EOSC-Life: Building a digital space for the life sciences

D2.3 — Implementation of a mechanism for publishing and sharing workflows across instances of the environment

WP2 – Tools Collaboratory: Deployment of life-science data integration and analysis workflows in EOSC
Lead Beneficiaries: CRMMP & EMBL (WP2) / UNIMAN (D2.3)
WP2 leader: Antonio Rosato and Jean-Karim Hériché
Contributing partner(s): UNIMAN, CRS4

Authors of this deliverable: **Carole Goble, Stian Soiland-Reyes, Finn Bacall, Stuart Owen (UNIMAN), Luca Pireddu, Simone Leo (CRS4)**

Contractual delivery date: **28 Feb 2023**
Actual delivery date: **2 May 2023**
H2020-INFRAEOSC-2018-2

Grant agreement no. 824087
Horizon 2020
Type of action: RIA

Table of Contents

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

Executive Summary

Just as data is expected to be FAIR (Findable, Accessible, Interoperable, Reusable), so are workflows, specialised forms of software that encode processing pipelines and analytical know-how. Sharing workflows opens them to review and reuse, supports their citability with due author credit and attribution, and provides steps towards reproducibility of the data results, with accurately documented method and provenance lineage of derived data. Reusing workflows, including adapting them for new data, parameters, or steps, improves the sharing of processing know-how, and improves productivity and reliability, saving labour. Using workflows as an access point to the tools and computational resources of EOSC democratises the use of computational platforms and the availability of computational methods that are beyond the skills of a large body of researchers. The production of workflows is labour-intensive and a highly skilled task, and as such they are important assets to be exchanged in EOSC. Across the BMS RIs, workflows are scattered in private or platform-specific repositories and hidden in containers, data repositories or supplementary materials.

The aim of this deliverable is to overcome this fragmentation without losing autonomy of the workflow developers and users by creating WorkflowHub[1] [1], a unifying workflow registry with rich metadata capture for findability and to support value-added services for workflow testing using the LifeMonitor[2] service. WorkflowHub is workflow manager agnostic, supporting workflows in their native repositories and instances and links to registries such as bio.tools, to support discovery over the fragmented EOSC ecosystem. To date 319 workflows have been registered from 14 different workflow managers: the most popular are Galaxy, Nextflow, Snakemake, Python Scripts and Jupyter Notebooks. 229 workflows are linked to an external URL, 71 of which are linked to a git repository. However, it also supports the manual uploading and storage of files, and as such also acts as a repository: currently, 90 workflows are uploaded as files. The registry supports a common API to simplify access for tool developers and uses the GA4GH TRS API for supporting the direct execution of workflows. It also supports workflow snapshot preservation, DOI publishing, citation and monitoring, partnering with DataCite in the scholarly communication landscape.

The Hub's emphasis is on machine- and human-readable metadata to drive FAIR capability, requiring a new workflow metadata framework to be developed and adopted. Standardised workflow identifiers and metadata descriptions support workflow discovery, reuse, preservation, interoperability and monitoring and metadata harvesting using standard protocols. As workflows are multi-component objects (requiring links to test data, example runs, explanatory documentation, etc.) we use the RO-Crate specification for packaging workflows, which is an implementation mechanism for FAIR Digital Objects as envisioned by the EOSC Interoperability Framework. These digital objects are used to exchange workflows, their metadata and their companion objects (e.g., workflow tests) between the Hub, Workflow Managers and workflow services such as LifeMonitor, and to deposit workflows in EOSC long-term preservation archives such as Zenodo.

---

[1] https://workflowhub.eu/

[2] https://lifemonitor.eu/

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

The WorkflowHub also acts as a Community of Practice Hub. 435 contributors are registered from 136 organisations, organised into 155 teams, and it features support workflow subscription and access to community channels.

WorkflowHub has been adopted by several EOSC-Life partners, most notably ELIXIR, by other European projects and centres such as BioExcel, and by Research Infrastructures outside of BMS, notably DISSCo and SKA. The Hub has also been enthusiastically adopted by the Australian BioCommons who are our development partners. WorkflowHub is a registered service in the EOSC Service Catalogue and Marketplace[3].

The WorkflowHub governance is operated by the WorkflowHub Club under the sponsorship of ELIXIR and the Australian BioCommons. An extensive number of follow-on projects from Horizon Europe have secured its sustainability for the next 5 years.

Project Objectives

This deliverable has contributed to the following project objectives:

a. Establish EOSC-Life by publishing FAIR life science data resources for cloud use.
    - Developed the EOSC-Life Workflow Collaboratory which provides a framework for researchers and workflow specialists to find, use and reuse workflows.
    - Supported the adoption of the Workflow Collaboratory and its metadata framework in key Workflow Managers used by the community.
    - Developed and operated the WorkflowHub workflow registry to enable FAIR Computational Workflows discoverability, sharing and publishing, and in some cases execution.
    - Developed and operated services to support the FAIRness of workflows including the LifeMonitor service for monitoring workflow execution viability and supporting the adoption of workflow maintenance best practices.

b. Create an ecosystem of innovative life-science tools in EOSC.
    - Designed and developed the EOSC-Life Workflow Collaboratory of services for supporting FAIR Computational workflows operating in EOSC.
    - Designed and deployed a workflow metadata framework including the adoption of RO-Crate[4], a community-developed standardised approach for research output packaging with rich metadata, an implementation approach for FAIR Digital Objects as proposed by the EOSC Interoperability Framework.
    - The WorkflowHub is a registered service of the European Open Science Cloud.

c. Enable ground-breaking data driven research in Europe by connecting life scientists to EOSC.
    - Provide a route to EOSC services and computing resources through workflow platform instances on EOSC.
    - Workflows registered by 5 BMS RIs and 5 / 8 EOSC-Life WP3 Demonstrator projects.

---

[3] https://marketplace.eosc-portal.eu/datasources/eosc.elixir-uk.5126ffcc8e23f65bbbe219d36128f2c8

[4] https://www.researchobject.org/ro-crate/

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

    - 319 workflows registered (42 associated with COVID-19), 8 explicit collections of workflows, including SARS-CoV-2 monitoring[5];
    - Since its launch the Hub has had **389,253** views and **5837** downloads. Accesses through the WorkflowHub APIs are not logged.
    - Part of the European COVID-19 Data Portal.
    - **435** registered people from **136** organisations, organised into **155** teams.
    - Adopted by Australian BioCommons and RIs outside BMS.

