Making Canonical Workflow Building Blocks interoperable across workflow languages

Stian Soiland-Reyes, Genís Bayarri, Pau Andrio, Robin Long, Douglas Lowe, Ania Niewielska, Adam Hospital

Extended abstract, submitted to Special issue on Canonical Workflow Frameworks for Research for Research in Data Intelligence Journal. This version: https://doi.org/10.5281/zenodo.4602855

We here introduce the concept of *Canonical Workflow Building Blocks* (**CWBB**), a methodology of describing and wrapping computational tools, in order for them to be utilized in a reproducible manner from multiple workflow languages and execution platforms. We argue such practice is a necessary requirement for FAIR Computational Workflows [Goble 2020] to improve widespread adoption and reuse of a computational method across workflow language barriers.

The need for *reproducibility* of research software usage is well established [Stodden 2016, Leipzig 2020, Katz 2021], and adaptation of *workflow management systems* (**WfMS**) together with *software packaging and containers* [Möller 2017] have been proposed as key ingredients for making research software usage *FAIR* [Cohen-Boulakia 2017] and reproducible [Leipzig 2020, Grüning 2018, Lamprecht 2020]. Recently it is also argued that computational workflows should also be treated as FAIR Digital Objects [De Smedt 2020] in their own right, with identifier, metadata and interoperability requirements [Goble 2020].

BioExcel is a European Centre of Excellence for Computational Biomolecular Research, having a particular workflow focus on the boundary of molecular dynamics simulations and bioinformatics analytics with use of *High Performance Computing* (**HPC**) to approach Exascale performance, while also improving usability. The *BioExcel Building Blocks* (**BioBB**) [Andrio 2019] have been created as portable wrappers of the open-source computational tools we identified as useful for BioExcel workflows, forming several families of documented and interoperable operations that can be called from multiple workflow systems, as shown with the BioBB demonstrator workflows [Hospital 2020], along with multiple tutorials and notebooks.

We here propose that these building blocks and their families can themselves be considered composite Digital Objects, as collections of *software packaging* (Pip, BioConda, BioContainers), *documentation* (ReadTheDocs), interactive *tutorials* (Jupyter Notebooks, myBinder), *registry & findability* (bio.tools, BioSchemas, WorkflowHub), *WfMS integration stubs* (CWL, Galaxy, PyCOMPSs), *Source Code* (GitHub) and *REST APIs* (OpenAPI, Swagger). In addition, the building blocks, as wrappers of upstream open source tools, benefit from and relate to the tools' existing documentation, support forums, academic publications and continued development.

While we started with documenting the collection of these views of the building blocks in human-readable text for users, we formalize how we can expose machine-readable workflow building blocks, using Bioschemas [Gray 2017] metadata to collate and register them as RO-Crate [Ó Carragáin 2019] packages to form FAIR Digital Objects. We explore how automatic generation of WFMS bindings for building blocks can help progress their FAIR metadata along with consistent documentation and tool usage patterns across workflow language barriers.

A question of granularity applies at the workflow tool level, particularly for Findability and Accessibility, as we can consider at lowest granularity the *scientific method* in general (e.g. any algorithm for sequence alignment), implemented by an *application suite* (bio.tools entry, homepage, documentation), instantiated as a particular *software installation* (Debian package, Docker container) with its dependencies at same level, which includes one or more *software executables* (a particular binary, a running service service), providing at the highest granularity level the specific types of *software functionality* (a particular mode of operation, choice of analysis).

While workflow management systems typically only operate at the highest granularity levels and are frequently unaware of or not exposing metadata at the more general level, we argue that in order for a Canonical Workflow [Wittenburg 2021] to follow and support FAIR principles for itself and its data, the workflow management system need to propagate structured metadata about the tools used by the workflow. We propose that in order to support the workflow's applicability to multiple WfMS, the tools themselves must also have a consistent packaging and formal description that enables consistent computational invocation.

We use our experiences with BioBB as a starting point to define the generalized methodology of *Canonical Workflow Building Blocks*: following a set of requirements and recommendations for how to formalize and develop a family of compatible computational tools as Digital Objects. These building blocks let researchers instantiate a Canonical Workflow in multiple workflow management systems, while retaining the FAIR aspects of the CWBB Digital Objects which may also assist the workflow designers in producing FAIR outputs in a consistent manner.

