

ShareFAIR-KG, a centralised knowledge base of scientific workflows

Marie SCHMIT¹, Melvin Selim ATAY², Khalid BELHAJJAME³, Ulysse LE CLANCHE⁴, Emmanuel COQUERY⁵, Olivier DAMERON⁴, Fabien DUCHATEAU⁵, Alban GAIGNARD^{6,7}, Mouna EL GARB⁵, Jaffar GURA³, Nicolas LUMINEAU⁵, George MARCHMENT⁸, Camille MAUMET², Clémence SEBE⁸, Frédéric LEMOINE¹, Sarah COHEN-BOULAKIA⁸ and Hervé MÉNAGER^{1,7}

¹ Institut Pasteur, Université Paris Cité, Bioinformatics of Biostatistics Hub, F-75015 Paris, France

² Univ Rennes, CNRS, Inria, Inserm, IRISA UMR 6074, EMPENN — ERL U 1228, F-35000 Rennes, France

³ LAMSADE, PSL, Paris Dauphine University, 75016 Paris, France

⁴ Université Rennes, Inria, CNRS, IRISA—UMR 6074, Rennes 35000, France

⁵ Université Lyon 1, INSA Lyon, CNRS, LIRIS, UMR 5205, Lyon, France

⁶ Nantes Université, CNRS, INSERM, l'Institut du Thorax, F-44000 Nantes, France

⁷ IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA, 94800 Villejuif, France

⁸ Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, 91405, Orsay, France

Corresponding author: herve.menager@pasteur.fr

Keywords Knowledge graphs, workflows, FAIR, SPARQL, metadata

Abstract *The emergence of high-throughput technologies in the Life Sciences has led to increasingly complex data analysis pipelines. Ensuring their transparency, reusability, and reproducibility requires adherence to the FAIR principles, which motivated the development of the ShareFAIR project. ShareFAIR aims to build a framework to make workflows more shareable, understandable, and reusable by improving their description through structure, provenance, and annotations.*

The ShareFAIR architecture consists of several modules addressing workflow structure extraction, provenance, annotation, and querying, with applications ranging from genomics to neuroimaging. At its core, this system depends on ShareFAIR-KG, a centralized and queryable knowledge graph (KG) that integrates all workflow components and tool metadata at various levels of granularity. Populated through automated harvesting and normalized into a graph-based representation, it consolidates the results and software produced in the project. With the help of ShareFAIR-KG, users can for instance research workflows implementing specific analytical steps, obtain the software packages used or wrapped in workflows or query citation and licensing information.

Introduction

During the last decade, an unprecedented amount of multi-scale data has been generated in the field of Life Sciences. While it offers new opportunities for biological discovery, its volume and complexity have introduced significant computational challenges. Notably, the difficulty in reproducing data analysis pipelines has become prevalent, eroding trust in scientific results and limiting the reuse of established analysis pipelines [1]. This issue is particularly relevant in bioinformatics and extends to the neuroimaging field [2].

Scientific workflows were developed to address these limitations, offering scalability, automation and reproducibility [3]. However, workflows alone are insufficient: the application of FAIR principles (Findability, Accessibility, Interoperability and Reuse) [4,5,6,7] to these pipelines is increasingly recognised

as essential for ensuring transparency, quality and reliability of computational results [4]. Other authors have highlighted the importance to develop dedicated best practices for research software that also apply to workflows [8].

The [ShareFAIR](#) project arose from this need to make scientific workflows FAIR, easier to share and to reproduce. It seeks to enhance the reliability of both datasets (e.g. finely annotated raw data and analysis results) and analysis protocols by explicitly tracking provenance relationship between data results and analytical methods. To achieve this objective, the ShareFAIR architecture includes multiple modules for the annotation, query, extraction and representation of workflow structure and execution traces. The basis of this architecture is a knowledge graph of annotated workflows from different languages and sources, gathering and allowing to query data generated in those modules. This knowledge base facilitates the discovery of workflows with their scientific context, structure and execution traces.