Specifically, this deliverable has contributed to WP2 objectives:

2. Make BMS RI tools ready for deployment in the EOSC following FAIR principles.
    - Supported the integration of the Bio.tools registry for BMS tools into the EOSC-Life Workflow Collaboratory.
    - Advised on how to incorporate tools into workflows.
3. Drive implementation of workflows that cross disciplines and RI boundaries.
    - The WorkflowHub registry is subject neutral, and already contains workflows for astronomy, collection digitisation, particle physics, biodiversity and climate modelling.
    - Introduced a Workflow Collaboratory and metadata framework that is workflow manager and discipline agnostic.
4. Promote and develop use of the EOSC by fostering tool-focused collaborations and sharing between communities via cloud implementation and sharing of workflows that integrate the data and data analysis methods of different BMS RI.
    - Developed the EOSC-Life Workflow Collaboratory which provides a framework for researchers and workflow specialists to find, use and reuse workflows.
    - Developed and operated the WorkflowHub workflow registry to enable FAIR Computational Workflows discoverability, sharing and publishing, and in some cases execution.

Detailed Report on the Deliverable

1. Introduction

The practice of performing computational processes using workflows has taken hold in the biosciences as the discipline becomes increasingly computational. Workflows are popular specialist software that handle the complexity of linking data flows across different codes and executing them on different computing platforms. The COVID-19 pandemic spotlighted the importance of systematic and shared analysis of SARS-CoV-2 and its data processing pipelines. There is also a drive in the community towards adopting FAIR practices (Findable, Accessible, Interoperable, and Reusable) not just for data, but also for workflows, and to improve the reproducibility of computational processes.

---

[5] https://workflowhub.eu/collections/2

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

EOSC-Life brought together 13 of the Life Science 'ESFRI' research infrastructures to create an open, digital, and collaborative space for biological and medical research to develop a cloud-based **Workflow Collaboratory** [2] to drive implementation of FAIR computational workflows [3] across disciplines and RI boundaries, and foster tool-focused collaborations and reuse between communities via the sharing of data analysis workflows.

EOSC-Life is made up of established research infrastructures, ranging from those specialized in biobanking and clinical trial management, through to others coordinating biomedical imaging and phenotyping plants, and conducting 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:

- popular cross-discipline platforms for tool chaining such as Galaxy, Snakemake, and Nextflow or job scheduling such as PyCOMPS;
- more specialist, subject-specific systems such as SCIPION for structural bioinformatics;
- workflow-like platforms such as executable notebooks – e.g. Jupyter Notebooks;
- general scripting languages, such as R and Python.

The Common Workflow Language (CWL) Community currently lists over 340 different workflow platforms of various kinds[6]. Given the diversity of the EOSC-Life community, the Workflow Collaboratory and its workflow sharing and publishing capability must be system agnostic, flexible and extensible.

Mechanisms for finding, sharing, and publishing workflows currently include the following.

1. Using a workflow management system, such as Galaxy, CWL, Nextflow, or Snakemake. This not only allows for the creation, execution, and sharing of reproducible workflows, but also has community congregation around curated repositories (e.g. InterGalactic Workflow Commission[7], CWL Viewer[8], nf-core[9], Snakemake catalogue[10]) and shared cloud-based execution platforms (e.g. GalaxyEurope, Nextflow Tower, BinderHub, though these are platform-specific).
2. Sharing workflows on platforms like GitHub or GitLab, which allow for version control and collaboration, which are also almost universally the platforms used by community repositories.
3. Depositing and publishing workflows in a public generalist repository such as Zenodo or DataVerse, which allows for their discovery and citation, sometimes with some integration with GitHub.
4. Depositing containerized workflows in container repositories or registries such as DockStore[11]. Container frameworks such as Docker or Singularity package the entire workflow environment and its dependencies, making it easy to share and run on different platforms.
5. Publishing workflows in conventional journals (e.g. as pioneered by GigaScience) or those specifically intended for software publications (e.g. the Journal of Open Science Software

---

[6] https://s.apache.org/existing-workflow-systems

[7] https://github.com/galaxyproject/iwc

[8] https://view.commonwl.org/workflows

[9] https://nf-co.re/

[10] https://snakemake.github.io/snakemake-workflow-catalog/

[11] https://dockstore.org/

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

(JOSS)) with links to repositories as supplementary materials, which allows for their citation, but with limited discoverability.

6. Register the workflow in a registry, such as Knime Community Hub or BinderHub, which are typically platform-specific.
7. Search using a public search engine such as Google or Bing. This requires good machine-processable metadata markup.

To serve the needs of this established and diverse community, EOSC-Life has developed the **WorkflowHub** [1] as an inclusive workflow registry, capable of supporting *any* Workflow Management System (WfMS) or workflow-like environments, with a focus on FAIR compliance metadata, and integration with added-value services and community repositories. Although developed by EOSC-Life, the registry is subject neutral, and already contains workflows for astronomy, collection digitisation, particle physics, biodiversity and climate modelling.

As an umbrella registry, the Hub makes workflows Findable and Accessible by indexing them across workflow management systems and their native repositories, while providing rich standardised metadata. Interoperability and Reusability are supported by standardised descriptions of workflows and packaging of workflow components, developed in close collaboration with the communities participating in EOSC-Life and also beyond. 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.

Links to other work packages

- WP1: contributed workflow metadata framework and the RO-Crate for FAIR data;
- WP3: provided support for registering workflows for WP3 demonstrator projects and open calls;
- WP6: workflow metadata framework and LifeMonitor services contributed to workflow provenance and FAIR metadata;
- WP9: contributions to workflow training, notably the “defrag” imaging training school funded by WP9 open call.

Related Deliverables

- Antonio Rosato, Jean-Karim Hériché (eds) (2022): **EOSC-Life A common environment that can run cross-RIs workflows.** EOSC-Life deliverable D2.2, *Zenodo*. https://doi.org/10.5281/zenodo.7217294

2. Description of Work

The EOSC-Life Workflow Collaboratory

The Workflow Collaboratory provides a framework for researchers and workflow specialists to find, use and reuse workflows (Figure 1). In this framework, WorkflowHub aims to incorporate

This project has received funding from the *European Union’s Horizon 2020 research and innovation programme* under grant agreement No 824087.

workflows, in partnership with the WfMS developers and communities, to embed the registration of workflows in the community processes by integrating with:

- pre-existing workflow repositories – typically in Git, GitLab and GitHub, supporting registration and auto-updating; and
- pre-existing workflow management systems for metadata collection and supporting access to workflow execution platforms.

Consequently, the registry adopts common practices (e.g. use of GitHub repositories), and supports integration with an ecosystem of workflow execution platforms assisted by complementary registries (e.g. bio.tools, biocontainers), services for testing and benchmarking workflows (e.g. LifeMonitor, OpenEBench), and generalist repositories and services for research outputs in EOSC (e.g. Zenodo, OpenAIRE Research Graph, DataCite PID Graph).

