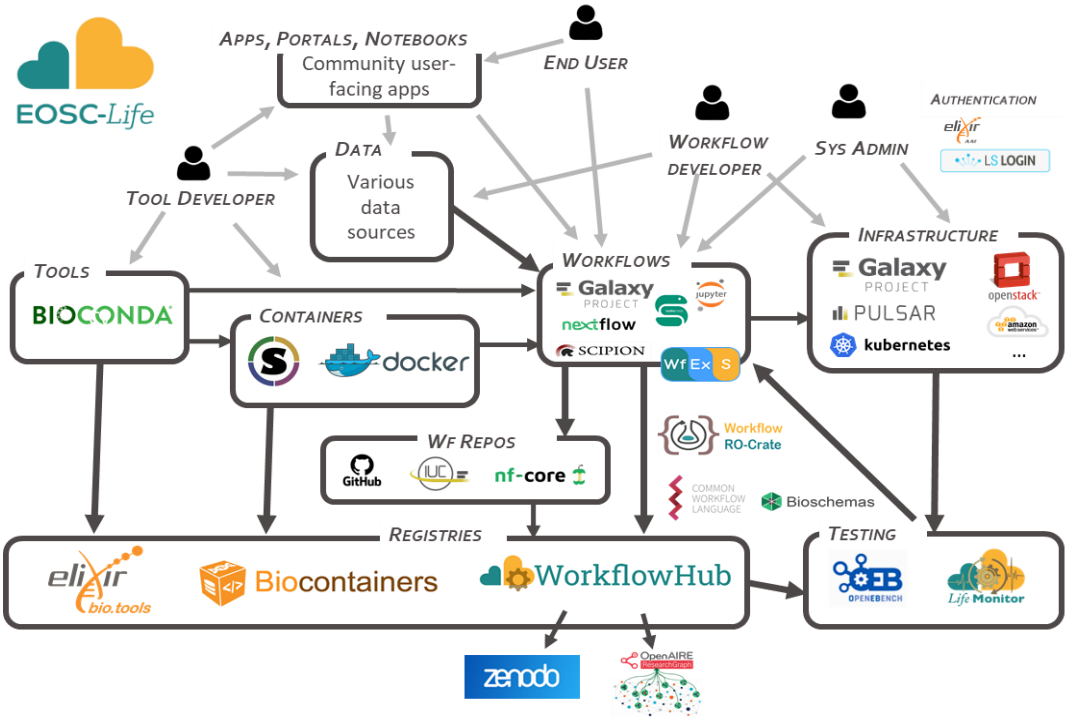

Figure 1. The EOSC-Life Workflow Collaboratory.

1.2 The Digital Object based Metadata Framework

A web-friendly metadata framework has been developed to support the description and exchange of workflows across the services.

Bioschemas [12] schema.org profiles for Computational Tool, Computational Workflow and Formal Parameter provide metadata about a workflow and its tools that are discipline independent, despite the "bio" prefix. The **EDAM Ontology** [13] adds informatics-specific metadata, such as strong typing of inputs and outputs and describes the overall workflow topics and operations to help find workflows.

The **Common Workflow Language** [7] is encouraged as a canonical workflow description to accompany the native workflow definitions when registering in the WorkflowHub. CWL represents the structure and steps of workflows in an interoperable way across workflow languages. Several workflow systems, such as Galaxy, support Abstract CWL for this purpose, though they are still executed using their native language on their native platforms.

RO-Crate [14], a community-developed standardised approach for FAIR Digital Objects [15], packages executable workflows, their components (e.g. example and test data), abstract CWL, diagrams and their metadata. RO-Crate makes workflows more readily re-usable, acting as the unit of currency of exchange between the services, recording the provenance of workflow runs and a format for archiving in public repositories such as Zenodo.

The GA4GH Tools Registry Service API supports the exchange of scientific tools and workflows and enables users to search for and retrieve metadata about registered tools, so that workflow execution platforms can search and import workflows from WorkflowHub and WorkflowHub can directly launch workflows on their platforms. Multi-language execution services such as Sapporo use the GA4GH Workflow Execution Service API.

2. Outlook

At the time of writing over 380 workflows have been registered in WorkflowHub from over 170 workflow teams. Adoption of the services has extended beyond Life Sciences to adoption by communities working in biodiversity, astronomy, astro-physics and climate change. These services continue to be supported by a new portfolio of Horizon Europe and national projects, and are sponsored by the European Life Science Research Infrastructures, ELIXIR, EuroBioImaging-ERIC, BBMRI-ERIC, and EU-IBISBA for their long-term sustainability. Many of the services are registered in the EOSC Service Catalogue and Marketplace[3] and have been adopted outside Europe by the Australian BioCommons[4].

Data availability statement

All workflows, tools and content are openly available. Software is open source. Standards are open. Available from the following: WorkflowHub: https://workflowhub.eu/; BioContainers: https://biocontainers.pro/; Bio.tools: https://bio.tools/; LifeMonitor: https://www.lifemonitor.eu/; OpenEBench: https://openebench.bsc.es/; Galaxy Europe: https://usegalaxy.eu/; Sapporo https://github.com/sapporo-wes/sapporo; WfExS: https://github.com/inab/WfExS-backend; Bioschemas: https://bioschemas.org; Common Workflow Language: https://www.commonwl.org/; EDAM Ontology: https://edamontology.org/page; RO-Crate: https://www.researchobject.org/ro-crate/; GA4GH Tools Research Service API: https://ga4gh.github.io/tool-registry-service-schemas/; GA4GH Workflow Execution Service API: https://ga4gh.github.io/workflow-execution-service-schemas/docs/

Underlying and related material

None

Author contributions

CG wrote the abstract and supervised the work. FB, SS-R, SO, IE, BD, HM, LR-N, JMF, BJ, SL, JG undertook the work.SC-G, FC supervised the work.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was funded by the European Union programmes Horizon 2020 under grant agreements H2020-INFRAEDI-02-2018 823830 (BioExcel-2), H2020-INFRAEOSC-2018-2 824087 (EOSC-Life), and Australian BioCommons which is enabled by NCRIS via Bioplatforms Australia funding.