In this work, we present ShareFAIR-KG, a knowledge-graph of normalised FAIR workflow metadata. It leverages existing state-of-the-art and community adopted standards to link the structural representation of scientific pipelines with their execution traces, software components and their parameters. Although there are several initiatives for sharing bioinformatics workflows and software such as the WorkflowHub directory [9] or the bio.tools registry [10,11], the aim of ShareFAIR-KG is to bring them together by adding a semantic layer on top of them. Thus, ShareFAIR-KG automatically harvests metadata from multiple existing repositories and enables their discovery through complex querying. The knowledge base supports several use cases. Researchers can explore workflows implementing specific analytical steps, facilitating the location and comparison of workflows across all supported languages. Tool developers can query software packages used or wrapped in workflows, for instance to guide the creation of new wrappers. Finally, data managers and curators can evaluate workflow FAIR compliance by querying licensing or citation information. As the groundwork of the ShareFAIR architecture, ShareFAIR-KG integrates functionalities and results from ongoing collaborative development across multiple modules of the ShareFAIR project.

This paper is organised as follows. First, we present the standards selected for workflow representation, that build the semantics of ShareFAIR-KG. Next, we explain how the knowledge graph is deployed and populated, showing examples of use cases to demonstrate its query-ability and how it explains workflows at every granularity level. Finally, we detail how the knowledge base serves as the integration layer for other modules of the ShareFAIR project, and how we plan to improve this joint work to expand its content and capabilities.

Scientific workflow and tool metadata standards

Comprehensive workflow description requires a representation model with sufficient granularity, as workflows can be studied at different scales: (i) the workflows and their scientific domains at the highest level, (ii) then their processes, either nested subworkflows or computational units calling a (iv) software with its parameters. Those components and the data-flow between them must also be considered from both the prospective (specification) and retrospective (execution traces) provenance perspective.

This representation is further subject to additional critical requirements. First, it must integrate existing ontologies, schemas or vocabularies while avoiding to redefine yet another standard. Then, it must facilitate metadata generation and exchange by adopting interoperable and reusable standards that are widely adopted by the community and already integrated by existing frameworks.

The ShareFAIR-KG semantic model [12] complies with those constraints and provides (i) scientific context and metadata for workflows and their constituent software and steps, using domain ontologies such as EDAM [13]; (ii) unique software referencing via bio.tools [10,11] identifiers; and (iii) prospective and retrospective provenance representation through Provenance Run Crate [14], an [RO-Crate](#) profile extending WorkflowHub's [workflow profile](#) (describing workflow files and metadata), Process Run Crate (software used for file creation) and Workflow Run Crate (workflow execution by a Workflow Management System or WMS).

Populating and querying ShareFAIR-KG

ShareFAIR-KG is a knowledge base of workflow metadata following this representation model. Queryable with [SPARQL](#), it is deployed in a [Jena Fuseki](#) server, inside a [Docker](#) container providing isolation and reproducibility.

The knowledge base integrates (i) EDAM domain-specific ontology for annotation, (ii) bio.tools registry for software identification and as a source of pre-existing EDAM annotations, (iii) WorkflowHub general metadata on workflows, (iv) metadata on the prospective and retrospective provenance of workflows from various WMS and sources, including: WorkflowHub, [nf-core](#), Common Workflow Language (CWL) [15,16] and the [Intergalactic Workflow Commission \(IWC\)](#) library. We chose those registries and libraries for their wide-spread adoption by the community and their alignment with FAIR principles, leveraging open registries (e.g. bio.tools) and standards (e.g. RO-Crate).

ShareFAIR-KG currently contains the prospective provenance of 49 Nextflow [17] and 73 Common Workflow Language (CWL) workflows; and both types of provenance for 102 Galaxy[18] workflows¹⁸ executed on Galaxy server. This dataset spans workflows covering very diverse domains of bioinformatics, from genome assembly to metabolite profiling. Querying this knowledge graph with SPARQL yields deeper insights into these workflows, as shown in the following use cases (further detailed in [12]):

- **Tool adoption** An example query is the search for the most frequently used tools¹⁹ from bio.tools among workflows in the knowledge base, with their EDAM operations. Querying ShareFAIR-KG reveals that the top two by frequency are MultiQC [19] and dada2 [20], performing the operations [sequencing quality control](#) and [DNA barcoding, variant calling](#) (Table 1).
- **Workflow using specific tools** Knowing that MultiQC is the most used tool amongst workflows in ShareFAIR-KG, the workflows using it can be queried. For example, the Galaxy workflows *Genome Assembly from Hifi reads - VGP3* and *Scaffolding with Hi-C data VGP8* both use MultiQC.
- **Workflow input format** Another example of query involves retrieving workflow input and output data. For instance, the inputs of the CWL workflow [gatk4W](#) have the formats [FASTA](#) and [FASTQ](#).