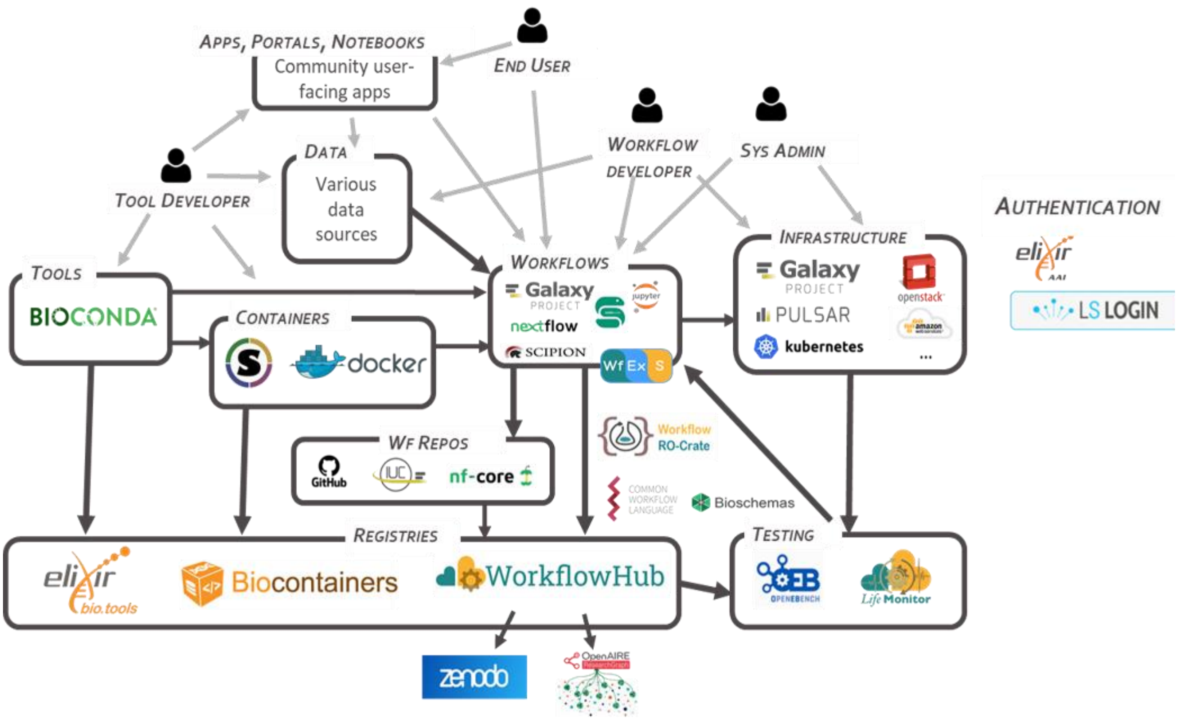


Figure 1: The EOSC-Life workflow collaboratory

COVID-19

In April 2020, the bioscience community, including EOSC-Life, ramped up to address the COVID-19 challenge, including the large-scale processing of SARS-COV-2 data using computational workflows and scripts for automated data analysis and for updating public data archives.

Although WorkflowHub was road mapped for development in 2020 and for a pilot release in 2021, we recognised the urgent need to share openly, find and reuse the COVID data processing pipelines being developed across Europe. Consequently, we considerably accelerated the work, releasing a Minimal Viable Product version after an April 2020 COVID Hackathon; an Alpha version in June 2020 and a beta version in September 2020. We finally removed the beta designation in September 2022. This extreme open development approach was stressful on the development team but returned benefits by enabling the registry to be immediately usable and allowing close ties to be forged with the developer and user communities in the RIs, outside the RIs, and with WP2 tasks 2.1 and 2.5 and Work packages (WP1, WP3 and WP6).

WorkflowHub Club

To convene an inclusive community of developers and users we established the WorkflowHub Club[12], which meets biweekly online. WorkflowHub evolved into a collaboration between several projects[13], beyond EOSC-Life, and any contributors are welcome to join our open community. Over 60 people are listed in the acknowledgements section[14].

The Club includes several members of EOSC-Life's Research Infrastructures: ELIXIR, MIRRI-ERIC, INSTRUCT-ERIC, ISBE, BBMRI-ERIC, EMBRC-ERIC, and EuroBioImaging-ERIC.

All minutes are open. Other communication channels include a mailing list, Google Drive folder, Slack channel, and GitHub[15] organisation, which lists 14 contributors. Through the WorkflowHub Club we have widened our contributor community; for example, the user guides have been contributed by the Australian BioCommons.

The Hub was developed openly and largely virtually, using open software development practices, hackathons, and virtual communication channels. It has a roadmap and regular release cycle.

The WorkflowHub governance is operated by the WorkflowHub Club under the sponsorship of ELIXIR and the Australian BioCommons.

WorkflowHub

The WorkflowHub is a central place to find and share workflows regardless of their type or where they are located. The registry is now well established. It is a registered service in the EOSC Service Catalogue and Marketplace[16]; an ELIXIR service sponsored by two ELIXIR Nodes (UK, Belgium) and a service of the ELIXIR Tools ecosystem[17]. The Hub has also been adopted by the Australian BioCommons[18] and is part of the European COVID-19 Data Portal[19].

The WorkflowHub is built on top of the FAIRDOM-SEEK[20] platform and inherits several properties from that platform, notably the support of sharing permissions whereby content can be registered, and its visibility limited to networks. The content of WorkflowHub is openly available

---

[12] https://about.workflowhub.eu/project/community/
[13] https://about.workflowhub.eu/project/acknowledgements
[14] https://about.workflowhub.eu/project/acknowledgements/ - workflowhub-club
[15] https://github.com/workflowhub-eu/about
[16] https://marketplace.eosc-portal.eu/datasources/eosc.elixir-uk.5126ffcc8e23f65bbbe219d36128f2c8
[17] https://elixir-europe.org/platforms/tools
[18] https://australianbiocommons.github.io/2_1_workflows.html
[19] https://www.covid19dataportal.org/related-resources
[20] https://seek4science.org/

to access without login. Contribution to the Hub requires registration and login to enable the tracking of ownership, credit, and citations.

To date, the Hub has:

- **303** workflows visible, out of a total of **319,** 42 of which are tagged with the term COVID.
- **8** explicit collections of workflows, including SARS-CoV-2 monitoring[21] which has over 700 views.

Since its launch, the Hub has had **389,253** views and **5837** downloads. Accesses through the WorkflowHub APIs are not logged so this is a significant underestimate.

Communities of practice are organised as Spaces and Teams. The workflows are curated by their creators and submitters. There are no centralised or additional resources to curate the Hub.

To date, the Hub has:

- **435** registered people from **136** organisations
- **21** spaces and **155** teams. Spaces are an umbrella grouping for teams; all resources are owned by one or more Teams. “Independent Team” is the default space with 109 teams currently under its umbrella.

The Hub is actually a hybrid. As a registry it is designed to have rich metadata and link to workflows held in their native repositories. However, it also supports the manual uploading and storage of files, and as such also acts as a repository.

Currently:

- **229** workflows are linked to an external URL, 71 of which are linked to a git repository.
- **90** workflows are uploaded files.

