Implementing FAIR Digital Objects in the EOSC-Life Workflow Collaboratory

Extended abstract for Special Issue of DI on CWFR

Carole Goble[1], Stian Soiland-Reyes[1,2], Finn Bacall[1], Stuart Owen[1], Alan Williams[1], Ignacio Eguinoa[3,4], Bert Droesbeke[3,4], Simone Leo[6], Luca Pireddu[6], Laura Rodriguez-Navas[7], José Mª Fernández[7], Salvador Capella-Gutierrez[7], Hervé Ménager[8], Björn Grüning[9], Beatriz Serrano-Solano[9], Philip Ewels[5], Frederik Coppens[3,4]

[1]Department of Computer Science, The University of Manchester, UK
[2]Informatics Institute, University of Amsterdam, The Netherlands
[3]Department of Plant Biotechnology and Bioinformatics, Ghent University, 9052 Ghent, Belgium
[4]VIB Center for Plant Systems Biology, 9052 Ghent, Belgium
[5]Science for Life Laboratory (SciLifeLab), Dept. of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden
[6]Center for Advanced Studies, Research and Development in Sardinia (CRS4), Pula (CA), Italy
[7]Life Sciences Department. Barcelona Supercomputing Center (BSC), Barcelona, Spain.
[8]Institut Pasteur, Paris, France
[9]Bioinformatics Group, University of Freiburg, Germany

The practice of performing computational processes using workflows has taken hold in the biosciences as the discipline becomes increasingly computational [1]. The COVID-19 pandemic has spotlighted the importance of systematic and shared analysis of SARS-CoV-2 and its data processing pipelines [2]. This is coupled with a drive in the community towards adopting FAIR practices (Findable, Accessible, Interoperable, and Reusable) not just for data, but also for workflows [3], and to improve the reproducibility of processes, both manual and computational.

EOSC-Life brings together 13 of the Life Science 'ESFRI' research infrastructures to create an open, digital and collaborative space for biological and medical research [4]. The project is developing a cloud-based **workflow collaboratory** to drive implementation of FAIR workflows across disciplines and RI boundaries, and foster tool-focused collaborations and reuse between communities via the sharing of data analysis workflows. The collaboratory aims to provide a framework for researchers and workflow specialists to use and reuse workflows. As such it is an example of the Canonical Workflow Frameworks for Research (CWFR) [5] vision in practice.

EOSC-Life is made up of established research infrastructures ranging from biobanking and clinical trial management, through to coordinating biomedical imaging and plant phenotyping to multi-omic and systems-based data analysis. The heterogeneity of the disciplines is reflected in the diversity of their data analysis needs and practices and the variety of workflow management systems they use. Many have specialist platforms developed over years. Workflow management systems in common use include Galaxy [6], Snakemake [7], and Nextflow [8], and more specialist, domain-specific systems such as SCIPION [9].

To serve the needs of this established and diverse community, EOSC-Life has developed **WorkflowHub** [14] as an inclusive workflow registry, agnostic to any *Workflow Management System* (**WfMS**). WorkflowHub aims to incorporate their workflows in partnership with the WfMS, to embed the registration of workflows in the community processes, e.g. based on pre-existing workflow repositories. The registry adopts common practices, e.g.use of GitHub repositories, and supports integration with the ecosystem of tool packages, assisted by registries (bio.tools [10],biocontainers [11]), and services for testing and benchmarking workflows (OpenEBench [12],LifeMonitor [13]).

As an umbrella registry, the Hub makes workflows Findable and Accessible by indexing workflows across workflow management systems and their native repositories, while providing rich standardized metadata. Interoperability and Reusability is supported by standardized descriptions of workflows and packaging of workflow components, developed in close collaboration with the communities. The WorkflowHub creates a place for registering and discovering libraries of workflows developed by collaborating teams, with suitable features for versioning, credit, analytics, and import/export needed to support the reuse of workflows, the development of sub-workflows as canonical steps and ultimately the identification of common patterns in the workflows.