¹⁸ Workflows were selected based on open-source licencing and technical compatibility with our extraction tools.

¹⁹ Tools that are the most frequently used in workflows, not the ones which have been run the most overall through these workflows.

- **Workflow tools output formats** The last query example infers workflow outputs from bio.tools EDAM annotations. Annotations of the bio.tools software executed in the Galaxy workflow *ATAC-seq* [21] indicate that it generates [BED](#), [WIG](#), [TSV](#), [HTML](#), [JSON](#), [YAML](#) and [CSV](#) data.

Tab. 1. Results of the SPARQL query retrieving most-used bio.tools tools, and their integration frequency in steps (# workflow steps). The names of the tools are from bio.tools. The namespace edam corresponds to <http://edamontology.org/>.

Tool name	bio.tools identifier	EDAM operation	EDAM operation label	# workflow steps
MultiQC	https://bio.tools/multiqc	edam:operation.3218, edam:operation.2428	Sequencing quality control, Validation	18
dada2	https://bio.tools/dada2	edam:operation.3200, edam:operation.3227	DNA barcoding, Variant calling	12
Bandage	https://bio.tools/bandage	edam:operation.3184	Sequence assembly visualisation	5

Current and Future Integration within the ShareFAIR Project

ShareFAIR-KG constitutes the foundational infrastructure of the ShareFAIR architecture, centralising output from other modules and providing data for query and representation purposes to others (Figure 1). While it already integrates several of their functionalities, ongoing efforts aim to strengthen this integration and consolidate the ShareFAIR framework. This collaboration will address key limitations by enhancing annotation depth and quality as well as expanding workflow coverage.

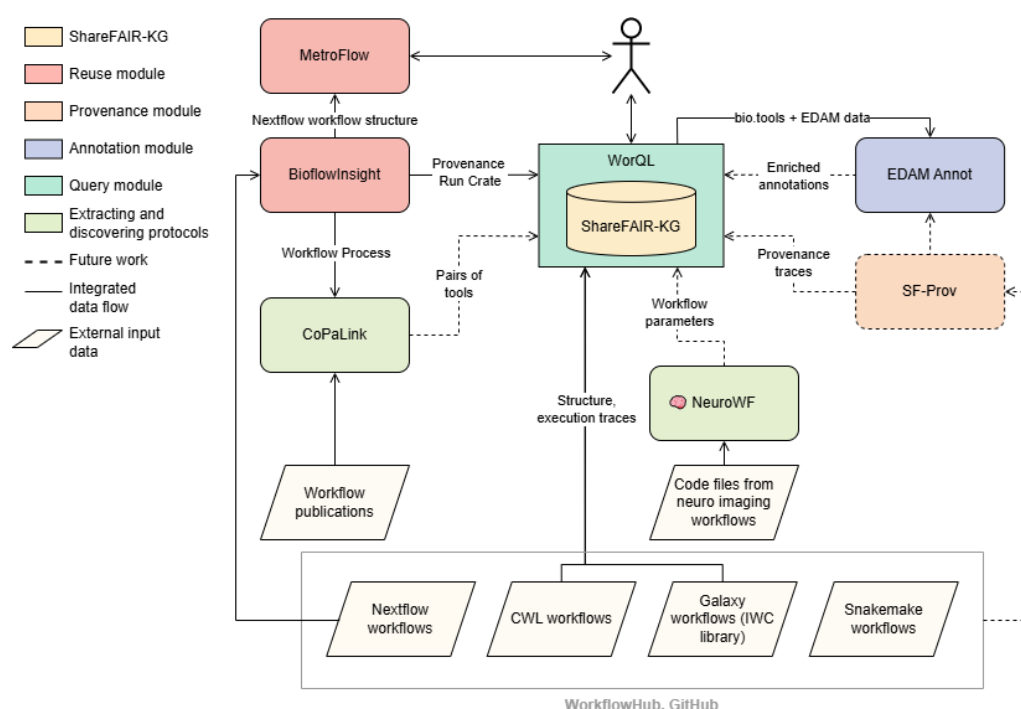


Fig. 1. ShareFAIR architecture with the software developed in different modules and the data flowing between them (type of data is specific on the arrows). Solid arrows represent current integration, dotted arrows show work in progress.