[3] https://marketplace.eosc-portal.eu/

[4] https://australianbiocommons.github.io/

Acknowledgement

We acknowledge the WorkflowHub Club (https://about.workflowhub.eu/project/community/) and the Galaxy community (https://galaxyproject.org/community/).

References

1. T. Reiter, P.T. Brooks, L. Irber, S.E.K. Joslin, C.M. Reid, C. Scott, C.T. Brown, N.T. Pierce-Ward, "Streamlining data-intensive biology with workflow systems", GigaScience, vol.10, no.1, pp:1-19, January 2021, https://doi.org/10.1093/gigascience/giaa140
2. C. Goble, S. Cohen-Boulakia, S. Soiland-Reyes, D. Garijo, Y. Gil, M.R. Crusoe, K. Peters, D. Schober, "FAIR Computational Workflows. Data Intelligence" vol.2, no.1, pp:108–121, 2020, https://doi.org/10.1162/dint_a_00033
3. T. Kluyver, B. Ragan-Kelley, F. Pérez, B.E. Granger, M. Bussonnier, J. Frederic, K. Kelley, J.B. Hamrick, J. Grout, S. Corlay et al "Jupyter notebooks—a publishing format for reproducible computational workflows" In F Loizides, B Scmidt (eds) International conference on electronic publishing. IOS Press, ELPUB, Göttingen, 2016, pp:87–90
4. P. Di Tommaso, M. Chatzou, E. Floden, P.P. Barja, E. Palumbo, C. Notredame, "Nextflow enables reproducible computational workflows". Nat Biotechnol vol.35, pp:316–319, 2017, https://doi.org/10.1038/nbt.3820
5. J. Köster, S. Rahmann, "Snakemake—a scalable bioinformatics workflow engine", Bioinformatics, vol.28, no.19, pp:2520–2522, October 2012, https://doi.org/10.1093/bioinformatics/bts480
6. E Afgan, D. Baker, B Batut, et al. (2018) "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update", Nucleic Acids Research, vol.46, pp:W537–W544, 2018, https://doi.org/10.1093/nar/gky379
7. M.R. Crusoe, S. Abeln, A. Iosup, P. Amstutz, J. Chilton, N. Tijanić, H. Ménager, S. Soiland-Reyes, B. Gavrilović, C. Goble, "The CWL Community Methods Included: Standardizing Computational Reuse and Portability with the Common Workflow Language", CACM, vol.65, no.6, pp:54-63 June 2022, https://doi.org/10.1145/3486897
8. F. da Veiga Leprevost et al, BioContainers: an open-source and community-driven framework for software standardization, *Bioinformatics*, vol.33, no.16, pp: 2580–2582, August 2017, https://doi.org/10.1093/bioinformatics/btx192
9. J. Ison, et al. Tools and data services registry: a community effort to document bioinformatics resources. Nucleic Acids Research. 2015, vol.44, no.D1, pp:D38–D47January 2016, https://doi.org/10.1093/nar/gkv1116
10. C. Goble, S. Soiland-Reyes, F. Bacall, S. Owen, L. Pireddu, S. Leo. EOSC-Life Implementation of a mechanism for publishing and sharing workflows across instances of the environment. 2023, Zenodo. https://doi.org/10.5281/zenodo.7886545
11. M. Barker, N.P. Chue Hong, D.S. Katz, A-L. Lamprecht, C. Martinez-Ortiz, F. Psomopoulos, J. Harrow, L.J. Castro, M. Gruenpeter, P. Andrea Martinez, T. Honeyman. "Introducing the FAIR Principles for research software". Sci Data 9, vol.622, 2022, https://doi.org/10.1038/s41597-022-01710-x
12. A. Gray, L.J. Castro, N. Juty, C. Goble "Schema.org for Scientific Data" in A. Choudhary, G. Fox, T. Hey (eds) Artificial Intelligence for Science, pp:495-514, 2023, https://doi.org/10.1142/9789811265679_0027
13. J. Ison, M. Kalas, I. Jonassen, D. Bolser, M. Uludag, H. McWilliam, J. Malone, R. Lopez, S. Pettifer, P. Rice, "EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats", Bioinformatics, vol.29, no.10, pp:1325-32, May 2013 https://doi.org/10.1093/bioinformatics/btt113
14. S. Soiland-Reyes, P. Sefton, M. Crosas, L.J. Castro, F. Coppens, J.M. Fernández, D. Garijo, B. Grüning, M. La Rosa, S. Leo, E. Ó Carragáin, M. Portier, A. Trisovic, RO-Crate Community, P. Groth, C. Goble "Packaging Research Artefacts with RO-Crate", Data Science, vol.5, no.2, pp: 97 – 138. 2022, https://doi.org/: 10.3233/DS-210053

15. S. Soiland-Reyes, P. Sefton, L.J. Castro, F. Coppens, D. Garijo, S. Leo, M. Portier, P. Groth, “Creating lightweight FAIR Digital Objects with RO-Crate”, Research Ideas and Outcomes vol.8, no.e93937, 2022, https://doi.org/10.3897/rio.8.e93937