At the heart of the collaboratory is a Digital Object framework for documenting and exchanging workflows annotated with machine processable metadata produced and consumed by the participating platforms. The Digital Object framework is founded on several needs:

- *Describing a workflow and its steps in a canonical, normalised and WfMS independent way*: we use the **Common Workflow Language (CWL)** [15], more specifically the *Abstract CWL* [20] (non-executable) description variant to accompany the native workflow definitions. This presents the structure, composed tools and external interface in an interoperable way across workflow languages. WfMS can generate abstract CWL, already demonstrated for Galaxy, next to the 'native' Galaxy workflow description. This language duality is an important retention aspect of *reproducibility*, as the structure and metadata of the workflow can be accessed independent of its native format as CWL, even if that may no longer be executable, capturing the *canonical workflow* in a FAIR format. The co-presence of the native format enables direct reuse in the specific WfMS, benefitting from all its features.
- *Metadata about a workflow and its tools using a minimal information model:* we use the **Bioschemas** [16] profiles Computational Tool, Computational Workflow and Formal Parameter which are discipline independent, opinionated conventions for using schema.org annotations. Bioschemas enables us to capture and publish workflow registrations and their metadata as FAIR Digital Objects. The EDAM Ontology [17] is further used to add bioinformatics-specific metadata, such as strong typing of inputs and outputs, within both Abstract CWL and Bioschemas annotations.
- *Organising and packaging the definitions and components of a workflow* with their associated objects such as test data: we use a Workflow profile specialisation of **RO-Crate** [18], a community developed standardised approach for research output packaging with rich metadata. RO-Crate provides us the ability to package executable workflows, their components such as example and test data, abstract CWL, diagrams and their documentation. This makes workflows more readily re-usable. RO-Crate is the base unit of upload and download at the WorkflowHub. As CWFR Digital Objects of workflows, RO-Crates are activation-ready and circulated between the different services for execution and testing.
- *Identifiers* for all the components: like FAIR Digital Objects [19], RO-Crates can be metadata-rich bags of identifiers and can themselves be assigned permanent identifiers. This enables the full description of a computational analysis, from input data, over tools and workflows, to final results.

Using these components we have built an environment that supports the Workflow Life Cycle, from abstract description, through to a specific rendering in a WfMS to its execution and the documentation of its run provenance, results and continued testing.

Final Paper

In the final paper we will expand on our EOSC-Life Digital Object framework using deployed examples and partnerships with WfMS. We will dig deeper into challenges such as the multiple levels and granularity of workflow objects and the management of different WfMS implementations of the same canonical workflow, as well as practical deployment integrations such as GitHub and GA4GH standard APIs. We will review how the EOSC-Life Collaboratory can be viewed as a CWFR exemplar and how RO-Crate can be used as a developer-friendly metadata framework for FDOs. Further we will explore how CWFR principles impact and assist WorkflowHub for identifier assignment, FDO mutability and FAIR workflow reuse.

Funding acknowledgement

This work has received funding from the European Commission's Horizon 2020 research and innovation programme under grant agreement numbers 824087 (EOSC-Life) and 823830 (BioExcel-2) and is supported by Research Foundation - Flanders (FWO) for ELIXIR Belgium (I002819N).

References

[1] Taylor Reiter, Phillip T Brooks, Luiz Irber, Shannon E K Joslin, Charles M Reid, Camille Scott, C Titus Brown, N Tessa Pierce-Ward, **Streamlining data-intensive biology with workflow systems**. *GigaScience* 10(1), giaa140. https://doi.org/10.1093/gigascience/giaa140

[2] Franziska Hufsky, et al **Computational strategies to combat COVID-19: useful tools to accelerate SARS-CoV-2 and coronavirus research**, *Briefings in Bioinformatics*, 2020;, bbaa232, https://doi.org/10.1093/bib/bbaa232

[3] Carole Goble, Sarah Cohen-Boulakia, Stian Soiland-Reyes,Daniel Garijo, Yolanda Gil, Michael R. Crusoe, Kristian Peters, and Daniel Schober FAIR Computational Workflows *Data Intelligence* 2020 2:1-2, 108-121 https://doi.org/10.1162/dint_a_00033

