On Managing Large Collections of Scientific Workflows

Nourhan Elfaramawy [1], Fatma Deniz [2], Lars Grunske [1], Marcus Hilbrich [1], Timo Kehrer [3], Anna-Lena Lamprecht [4], Jan Mendling [1], and Matthias Weidlich [1]

1 The Context of Scientific Workflows

Many scientific disciplines, spanning from neuroscience and earth observation to astrophysics and materials science, continuously produce ever-increasing amounts of data. Scientific discoveries then emerge from the analysis of large-scale datasets that are processed with scientific workflows [At17; Le21], i.e., models that capture series of analysis programs that are arranged in pipelines. Adopting the paradigm of scientific workflows has various advantages over an ad-hoc implementation of the analysis: Workflows provide abstractions to change the infrastructure used for executing the analysis, from single machines to large compute clusters; they facilitate the reuse of standardized operators by providing dedicated operator libraries; and they foster reproducibility and traceability of analysis results by making the chaining of operators and their configurations explicit.

As an example, consider Fig. 1, which illustrates the PopIns workflow [KMH16] for variant calling, i.e., the identification of insertions in sequencing data to identify genetic variations. The respective model captures programs (denoted as circles) to *assemble*, *merge*, *align*, and *place* sequencing data, along with their input and output data dependencies.

In recent years, large collections of scientific workflows emerged, including those targeting the general exchange of workflows, e.g., myExperiment [Go10] and WorkflowHub [Go21], as well as those focusing on workflows implemented for a specific workflow engine, such as nf-core [Ew20] for Nextflow and the Snakemake workflow catalog [Kö24]. While these repositories denote a valuable source of information and best practices, they typically include only rudimentary functionality for managing the workflows, such as classification schemes (e.g., by domain, workflow engine, and maturity) and full-text search. As such, the reuse and adaptation of (partial) workflows for new types of analysis and their integration into an existing pipeline is a largely manual, cumbersome process. Hence, opportunities for

1 Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany,
nourhan.elfaramawy@hu-berlin.de, https://orcid.org/0000-0001-9444-5163;
lars.grunske@hu-berlin.de, https://orcid.org/0000-0002-8747-3745;
marcus.hilbrich@hu-berlin.de, https://orcid.org/0000-0003-3717-9449;
jan.mendling@hu-berlin.de, https://orcid.org/0000-0002-7260-524X;
matthias.weidlich@hu-berlin.de, https://orcid.org/0000-0003-3325-7227
2 Technische Universität Berlin, Straße des 17. Juni 135, 10623 Berlin, Germany,
deniz@tu-berlin.de, https://orcid.org/0000-0001-6051-7288
3 Universität Bern, 3012 Bern, Schweiz, timo.kehrer@unibe.ch, https://orcid.org/0000-0002-2582-5557
4 Universität Potsdam, Am Neuen Palais 10, 14469 Potsdam, Germany,
anna-lena.lamprecht@uni-potsdam.de, https://orcid.org/0000-0003-1953-5606

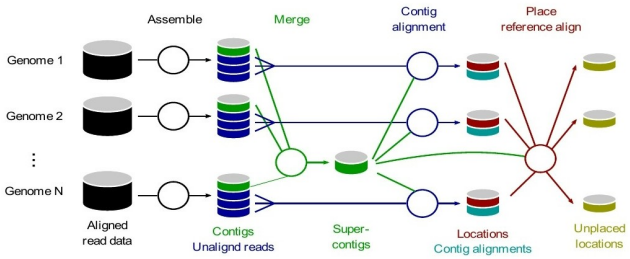

Fig. 1: The PopIns workflow [KMH16] for variant calling, taken from [El22]. Here, the circles denote programs that are executed on specific types of genomic data.

reuse and adaptation are often not exploited, leading to an increased development effort and, generally, lower quality of the resulting workflows.

To fully leverage the potential of collections of scientific workflows, research efforts that aim at a more holistic management of scientific workflows are needed. Similar to other artifacts that occur in a scientific process, scientific workflows should be subject to research data management along the FAIR principles [Go20], ensuring findability, accessibility, interoperability, and reusability. Below, we discuss three directions for such research efforts that focus on the comprehensive specification of workflows, traceability of the design choices made during their development and means for comparative analysis.