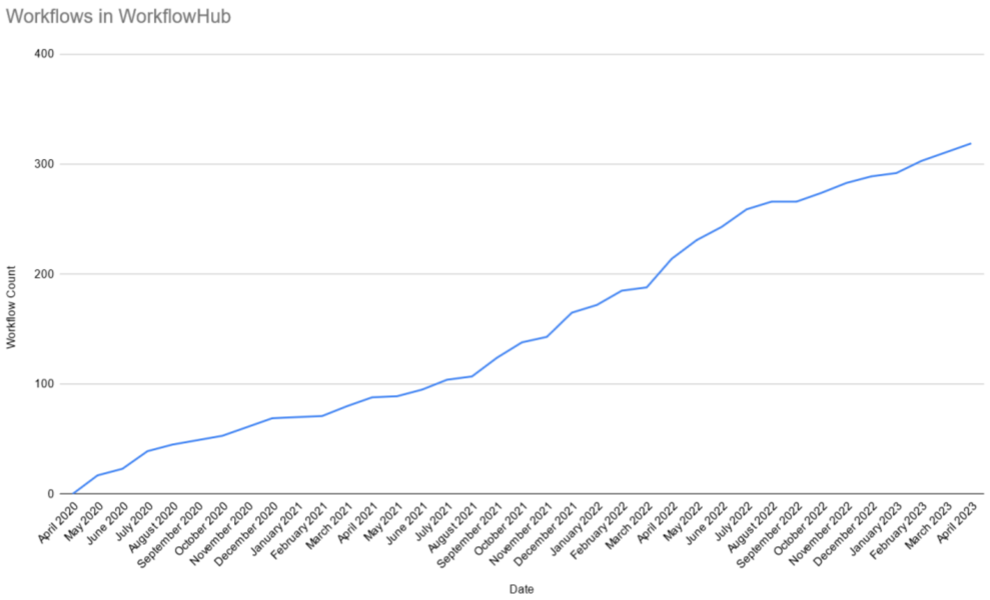


Figure 2: The growth of workflow deposition in WorkflowHub (05/04/2023)

---

[21] https://workflowhub.eu/collections/2

This project has received funding from the *European Union’s Horizon 2020 research and innovation programme* under grant agreement No 824087.

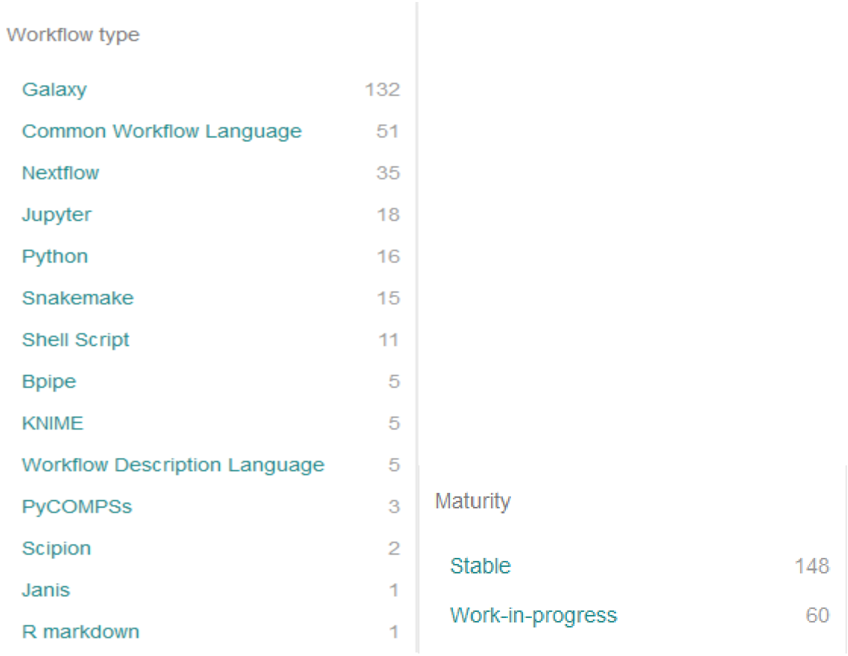
WorkflowHub is intended to be workflow system agnostic, that is, onboard workflows of all kinds, though popular workflow systems are supported with additional features. Galaxy and CWL have been close development partners, and dedicated support for Snakemake and Nextflow is planned. Additional features include:

- Semi-automated metadata extraction for workflow registration;
- Customised support for community repositories;
- RO-Crate and Abstract CWL support by the workflow manager.

To date, **18 different workflow types** are supported, with 14 represented by registered workflows (Figure 3).

Galaxy is by far the most registered, and the WorkflowHub team have worked closely with Galaxy for seamless integration and additional features, including semi-automated workflow registration (including workflow updates) from the Galaxy Intergalactic Workflow Collection[22], and semi-automated registration and test monitoring in LifeMonitor.

The Common Workflow Language is second most popular, and again WorkflowHub has special support for parsing CWL. The next are Nextflow, Jupyter Notebooks, Python scripts and Snakemake.



Workflow type	
Galaxy	132
Common Workflow Language	51
Nextflow	35
Jupyter	18
Python	16
Snakemake	15
Shell Script	11
Bpipe	5
KNIME	5
Workflow Description Language	5
PyCOMPSs	3
Scipion	2
Janis	1
R markdown	1

Maturity	
Stable	148
Work-in-progress	60

Figure 3: Workflow types registered and maturity categorisation (05/04/2023)

An indication of maturity of the workflow can be assigned by the creators. Workflows that are work in progress are encouraged, not just “finished” workflows.

---

[22] https://github.com/galaxyproject/iwc

Other features include:

Workflow registration and update management

- manual by file upload; import from git repository; import by RO-Crate digital object[23], also via REST API;
- metadata extraction for selected workflow managers (Galaxy, Snakemake, CWL);
- workflow versioning support.

Git Repo link up

- as many workflows are developed in Git repositories, WorkflowHub supports import from a Git[24]. New versions of workflows hosted on GitHub can be automatically uploaded to the Hub by the LifeMonitor GitHub app; in other cases, they currently have to be uploaded manually.
- Currently 71 workflows are linked to external git repositories.

Workflow publishing, citation and contribution to CRIS infrastructures

- DOI assignment to workflows for citation and publication using DataCite;
- author credit, gathered from citation.cff files and manual uploads;
- licensing information for workflows;
- analytics on views and downloads;
- digital objects (RO-Crate) deposition into Zenodo.
- contribution to the DataCite PID Graph, and the DataCite MDSchema used by OpenAire and others for OAI-harvesting MD from other MD Aggregators[25].

Workflow discovery

- tag and keyword search
- faceted browsing on tags and other properties with sorting by dates, titles and analytics;
- subscription to workflows to be notified of changes;
- workflows organised into collections;
- discovery by a workflow's attribution.

Workflow companion objects

- registering documents, data files and SOPs;
- linking workflows to associated data, documents, SOPs and electronic lab notebooks;
- linking workflows to other workflows that they may be derived from (attribution).

Community of Practice support

- spaces and teams for organising people and sharing workflows;
- registering events and presentations;
- support for community channels linking developers with users;
- explicitly curated collections.