Final Paper

In the final paper we will relate how our building block method compares with and improves on existing workflow fragment library approaches, and how it takes advantage of modern best practices for reproducibility of research software usage.

Further we will argue how our proposed methodology of Canonical Workflow Building Blocks are important not just for usability aspects in workflow design, but also for reproducibility and portability across workflow management systems, as well as for propagating FAIR metadata of computational tools across multiple instantiations of a Canonical Workflow on different workflow management systems.

We will cover in detail how existing standards and practices like Common Workflow Language [Amstutz 2016], Bioschemas [Gray 2017], RO-Crate [Ó Carragáin 2019] and PIDs [McMurry 2017] provide a reliable and extensible metadata framework for Canonical Workflow Building Blocks, but also highlight their current limitations and challenges. We will be exploring the implications of CWBB for the future directions of Canonical Workflow Framework for Research (CWFR) [Wittenburg 2021] and how a CWFR approach can support and be supported by current Workflow Management Systems.

Finally we will formalize the concept of Canonical Workflow Building Blocks as a set of requirements, demonstrate to what extent our current approach fulfils these and in what way the CWFR approach with the help of CWBB can help achieve FAIR Computational Workflows across the workflow language barriers.

Acknowledgements

This work has been done as part of the BioExcel CoE (www.bioexcel.eu), a project funded by the European Union contract H2020-INFRAEDI-02-2018-823830.

Author affiliations

Stian Soiland-Reyes https://orcid.org/0000-0001-9842-9718
Department of Computer Science, The University of Manchester, Manchester, UK;
Informatics Institute, University of Netherlands, NL

Genís Bayarri
Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain

Pau Andrio https://orcid.org/0000-0003-2116-3880
Barcelona Supercomputing Center (BSC), Barcelona, Spain

Robin Long https://orcid.org/0000-0003-2249-645X
Research IT, The University of Manchester, Manchester, UK

Douglas Lowe https://orcid.org/0000-0002-1248-3594
Research IT, The University of Manchester, Manchester, UK

Ania Niewielska https://orcid.org/0000-0003-0989-3389
European Bioinformatics Institute (EMBL-EBI), Cambridge, UK

Adam Hospital https://orcid.org/0000-0002-8291-8071
Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain

References

[Amstutz 2016] 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://w3id.org/cwl/v1.0/ https://doi.org/10.6084/m9.figshare.3115156.v2

[Andrio 2019] Pau Andrio, Adam Hospital, Javier Conejero, Luis Jordá, Marc Del Pino, Laia Codo, Stian Soiland-Reyes, Carole Goble, Daniele Lezzi, Rosa M. Badia, Modesto Orozco, Josep Ll. Gelpi (2019): **BioExcel Building Blocks, a software library for interoperable biomolecular simulation workflows**. *Scientific Data* **6**:169 https://doi.org/10.1038/s41597-019-0177-4

[Cohen-Boulakia 2017] Sarah Cohen-Boulakia, Khalid Belhajjame, Olivier Collin, Jérôme Chopard, Christine Froidevaux, Alban Gaignard, Konrad Hinsen, Pierre Larmande, Yvan Le Bras, Frédéric Lemoine, Fabien Mareuil, Hervé Ménager, Christophe Pradal, Christophe Blanchet (2017): **Scientific Workflows for Computational Reproducibility in the Life Sciences: Status, Challenges and Opportunities**. *Future Generation Computer Systems* **75,** pp 284–298. https://doi.org/10.1016/j.future.2017.01.012

[De Smedt 2020] Koenraad De Smedt, Dimitris Koureas, Peter Wittenburg (2020): **FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units**. *Publications* **8**(2):21 https://doi.org/10.3390/publications8020021

[Gray 2017] Alasdair Gray, Carole Goble, Rafael Jimenez, Bioschemas Community (2017): **Bioschemas: From Potato Salad to Protein Annotation**. Poster, *International Semantic Web Conference (ISWC)*, Vienna Austria, 2017-10-23. https://iswc2017.semanticweb.org/paper-579/

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

[Grüning 2018] Björn Grüning, John Chilton, Johannes Köster, Ryan Dale, Nicola Soranzo, Marius van den Beek, Jeremy Goecks, Rolf Backofen, Anton Nekrutenko, James Taylor (2018): **Practical Computational Reproducibility in the Life Sciences**. *Cell Systems* **6**(6) pp 631–635. https://doi.org/10.1016/j.cels.2018.03.014