2 Towards Research Data Management for Scientific Workflows

Research data management for scientific workflows shall ensure that (i) users can discover the workflows relevant for a certain type of analysis; (ii) users can access the workflows using standardized protocols; (iii) users know how to integrate the workflows in their specific environment; and (iv) users may derive how to reuse the workflows for a specific research task. Accessibility of scientific workflows may primarily be seen as a technical concern, involving standardized serialization formats and protocols for exchange, whereas authentication and authorization are often negligible. Yet, to achieve findability and support users in the assessment of interoperability and reusability, we see open issues along at least the following three dimensions.

Comprehensiveness of Workflow Definitions. The discovery of workflows for a specific task requires that workflows are specified comprehensively. A workflow definition shall not only cover functional aspects, e.g., the workflow objective and the semantics of the contained programs [Be05], but also include information on validity assumptions [Sc23]. Such assumptions may refer to the input data of the workflow or the execution infrastructure. Yet, they are often kept implicit, which compromises dependability. The use of a workflow in a context for which it was not designed, potentially leads to manifold errors at runtime or, even worse, flawed results that remain undetected.

The comprehensive definition of workflows including validity assumptions imposes additional efforts for users. Hence, it is of crucial importance to provide support for the

extraction and definition of these assumptions. To this end, automated discovery of patterns in monitoring data [Kl23] or contract-driven approaches to workflow specification [Vu23] may turn out to be beneficial.

Traceability of Workflow Design. While freely accessible repositories of scientific workflows are a prerequisite for knowledge transfer within a research community, effective reuse of functionality and the establishment of best practices also require traceability of the workflow design. Meta-information on how the published workflow has been obtained enables users to understand the rationale behind incorporated design choices, thereby establishing trust and eventually fostering their adoption.

Research data management of scientific workflows shall therefore include not only the final artifact but also cover the process of its design that eventually led to the published form. To achieve such traceability, the analysis of the evolution of a scientific workflow (e.g., based on GIT revisions, Research Data Management Plans) as well as logging information on executions of earlier versions of a workflow may provide valuable starting points. At the same time, a better understanding of the conception and composition phases of the workflow life cycle [La21] is needed. In these phases, an abstract, methodical sketch of a workflow is first derived for a specific scientific hypothesis, before it is implemented by composing analysis programs. However, the question of which type of information about these early phases of the workflow life cycle is eventually valuable to guide users in the reuse of a workflow remains largely open.

Comparability of Workflows. Discovery and integration of existing workflows calls for means to compare workflow specifications. The identification of equivalent or similar parts of a workflow serves to judge its suitability for a specific analysis task. Existing means to query workflow repositories and extract recommendations [Zh18] rely on similarity measures for scientific workflows [St16]. Yet, they are limited in their ability to cope with the heterogeneity observed in the languages used for the definition of scientific workflows and typically neglect differences in the modularization of the workflow functionality.

Further research efforts are needed to facilitate the comparison of a broad range of workflows. Specifically, such efforts may build on meta-models for scientific workflows [Am16; Hi22] to bridge the differences between various specification languages and technical environments. To address the challenges related to the differences in the adopted modularization strategies, empirical studies on common workflow patterns [Ga14; Po23] may provide valuable insights that can be incorporated in approaches for workflow comparison.

3 Conclusion

Scientific workflows facilitate the analysis of large-scale data sets, thereby providing the backbone of research efforts in many scientific disciplines. Such workflows are models that describe the composition of analysis programs. In light of this, we suggest that they should

be incorporated into robust and suitable research data management practices. To manage collections of scientific workflows effectively, our recommendations incorporate three areas for further exploration. These areas encompass enhancing the workflow specification, the traceability of workflow design, and the comparative analysis of various workflows.

References

[Am16] Amstutz, P.; Crusoe, M. R.; Tijanić, N.; Chapman, B.; Chilton, J.; Heuer, M.; Kartashov, A.; Leehr, D., et al.: Common workflow language, v1. 0, 2016.

[At17] Atkinson, M. P.; Gesing, S.; Montagnat, J.; Taylor, I. J.: Scientific workflows: Past, present and future. Future Gener. Comput. Syst. 75/, pp. 216–227, 2017.

[Be05] Berkley, C.; Bowers, S.; Jones, M. B.; Ludäscher, B.; Schildhauer, M.; Tao, J.: Incorporating Semantics in Scientific Workflow Authoring. In: SSDBM 2005. Pp. 75–78, 2005.

[El22] Elfaramawy, N.: Interactive Workflows for Exploratory Data Analysis. In: Proceedings of the VLDB 2022 PhD Workshop. Vol. 3186. CEUR, CEUR-WS.org, 2022.