Access options to the workflows from the WorkflowHub entry (see Figure 4) include:

- View on Git (linked to their repository);
- Download RO-Crate (which packages the files, references and metadata for a workflow);

---

[23] https://about.workflowhub.eu/docs/adding-files/

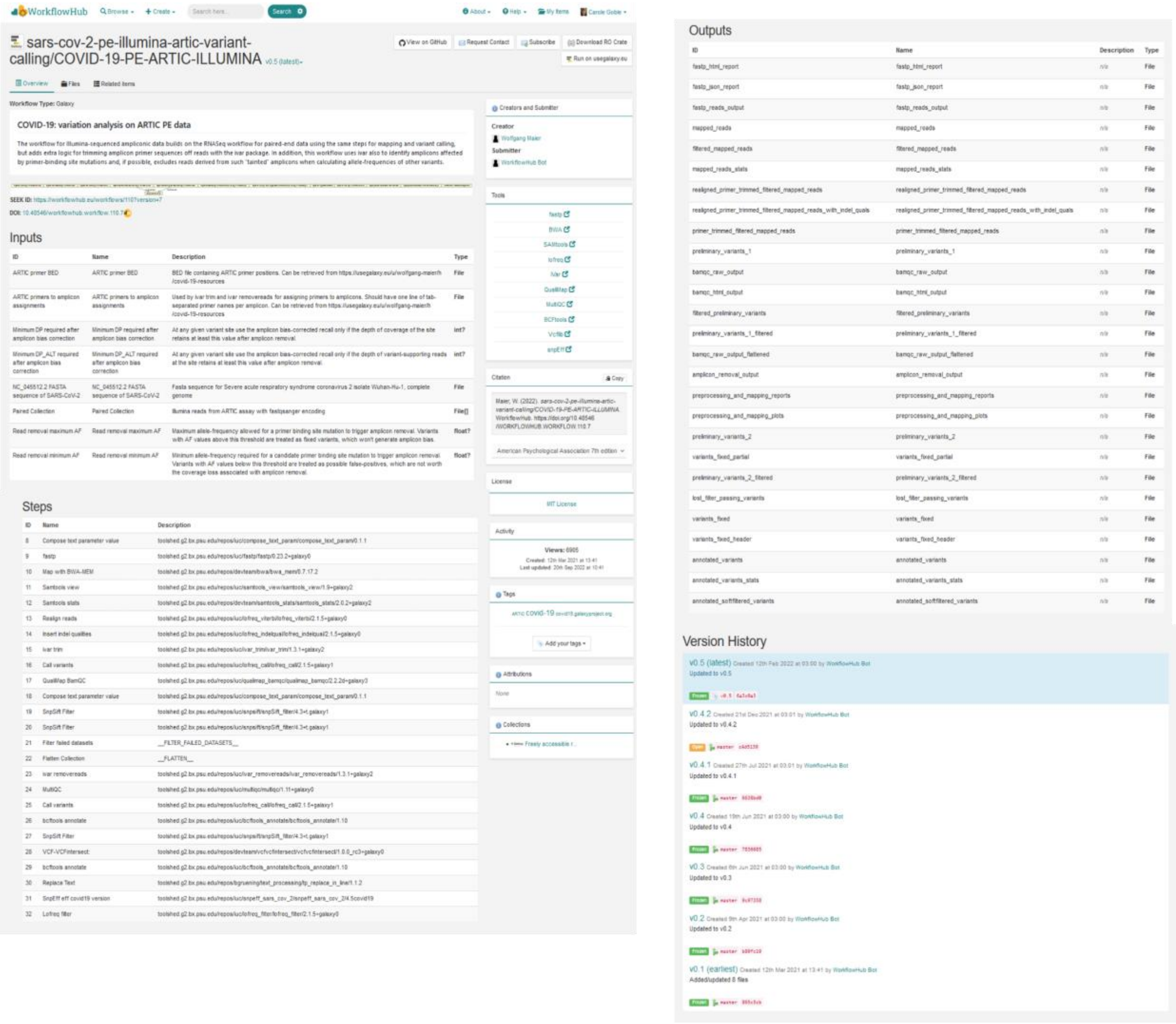
[24] https://about.workflowhub.eu/docs/adding-files/#git

[25] https://guidelines.openaire.eu/en/latest/data/index.html

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

- Visit site (for workflows that have a Gitlab repository or other site);
- Request Contact (with registered workflow developers);
- Run on a Galaxy Instance such as UseGalaxy.eu (launch using the TRS API);
- Request to REST API.

Access is also possible through communication channels (for example, Binder - See Figure 10).



sars-cov-2-pe-illumina-artic-variant-calling/COVID-19-PE-ARTIC-ILLUMINA v0.5 (latest)

Workflow Type: Galaxy

COVID-19: variation analysis on ARTIC PE data

The workflow for Illumina-sequenced ampliconic data builds on the RNASeq workflow for paired-end data using the same steps for mapping and variant calling, but adds extra logic for trimming amplicon primer sequences off reads with the ivar package. In addition, this workflow uses ivar also to identify amplicons affected by primer-binding site mutations and, if possible, excludes reads derived from such "tainted" amplicons when calculating allele-frequencies of other variants.

SEEK ID: https://workflowhub.eu/workflows/110?version=7

DOI: 10.48546/workflowhub.workflow.110.7

Inputs

ID	Name	Description	Type
ARTIC primer BED	ARTIC primer BED	BED file containing ARTIC primer positions. Can be retrieved from https://usegalaxy.eu/u/wolfgang-maier/h/covid-19-resources	File
ARTIC primers to amplicon assignments	ARTIC primers to amplicon assignments	Used by ivar trim and ivar removereads for assigning primers to amplicons. Should have one line of tab-separated primer names per amplicon. Can be retrieved from https://usegalaxy.eu/u/wolfgang-maier/h/covid-19-resources	File
Minimum DP required after amplicon bias correction	Minimum DP required after amplicon bias correction	At any given variant site use the amplicon bias-corrected recall only if the depth of coverage of the site retains at least this value after amplicon removal	int?
Minimum DP_ALT required after amplicon bias correction	Minimum DP_ALT required after amplicon bias correction	At any given variant site use the amplicon bias-corrected recall only if the depth of variant-supporting reads at the site retains at least this value after amplicon removal	int?
NC_045512.2 FASTA sequence of SARS-CoV-2	NC_045512.2 FASTA sequence of SARS-CoV-2	Fasta sequence for Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome	File
Paired Collection	Paired Collection	Illumina reads from ARTIC assay with fastqsanger encoding	File[]
Read removal maximum AF	Read removal maximum AF	Maximum allele-frequency allowed for a primer binding site mutation to trigger amplicon removal. Variants with AF values above this threshold are treated as fixed variants, which won't generate amplicon bias.	float?
Read removal minimum AF	Read removal minimum AF	Minimum allele-frequency required for a candidate primer binding site mutation to trigger amplicon removal. Variants with AF values below this threshold are treated as possible false-positives, which are not worth the coverage loss associated with amplicon removal.	float?