[Hospital 2020] Adam Hospital, Genís Bayarri, Stian Soiland-Reyes, Jose Lluis Gelpi, Pau Andrio, Daniele Lezzi, Sarah Butcher, Ania Niewielska, Yvonne Westermaier, Rosa Maria Badia, Rodrigo Vargas, Alexandre Bonvin (2020): **BioExcel-2 Deliverable 2.3 – First release of demonstration workflows.** Project deliverable, *Zenodo*. https://doi.org/10.5281/zenodo.4540432

[Katz 2021] Daniel S. Katz, Morane Gruenpeter, Tom Honeyman, Lorraine Hwang, Mark D. Wilkinson, Vanessa Sochat, Hartwig Anzt, Carole Goble, FAIR4RS Subgroup 1 (2021): **A Fresh Look at FAIR for Research Software**. arXiv:2101.10883

[Lamprecht 2020] Anna-Lena Lamprecht, Leyla Garcia, Mateusz Kuzak, Carlos Martinez, Ricardo Arcila, Eva Martin Del Pico, Victoria Dominguez Del Angel, et al. **Towards FAIR Principles for Research Software**. *Data Science* **3**(1), pp 37–59. https://doi.org/10.3233/ds-190026

[Leipzig 2020] Jeremy Leipzig, Daniel Nüst, Charles Tapley Hoyt, Stian Soiland-Reyes, Karthik Ram, Jane Greenberg (2020): **The role of metadata in reproducible computational research**. arXiv:2006.08589

[McMurry 2017] Julie A McMurry, Nick Juty, Niklas Blomberg, Tony Burdett, Tom Conlin, Nathalie Conte, Mélanie Courtot, John Deck, Michel Dumontier, Donal K Fellows, Alejandra Gonzalez-Beltran, Philipp Gormanns, Jeffrey Grethe, Janna Hastings, Jean-Karim Hériché, Henning Hermjakob, Jon C Ison, Rafael C Jimenez, Simon Jupp, John Kunze, Camille Laibe, Nicolas Le Novère, James Malone, Maria Jesus Martin, Johanna R McEntyre, Chris Morris, Juha Muilu, Wolfgang Müller, Philippe Rocca-Serra, Susanna-Assunta Sansone, Murat Sariyar, Jacky L Snoep, Stian Soiland-Reyes, Natalie J Stanford, Neil Swainston, Nicole Washington, Alan R Williams, Sarala M Wimalaratne, Lilly M Winfree, Katherine Wolstencroft, Carole Goble, Cristopher J Mungall, Melissa A Haendel, Helen Parkinson (2017): **Identifiers for the 21st century: How to design, provision, and reuse identifiers to maximize utility and impact of life science data**. *PLOS Biology* **15**(6):e2001414 https://doi.org/10.1371/journal.pbio.2001414

[Möller 2017] Steffen Möller, Stuart W. Prescott, Lars Wirzenius, Petter Reinholdtsen, Brad Chapman, Pjotr Prins, Stian Soiland-Reyes, Fabian Klötzl, Andrea Bagnacani, Matúš Kalaš, Andreas Tille, Michael R. Crusoe (2017): **Robust cross-platform workflows: How technical and scientific communities collaborate to develop, test and share best practices for data analysis**. *Data Science and Engineering* **2**, pp 232–244. https://doi.org/10.1007/s41019-017-0050-4

[Ó Carragáin 2019] Eoghan Ó Carragáin; Carole Goble; Peter Sefton; Stian Soiland-Reyes (2019): **A lightweight approach to research object data packaging**. *Bioinformatics Open Source Conference (BOSC)*, ISMB/ECCB 2019, Basel, Switzerland, 24–25 July 2019. https://doi.org/10.5281/zenodo.3250687

[Stodden 2016] Victoria Stodden, Marcia McNutt, David H. Bailey, Ewa Deelman, Yolanda Gil, Brooks Hanson, Michael A. Heroux, John P.A. Ioannidis, Michela Taufer (2016): **Enhancing reproducibility for computational methods.** *Science* **354**(6317), pp 1240–1241 https://doi.org/10.1126/science.aah6168

[Wittenburg 2021] Peter Wittenburg et al. (2021): **CWFR Position Paper**. *OSF*, January 6, 2021. https://osf.io/3rekv/