[Ew20] Ewels, P. A.; Peltzer, A.; Fillinger, S.; Patel, H.; Alneberg, J.; Wilm, A.; Garcia, M. U.; Di Tommaso, P., et al.: The nf-core framework for community-curated bioinformatics pipelines. Nature biotechnology 38/3, pp. 276–278, 2020.

[Ga14] Garijo, D.; Alper, P.; Belhajjame, K.; Corcho, Ó.; Gil, Y.; Goble, C. A.: Common motifs in scientific workflows: An empirical analysis. Future Gener. Comput. Syst. 36/, pp. 338–351, 2014.

[Go10] Goble, C. A.; Bhagat, J.; Aleksejevs, S.; Cruickshank, D.; Michaelides, D. T.; Newman, D. R.; Borkum, M.; Bechhofer, S., et al.: myExperiment: a repository and social network for the sharing of bioinformatics workflows. Nucleic Acids Res. 38/Web-Server-Issue, pp. 677–682, 2010.

[Go20] Goble, C.; Cohen-Boulakia, S.; Soiland-Reyes, S.; Garijo, D.; Gil, Y.; Crusoe, M. R.; Peters, K.; Schober, D.: FAIR Computational Workflows. Data Intelligence 2/1-2, pp. 108–121, 2020.

[Go21] Goble, C.; Soiland-Reyes, S.; Bacall, F.; Owen, S.; Williams, A.; Eguinoa, I.; Droesbeke, B.; Leo, S., et al.: Implementing FAIR Digital Objects in the EOSC-Life Workflow Collaboratory, 2021.

[Hi22] Hilbrich, M.; Müller, S.; Kulagina, S.; Lazik, C.; Mecquenem, N. D.; Grunske, L.: A Consolidated View on Specification Languages for Data Analysis Workflows. In: ISoLA 2022. Vol. 13702. LNCS, Springer, pp. 201–215, 2022.

[Kl23] Kleest-Meißner, S.; Sattler, R.; Schmid, M. L.; Schweikardt, N.; Weidlich, M.: Discovering Multi-Dimensional Subsequence Queries from Traces - From Theory to Practice. In: (BTW 2023), Proceedings. Vol. P-331. LNI, GI e.V., pp. 511–533, 2023.

[KMH16] Kehr, B.; Melsted, P.; Halldórsson, B. V.: PopIns: population-scale detection of novel sequence insertions. Bioinform. 32/7, pp. 961–967, 2016.

[Kö24] Köster, J., et al.: Snakemake workflow catalog, https://snakemake.github.io/snakemake-workflow-catalog/, 2024.

[La21] Lamprecht, A.; Palmblad, M.; Ison, J. C.; Schwämmle, V.; Manir, M. S. A.; Altintas, I.; Baker, C. J. O.; Amor, A. B. H., et al.: Perspectives on automated composition of workflows in the life sciences. F1000Research 10/, p. 897, 2021.

[Le21] Leser, U.; Hilbrich, M.; Draxl, C.; Eisert, P.; Grunske, L.; Hostert, P.; Kainmüller, D.; Kao, O., et al.: The Collaborative Research Center FONDA. Datenbank-Spektrum 21/3, pp. 255–260, 2021.

[Po23] Pohl, S.; Elfaramawy, N.; Cao, K.; Kehr, B.; Weidlich, M.: How do users design scientific workflows? The Case of Snakemake. CoRR abs/2309.14097/, 2023, arXiv: 2309.14097.

[Sc23] Schintke, F.; Mecquenem, N. D.; Frantz, D.; Guarino, V. E.; Hilbrich, M.; Lehmann, F.; Sattler, R.; Sparka, J. A., et al.: Validity Constraints for Data Analysis Workflows. CoRR abs/2305.08409/, 2023, arXiv: 2305.08409.

[St16] Starlinger, J.; Boulakia, S. C.; Khanna, S.; Davidson, S. B.; Leser, U.: Effective and efficient similarity search in scientific workflow repositories. Future Gener. Comput. Syst. 56/, pp. 584–594, 2016.

[Vu23] Vu, A. D.; Sparka, J. A.; Mecquenem, N. D.; Kehrer, T.; Leser, U.; Grunske, L.: Contract-Driven Design of Scientific Data Analysis Workflows. In: IEEE e-Science 2023. IEEE, pp. 1–10, 2023.

[Zh18] Zhou, Z.; Cheng, Z.; Zhang, L.; Gaaloul, W.; Ning, K.: Scientific Workflow Clustering and Recommendation Leveraging Layer Hierarchical Analysis. IEEE Trans. Serv. Comput. 11/1, pp. 169–183, 2018.