Steps

ID	Name	Description
8	Compose text parameter value	toolshed.g2.bx.psu.edu/repos/iuc/compose_text_param/compose_text_param/0.1.1
9	fastp	toolshed.g2.bx.psu.edu/repos/iuc/fastp/fastp/0.23.2+galaxy0
10	Map with BWA-MEM	toolshed.g2.bx.psu.edu/repos/devteam/bwa/bwa_mem/0.7.17.2
11	Samtools view	toolshed.g2.bx.psu.edu/repos/iuc/samtools_view/samtools_view/1.9+galaxy2
12	Samtools stats	toolshed.g2.bx.psu.edu/repos/devteam/samtools_stats/samtools_stats/2.0.2+galaxy2
13	Realign reads	toolshed.g2.bx.psu.edu/repos/iuc/lofreq_viterbi/lofreq_viterbi/2.1.5+galaxy0
14	Insert indel qualities	toolshed.g2.bx.psu.edu/repos/iuc/lofreq_indelqual/lofreq_indelqual/2.1.5+galaxy0
15	ivar trim	toolshed.g2.bx.psu.edu/repos/iuc/ivar_trim/ivar_trim/1.3.1+galaxy2
16	Call variants	toolshed.g2.bx.psu.edu/repos/iuc/lofreq_call/lofreq_call/2.1.5+galaxy1
17	QualiMap BamQC	toolshed.g2.bx.psu.edu/repos/iuc/qualimap_bamqc/qualimap_bamqc/2.2.2d+galaxy3
18	Compose text parameter value	toolshed.g2.bx.psu.edu/repos/iuc/compose_text_param/compose_text_param/0.1.1
19	SnpSift Filter	toolshed.g2.bx.psu.edu/repos/iuc/snpsift/snpSift_filter/4.3+t.galaxy1
20	SnpSift Filter	toolshed.g2.bx.psu.edu/repos/iuc/snpsift/snpSift_filter/4.3+t.galaxy1
21	Filter failed datasets	__FILTER_FAILED_DATASETS__
22	Flatten Collection	__FLATTEN__
23	ivar removereads	toolshed.g2.bx.psu.edu/repos/iuc/ivar_removereads/ivar_removereads/1.3.1+galaxy2
24	MultiQC	toolshed.g2.bx.psu.edu/repos/iuc/multiqc/multiqc/1.11+galaxy0
25	Call variants	toolshed.g2.bx.psu.edu/repos/iuc/lofreq_call/lofreq_call/2.1.5+galaxy1
26	bcftools annotate	toolshed.g2.bx.psu.edu/repos/iuc/bcftools_annotate/bcftools_annotate/1.10
27	SnpSift Filter	toolshed.g2.bx.psu.edu/repos/iuc/snpsift/snpSift_filter/4.3+t.galaxy1
28	VCF-VCFintersect:	toolshed.g2.bx.psu.edu/repos/devteam/vcfvcfintersect/vcfvcfintersect/1.0.0_rc3+galaxy0
29	bcftools annotate	toolshed.g2.bx.psu.edu/repos/iuc/bcftools_annotate/bcftools_annotate/1.10
30	Replace Text	toolshed.g2.bx.psu.edu/repos/bgruening/text_processing/tp_replace_in_line/1.1.2
31	SnpEff eff covid19 version	toolshed.g2.bx.psu.edu/repos/iuc/snpeff_sars_cov_2/snpeff_sars_cov_2/4.5covid19
32	Lofreq filter	toolshed.g2.bx.psu.edu/repos/iuc/lofreq_filter/lofreq_filter/2.1.5+galaxy0

Creator
Wolfgang Maier
Submitter
WorkflowHub Bot

Tools
fastp
BWA
SAMtools
lofreq
iVar
QualiMap
MultiQC
BCFtools
Vcflib
snpEff

Citation
Maier, W. (2022). sars-cov-2-pe-illumina-artic-variant-calling/COVID-19-PE-ARTIC-ILLUMINA. WorkflowHub. https://doi.org/10.48546/WORKFLOWHUB.WORKFLOW.110.7
American Psychological Association 7th edition

License
MIT License

Activity
Views: 6905
Created: 12th Mar 2021 at 13:41
Last updated: 20th Sep 2022 at 10:41

Tags
ARTIC COVID-19 covid19.galaxyproject.org

Attributions
None

Collections
Freely accessible r...

Outputs

ID	Name	Description	Type
fastp_html_report	fastp_html_report	n/a	File
fastp_json_report	fastp_json_report	n/a	File
fastp_reads_output	fastp_reads_output	n/a	File
mapped_reads	mapped_reads	n/a	File
filtered_mapped_reads	filtered_mapped_reads	n/a	File
mapped_reads_stats	mapped_reads_stats	n/a	File
realigned_primer_trimmed_filtered_mapped_reads	realigned_primer_trimmed_filtered_mapped_reads	n/a	File
realigned_primer_trimmed_filtered_mapped_reads_with_indel_quals	realigned_primer_trimmed_filtered_mapped_reads_with_indel_quals	n/a	File
primer_trimmed_filtered_mapped_reads	primer_trimmed_filtered_mapped_reads	n/a	File
preliminary_variants_1	preliminary_variants_1	n/a	File
bamqc_raw_output	bamqc_raw_output	n/a	File
bamqc_html_output	bamqc_html_output	n/a	File
filtered_preliminary_variants	filtered_preliminary_variants	n/a	File
preliminary_variants_1_filtered	preliminary_variants_1_filtered	n/a	File
bamqc_raw_output_flattened	bamqc_raw_output_flattened	n/a	File
amplicon_removal_output	amplicon_removal_output	n/a	File
preprocessing_and_mapping_reports	preprocessing_and_mapping_reports	n/a	File
preprocessing_and_mapping_plots	preprocessing_and_mapping_plots	n/a	File
preliminary_variants_2	preliminary_variants_2	n/a	File
variants_fixed_partial	variants_fixed_partial	n/a	File
preliminary_variants_2_filtered	preliminary_variants_2_filtered	n/a	File
lost_filter_passing_variants	lost_filter_passing_variants	n/a	File
variants_fixed	variants_fixed	n/a	File
variants_fixed_header	variants_fixed_header	n/a	File
annotated_variants	annotated_variants	n/a	File
annotated_variants_stats	annotated_variants_stats	n/a	File
annotated_softfiltered_variants	annotated_softfiltered_variants	n/a	File