The Reuse Module aims to extract and visualise workflow structure. As part of this module, BioflowInsight [22] was developed to capture Nextflow workflow structures and generate their Provenance Run Crate metadata within ShareFAIR-KG. Additionally, MetroFlow [23] provides a unique visualization of Nextflow workflows, representing them as intuitive metro maps.

The Provenance Module ingests and exploits provenance traces captured from workflow executions. Only a limited number of WMS supports provenance capture, such as Galaxy and Nextflow with the plugin [nf-prov](#). This module currently aligns nf-prov provenance with prospective specifications. In the short-term, our road map includes the verification and repair of the completeness of nf-prov outputs. In the longer term, we investigate multi-granular representations of provenance, including fine-grained execution traces. Both the repair and the refined provenance derivations (at different levels of granularity) will be reflected in ShareFAIR-KG.

The Annotation Module aims at inferring missing metadata to better annotate and explain scientific pipelines, as well as analyzed data. Bio.tools metadata and the EDAM ontology from ShareFAIR-KG are at the core of the EDAMAnnot [\[24\]](#) toolbox and were used to investigate metadata quality in terms of information content. Neuro-symbolic approaches leveraging both retrieval augmented generation and knowledge graphs are under investigation to i) increase the quality of bioinformatics tools and workflow metadata and ii) annotate analysed data for increase potential reuse.

The Query Module leverages BioFlow-Ontology [\[25\]](#), a comprehensive workflow representation model, to develop WorQL, a dedicated query language for workflow interrogation. WorQL employs *query relaxation mechanisms*, such that if a query yields no results, it is automatically reformulated to return meaningful matches [\[26\]](#). This enhances ShareFAIR-KG accessibility, enabling users from diverse backgrounds to efficiently query workflows and fostering broader adoption. Bioflow-Ontology and the ShareFAIR-KG representation model serve complementary purposes: one extends existing standards for seamless metadata generation and interoperability, the other optimises workflow querying. The integration of this module with ShareFAIR-KG will leverage inference rules and reasoning to align both data representations.

Linking bioinformatics tools across publication and workflow code Workflow descriptions in publications and actual code often diverge. For example, the same tool may appear under different names, or can be omitted from the publication. To address this, CoPaLink [\[27\]](#) has been developed to extract all bioinformatics tool names from both sources (publications and code). This allows users to compare tools detected in each source and identify tools that appear in only one source. CoPaLink will be integrated in ShareFAIR-KG.

Neuroimaging Workflow Integration (NeuroWF), currently in development Besides bioinformatics workflows and provenance tools, ShareFAIR-KG will enable querying and exploring the collection of open source neuroimaging workflows. In particular, we will generate Provenance Run Crate metadata to describe functional magnetic resonance imaging (task-fMRI) pipelines of the most commonly used software tools. By using the previously described modules, NeuroWF aims to enable parameter metadata [\[28\]](#) querying for further analysis of neuroimaging workflows in a reproducible manner.

Conclusion

This work paves the way for FAIR sharing of scientific workflows. It leverages the results and functionalities developed in the ShareFAIR project modules to gather and normalise workflow annotations and metadata from multiple sources and languages, providing details on their scientific context, structure and execution traces to facilitate workflow discovery, understanding, sharing and reuse.

ShareFAIR-KG still faces specific constraints, mainly related to the quality and completeness of annotations and workflow descriptions. Query capabilities are currently limited by the variable depth and coverage of workflow description. Additionally, the knowledge graph is rebuilt from scratch during each update rather than maintained incrementally with workflow versioning support, limiting its scalability for evolving pipelines.

Future work will first address the lack of complete workflow annotations by finalising the integration with the other ShareFAIR modules, to integrate a larger diversity of standards and increase annotation quality, for instance by enriching bio.tools annotations. Furthermore, we plan to increase the scale of workflow integration to encompass neuroimaging workflows, while expanding the coverage of execution traces beyond Galaxy workflows. Joint efforts will also focus on improving query accessibility and user-friendliness. These advances will establish the technical foundation for the ShareFAIR framework.