[4] EOSC-Life https://www.eosc-life.eu/

[5] Alex Hardisty, Peter Wittenburg eds, Canonical Workflow Framework for Research CWFR- Position Paper V2, https://codata.org/wp-content/uploads/2021/01/CWFR-position-paper-v3.pdf, Dec 2020

[6] Enis Afgan, Dannon Baker, Bérénice Batut, Marius van den Beek, Dave Bouvier, Martin Čech, John Chilton, Dave Clements, Nate Coraor, Björn Grüning, Aysam Guerler, Jennifer Hillman-Jackson, Vahid Jalili, Helena Rasche, Nicola Soranzo, Jeremy Goecks, James Taylor, Anton Nekrutenko, and Daniel Blankenberg (2018): **The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update**, *Nucleic Acids Research*, Volume 46, Issue W1, 2 July 2018, Pages W537–W544 https://doi.org/10.1093/nar/gky379

[7] Johannes Köster, Sven Rahmann, **Snakemake—a scalable bioinformatics workflow engine**, *Bioinformatics*, Volume 28, Issue 19, 1 October 2012, Pages 2520–2522, https://doi.org/10.1093/bioinformatics/bts480

[8] Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). **Nextflow enables reproducible computational workflows.** *Nature Biotechnology*, 35(4), 316–319. https://doi.org/10.1038/nbt.3820

[9] J. Gómez-Blanco, J.M. de la Rosa-Trevín, R. Marabini, L. del Cano, A. Jiménez, M. Martínez, R. Melero, T. Majtner, D. Maluenda, J. Mota, Y. Rancel, E Ramírez-Aportela, J.L. Vilas, M. Carroni, et al, **Using Scipion for stream image processing at Cryo-EM facilities**, *Journal of Structural Biology*, Volume 204, Issue 3, 2018, Pages 457-463, https://doi.org/10.1016/j.jsb.2018.10.001

[10] Ison, J., Ienasescu, H., Chmura, P. *et al.* **The *bio.tools* registry of software tools and data resources for the life sciences**. *Genome Biol* **20,** 164 (2019). https://doi.org/10.1186/s13059-019-1772-6

[11] Felipe da Veiga Leprevost, Björn A Grüning, Saulo Alves Aflitos, Hannes L Röst, Julian Uszkoreit, Harald Barsnes, Marc Vaudel, Pablo Moreno, Laurent Gatto, Jonas Weber, et al, **BioContainers: an open-source and community-driven framework for software standardization**, *Bioinformatics*, Volume 33, Issue 16, 15 August 2017, Pages 2580–2582, https://doi.org/10.1093/bioinformatics/btx192

[12] OpenEBench https://openebench.bsc.es/about

[13] Life Monitor https://github.com/crs4/life_monitor

[14] WorkflowHub https://workflowhub.eu

[15] Peter Amstutz, Michael R. Crusoe, Nebojša Tijanić (editors), Brad Chapman, John Chilton, Michael Heuer, Andrey Kartashov, Dan Leehr, Hervé Ménager, Maya Nedeljkovich, Matt Scales, Stian Soiland-Reyes, Luka Stojanovic (2016): **Common Workflow Language, v1.0.** Specification, *Common Workflow Language working group*. https://doi.org/10.6084/m9.figshare.3115156.v2

[16] Bioschemas http://bioschemas.org

[17] Ison J, Kalas M, Jonassen I, et al. **EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats**. *Bioinformatics*. 2013;29(10):1325-1332. https://doi.org/10.1093/bioinformatics/btt113

[18] Eoghan Ó Carragáin, Carole Goble, Peter Sefton, Stian Soiland-Reyes (2019): **A lightweight approach to research object data packaging.** *Bioinformatics Open Source Conference* (BOSC2019) https://doi.org/10.5281/zenodo.3250687

[19] De Smedt K, Koureas D, Wittenburg P (2020) **FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units**. *Publications* 8 (2). https://doi.org/10.3390/publications8020021

[20] BioExcel (2020): **Creating workflows with Common Workflow Language**. *BioExcel Best Practice Guides.* https://docs.bioexcel.eu/cwl-best-practice-guide/devpractice/partial.html