Version History

v0.5 (latest) Created 12th Feb 2022 at 03:00 by WorkflowHub Bot
Updated to v0.5

v0.4.2 Created 21st Dec 2021 at 03:01 by WorkflowHub Bot
Updated to v0.4.2

v0.4.1 Created 27th Jul 2021 at 03:01 by WorkflowHub Bot
Updated to v0.4.1

v0.4 Created 19th Jun 2021 at 03:00 by WorkflowHub Bot
Updated to v0.4

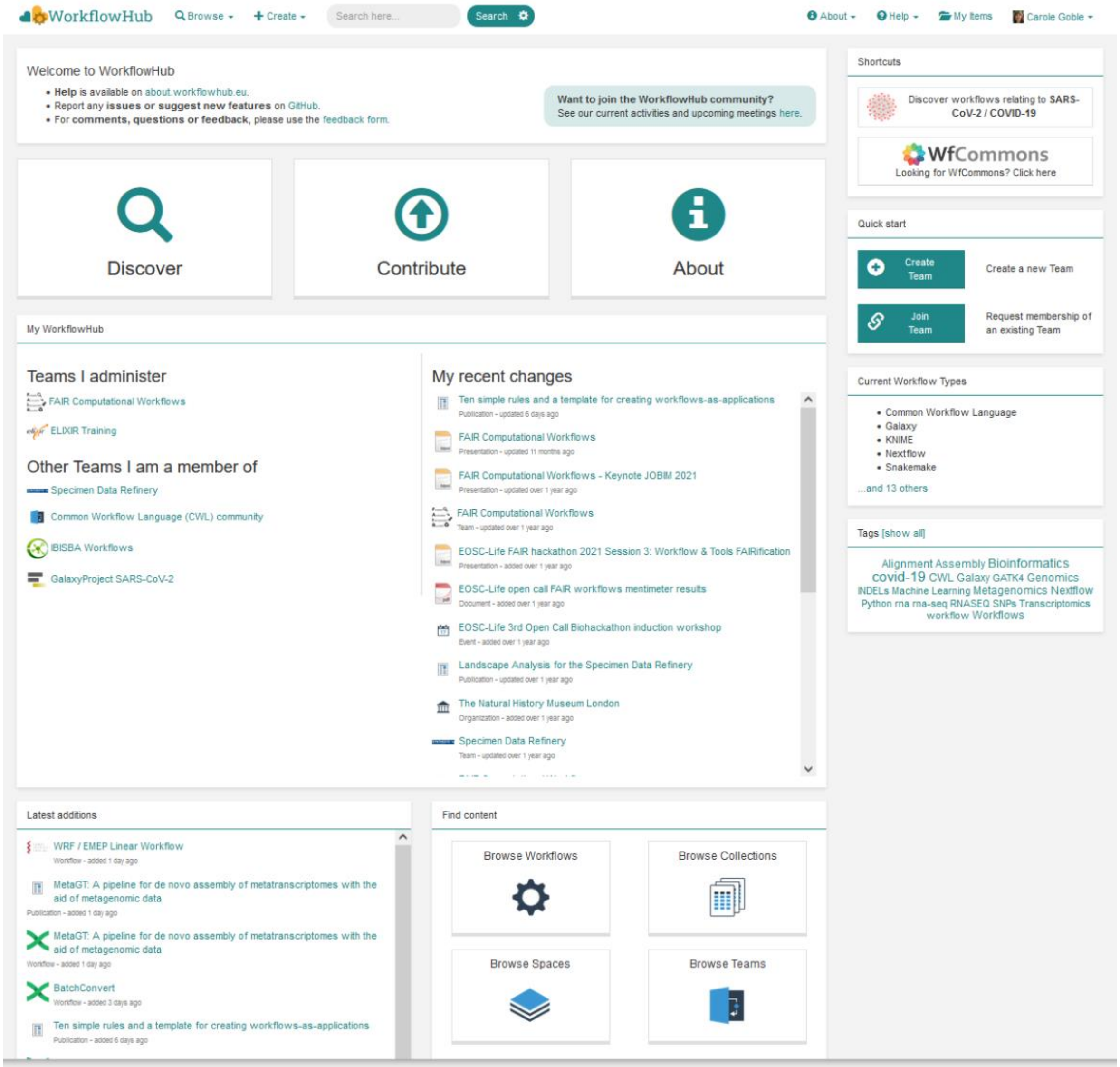
v0.3 Created 6th Jun 2021 at 03:00 by WorkflowHub Bot
Updated to v0.3

v0.2 Created 9th Apr 2021 at 03:00 by WorkflowHub Bot
Updated to v0.2

v0.1 (earliest) Created 12th Mar 2021 at 13:41 by WorkflowHub Bot
Added/updated 8 files

Figure 4: SARS-CoV-2 Workflow using the Galaxy Europe instance

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.



Welcome to WorkflowHub

- **Help** is available on about.workflowhub.eu.
- Report any **issues or suggest new features** on GitHub.
- For **comments, questions or feedback**, please use the feedback form.

Want to join the WorkflowHub community?
See our current activities and upcoming meetings here.

My WorkflowHub

Teams I administer

FAIR Computational Workflows
ELIXIR Training

Other Teams I am a member of

Specimen Data Refinery
Common Workflow Language (CWL) community
IBISBA Workflows
GalaxyProject SARS-CoV-2

My recent changes

Ten simple rules and a template for creating workflows-as-applications
Publication - updated 6 days ago
FAIR Computational Workflows
Presentation - updated 11 months ago
FAIR Computational Workflows - Keynote JOBIM 2021
Presentation - updated over 1 year ago
FAIR Computational Workflows
Team - updated over 1 year ago
EOSC-Life FAIR hackathon 2021 Session 3: Workflow & Tools FAIRification
Presentation - added over 1 year ago
EOSC-Life open call FAIR workflows mentimeter results
Document - added over 1 year ago
EOSC-Life 3rd Open Call Biohackathon induction workshop
Event - added over 1 year ago
Landscape Analysis for the Specimen Data Refinery
Publication - updated over 1 year ago
The Natural History Museum London
Organization - added over 1 year ago
Specimen Data Refinery
Team - updated over 1 year ago

Shortcuts

Discover workflows relating to **SARS-CoV-2 / COVID-19**

WfCommons
Looking for WfCommons? Click here

Quick start

Create a new Team
Request membership of an existing Team

Current Workflow Types

- Common Workflow Language
- Galaxy
- KNIME
- Nextflow
- Snakemake

...and 13 others

Tags [show all]

Alignment Assembly Bioinformatics covid-19 CWL Galaxy GATK4 Genomics INDELs Machine Learning Metagenomics Nextflow Python rna rna-seq RNASEQ SNPs Transcriptomics workflow Workflows

Latest additions

WRF / EMEP Linear Workflow
Workflow - added 1 day ago
MetaGT: A pipeline for de novo assembly of metatranscriptomes with the aid of metagenomic data
Publication - added 1 day ago
MetaGT: A pipeline for de novo assembly of metatranscriptomes with the aid of metagenomic data
Workflow - added 1 day ago
BatchConvert
Workflow - added 3 days ago
Ten simple rules and a template for creating workflows-as-applications
Publication - added 6 days ago

Find content

Figure 5: WorkflowHub Landing page

WorkflowHub Metadata framework

FAIR workflows emphasise the availability and use of machine-processable metadata. The WorkflowHub and EOSC-Life Workflow Collaboratory developed a new metadata framework for workflows.