Availability and implementation

All resources are openly available: **ShareFAIR-KG** with examples of queries and competency questions ([GitLab](#), v1.0.0, GNUGPLv3.0, [swh:1:dir:4e761e3](#)), **EDAMAnnot** ([GitHub](#), GNUGPLv3.0, [swh:1:dir:bf911b6](#)), **BioFlowInsight** ([GitLab](#), GNUAGPLv3.0), **NeuroWF** ([GitLab](#), MIT License, [swh:1:dir:467be46](#)), and **CoPaLink** ([tool](#) and [experiment](#) repositories, GNUAGPLv3.0) are hosted on GitLab and GitHub with persistent Software Heritage archives; the **knowledge graph content** ([Zenodo](#), DOI:[10.5281/zenodo.17737888](#)) and **BioFlow-Ontology** ([Zenodo](#), v7, CC-BY4.0) are available on Zenodo.

Acknowledgments

This work was supported in part by the National Research Agency under the France 2030 program, with reference to ANR-22-PESN-0007 (ShareFAIR project).

Generative AI Statement

During the preparation of this work, the authors used Kimi K2 (Moonshot AI) and Claude Sonnet 4.5 to improve textual clarity through reformulation, as well as to verify grammar and spelling. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Citations

References

- [1] Cohen-Boulakia S, Belhajjame K, Collin O, Chopard J, Froidevaux C, Gaignard A, et al. Scientific Workflows for Computational Reproducibility in the Life Sciences: Status, Challenges and Opportunities. *Future Generation Computer Systems*. 2017 Oct;75:284-98.
- [2] Botvinik-Nezer R, Holzmeister F, Camerer CF, Dreber A, Huber J, Johannesson M, et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*. 2020 Jun;582(7810):84-8. Available from: <https://doi.org/10.1038/s41586-020-2314-9>.
- [3] Djaffardjy M, Marchment G, Sebe C, Blanchet R, Belhajjame K, Gaignard A, et al. Developing and reusing bioinformatics data analysis pipelines using scientific workflow systems. *Computational and Structural Biotechnology Journal*. 2023 Jan;21:2075-85. Available from: <https://www.sciencedirect.com/science/article/pii/S2001037023001010>.

- [4] Goble C, Cohen-Boulakia S, Soiland-Reyes S, Garijo D, Gil Y, Crusoe MR, et al. FAIR Computational Workflows. *Data Intelligence*. 2020;2(1-2):108-21. Available from: https://doi.org/10.1162/dint_a_00033.
- [5] Wilkinson SR, Aloqalaa M, Belhajjame K, Crusoe MR, de Paula Kinoshita B, Gadelha L, et al. Applying the FAIR principles to computational workflows. *Scientific Data*. 2025;12(1):328.
- [6] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data*. 2016 Mar;3(1):160018.
- [7] Barker M, Chue Hong NP, Katz DS, Lamprecht AL, Martinez-Ortiz C, Psomopoulos F, et al. Introducing the FAIR Principles for Research Software. *Scientific Data*. 2022 Oct;9(1):622.
- [8] Di Cosmo R, Granger S, Hinsén K, Jullien N, Le Berre D, Louvet V, et al. CODE beyond FAIR: a roadmap for reusable research software. *Scientific Data*. 2025;In Press.
- [9] Gustafsson OJR, Wilkinson SR, Bacall F, Pireddu L, Soiland-Reyes S, Leo S, et al.. WorkflowHub: a registry for computational workflows. *arXiv*; 2024. ArXiv:2410.06941 [cs]. Available from: <http://arxiv.org/abs/2410.06941>.
- [10] Ison J, Rapacki K, Ménager H, Kalaš M, Rydzka E, Chmura P, et al. Tools and data services registry: a community effort to document bioinformatics resources. *Nucleic Acids Research*. 2016 Jan;44(D1):D38-47. Available from: <https://doi.org/10.1093/nar/gkv1116>.
- [11] Ison J, Ienasescu H, Chmura P, Rydzka E, Ménager H, Kalaš M, et al. The bio.tools registry of software tools and data resources for the life sciences. *Genome Biology*. 2019 Aug;20(1):164. Available from: <https://doi.org/10.1186/s13059-019-1772-6>.
- [12] Schmit M, Le Clanche U, Marchment G, Cohen-Boulakia S, Dameron O, Gaignard A, et al. A Standards-Based Knowledge Graph that Bridges Scientific Workflows, Run-Time Provenance, and Tool Registries. In: *SWAT4HCLS 2026*. Amsterdam, Netherlands; 2026. Available from: <https://hal.science/hal-05517640>.
- [13] Gaignard A, Skaf-Molli H, Belhajjame K. Findable and reusable workflow data products: A genomic workflow case study. *Semantic Web*. 2020;11(5):751-63. Available from: <https://journals.sagepub.com/doi/abs/10.3233/SW-200374>.
- [14] Leo S, Crusoe MR, Rodríguez-Navas L, Sirvent R, Kanitz A, De Geest P, et al.. Recording Provenance of Workflow Runs with RO-Crate. *arXiv*; 2024.
- [15] Amstutz P, Crusoe MR, Tijanić N, Chapman B, Chilton J, Heuer M, et al.. Common Workflow Language, v1.0; 2016. Publisher: figshare. Available from: <https://research.manchester.ac.uk/en/publications/common-workflow-language-v10/>.
- [16] Crusoe MR, Abeln S, Iosup A, Amstutz P, Chilton J, Tijanić N, et al. Methods included: standardizing computational reuse and portability with the common workflow language. *Communications of the ACM*. 2022;65(6):54-63.
- [17] Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. *Nature Biotechnology*. 2017 Apr;35(4):316-9. Publisher: Nature Publishing Group. Available from: <https://www.nature.com/articles/nbt.3820>.
- [18] The Galaxy Community. The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2022 Update. *Nucleic Acids Research*. 2022 Jul;50(W1):W345-51.
- [19] Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. 2016 06;32(19):3047-8. Available from: <https://doi.org/10.1093/bioinformatics/btw354>.
- [20] Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. *Nature methods*. 2016;13(7):581-3.
- [21] Lopez-Delisle L. github.com/iwc-workflows/atacseq/main. Zenodo; 2025. Available from: <https://doi.org/10.5281/zenodo.15078780>.
- [22] Marchment G, Brancotte B, Schmit M, Lemoine F, Cohen-Boulakia S. BioFlow-Insight: facilitating reuse of Nextflow workflows with structure reconstruction and visualization. *NAR Genomics and Bioinformatics*. 2024 08;6(3):lqae092. Available from: <https://doi.org/10.1093/nargab/lqae092>.

- [23] Marchment G, Brancotte B, Cohen-Boulakia S, Gura J, Lemoine F. MetroFlow : Generating Interactive Metro-Maps from Workflow Code. Nextflow Summit 2025. 2025. Available from: <https://www.youtube.com/watch?v=mgR7wQIGUDo>.
- [24] Le Clanche U, Cohen-Boulakia S, Cunff YL, Dameron O, Gaignard A. Assessing bioinformatics software annotations: bio.tools case-study. In: Proceedings JOBIM. Bordeaux & Online, France; 2025. p. 1-7. Available from: <https://inria.hal.science/hal-05096849>.
- [25] El Garb M, Coquery E, Duchateau F, Lumineau N. Improving reproducibility in bioinformatics workflows with BioFlow-Model. In: ACM Conference on Reproducibility and Replicability (ACM REP '25),. Vancouver, Canada: ACM; 2025. Available from: <https://hal.science/hal-05240352>.
- [26] Fakih G, Serrano-Alvarado P. A survey on SPARQL query relaxation under the lens of RDF reification. Semantic Web. 2024;15(6):2507-54.
- [27] Sebe C, Ferret O, Névél A, Esmailoghli M, Leser U, Cohen-Boulakia S. Supporting Workflow Reproducibility by Linking Bioinformatics Tools across Papers and Executable Code. arXiv; 2026. ArXiv:2603.08195 [cs]. Available from: <http://arxiv.org/abs/2603.08195>.
- [28] Atay MS, Clenet B, Bannier E, Maumet C. Real-world parameter preferences in task-fMRI pre-processing pipelines. In: OHBM 2026 - Annual Meeting of the Organization for Human Brain Mapping. Bordeaux, France: OHBM; 2026. Available from: <https://inria.hal.science/hal-05545909>.