- Bioschemas[26] schema.org profiles for Computational Tool[27], Computational Workflow[28] and Formal Parameter[29] provide metadata about a workflow and its tools that are discipline independent.

---

[26] https://bioschemas.org/
[27] https://bioschemas.org/profiles/ComputationalTool/1.0-RELEASE
[28] https://bioschemas.org/profiles/ComputationalWorkflow/1.0-RELEASE
[29] https://bioschemas.org/profiles/FormalParameter/1.0-RELEASE

- The Common Workflow Language[30] [4] as a canonical workflow description to accompany native workflow definitions – both *CWL workflows* (which are executable) and *Abstract CWL* (non-executable) description variants. This description presents the structure of workflows, composed of tools and external interface, in an interoperable way across workflow languages. Currently 33 workflows in other formats use Abstract CWL.
- The EDAM Ontology[31] is used to add bioinformatics-specific metadata, such as strong typing of inputs and outputs, within Abstract CWL and Bioschemas annotations. EDAM is also used to describe the overall workflow Topics and Operations (requested by users) to help find workflows. It is possible to search and filter using these properties. Note that the approach is not limited to EDAM, e.g. in terms of attribute names. This supports expansion beyond EDAM in the future to support non-life-sciences - although EDAM themselves are also currently going in that direction - topics are starting to support ecology, mathematics, language, physics, chemistry plus many more non-life science terms.
- RO-Crate[32] [5,6,7] a community-developed standardised approach for research output packaging with rich metadata. RO-Crate provides us with 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. RO-Crates are used to import into and export from WorkflowHub and are the currency of exchange between the services of the Workflow Collaboratory.

  Three RO-Crate profiles have been developed for Workflows:
    - Workflow-RO-Crate[33] which WorkflowHub can register and mint;
    - Workflow-Testing-RO-Crate[34] which includes test information for LifeMonitor;
    - Workflow-Run-RO-Crate[35] which includes provenance information after the execution of a workflow.
- Identifiers for all the components: 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.

The WorkflowHub metadata framework and Global Alliance for Genomics and Health APIs enable interoperability with the services of the Collaboratory. The Tools Registry Service API[36] supports the exchange of scientific tools and workflows and enables users to search for and retrieve metadata about registered tools, including the tool's name, version, description, author, input and output parameters, and Docker image details. WorkflowHub implements TRS so that workflow execution platforms that make use of the TRS API can import workflows for execution, without requiring bespoke integration with the source repository, and search WorkflowHub from within the execution platform.

WorkflowHub also supports a JSON-API[37] and RO-Crate Submission API[38].

---

[30] https://www.commonwl.org/
[31] https://edamontology.org/page
[32] https://www.researchobject.org/ro-crate/
[33] https://about.workflowhub.eu/Workflow-RO-Crate/
[34] https://crs4.github.io/life_monitor/workflow_testing_ro_crate
[35] https://www.researchobject.org/workflow-run-crate/
[36] https://ga4gh.github.io/tool-registry-service-schemas/
[37] https://workflowhub.eu/api
[38] https://about.workflowhub.eu/developer/ro-crate-api/

Services and Execution Platforms

EOSC-Life Services that integrate with WorkflowHub include:

- **bio.tools**[39], the ELIXIR registry of software tools, databases and services for bioinformatics and the life sciences: workflow entries on WorkflowHub can be annotated with tool identifiers, either by automatic parsing of the workflow description, or through a tagging mechanism that queries the bio.tools API. Users of WorkflowHub can filter through workflows that use a certain tool, and from a workflow entry that has been annotated with tools, they can click through to the respective bio.tools entry to see extra metadata. bio.tools can link to workflows that use a given tool, to show examples of use.
- **LifeMonitor**[40] EOSC-Life service developed by BBMRI-ERIC (Task 2.5) that ensures the correct operation of the workflow over time through monitoring and triggering automated workflow tests and automated checks on metadata and adherence to best practices on the workflow's source code Git repository. The service supports the maintenance of the Galaxy Intergalactic Workflow Collection (IWC) – the curated collection of Galaxy workflows – by monitoring the test status of the workflows and, in the future, by automatically publishing new or updated IWC workflow releases to WorkflowHub and LifeMonitor.
- **LS-Login**[41] enables researchers to use their home organisation credentials or community or other identities (e.g. Google, Apple, ORCID, LS ID) to sign in and access data and services they need. It also allows service providers to control and manage access rights of their users and create different access levels for research groups or international projects.
- **citation.cff**[42] is a plain text file format used to describe software or research artefacts such as datasets, models, and algorithms, to make it easier for researchers to cite in scholarly publications. It is designed to be machine- and human-readable, making it easy to include in software repositories, code repositories, or other places where research artefacts are shared. By harvesting a citation.cff file from Git repositories, WorkflowHub makes it easier for others to cite their workflow, increasing their visibility and impact.
- **ORCID**[43] for researcher identifiers.

External Services and resources integrate with WorkflowHub include:

- **DataCite**[44] provides persistent identifiers (DOIs) for research data and other research outputs. WorkflowHub uses DataCite DOIs and contributed to the DataCite PID Graph[45].
- **FAIR Signposting**[46] is an approach to make the scholarly web more friendly to machines. It uses Typed Links[47] as a means to clarify patterns that occur repeatedly in scholarly portals. For resources of any media type, these typed links are provided in HTTP Link headers[48]. Workflow

---

[39] https://bio.tools/
[40] https://www.lifemonitor.eu/
[41] https://lifescience-ri.eu/ls-login/
[42] https://citation-file-format.github.io/
[43] https://orcid.org/
[44] https://datacite.org/
[45] https://support.datacite.org/docs/datacite-graphql-api-guide
[46] https://signposting.org/
[47] https://www.iana.org/assignments/link-relations/link-relations.xhtml
[48] http://tools.ietf.org/html/rfc5988

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

is an early implementer[49] using the patterns Bibliographic Metadata[50] and Publication Boundary[51] to signpost the RO-Crate.

Workflow execution platforms that integrate with WorkflowHub and its metadata framework include:

- **Galaxy Europe**[52] implements TRS to allow it to import and execute workflows from WorkflowHub. Galaxy also exports RO-Crates[53] following the Workflow Run Crate profile[54], embedding the input/output data from the workflow history, along with Galaxy log information and the executed Galaxy workflow definitions in several formats including Abstract CWL. The GTN Smörgåsbord 2023[55] will include a training module on RO-Crate and Galaxy[56].
- **Sapporo**[57] back-end service developed by DNA Data Bank of Japan (DDBJ)[58] for executing workflows on a number of execution systems, such as Nextflow, Snakemake, CWL and others by using the GA4GH Workflow Execution Service (WES) API. It also provides an optional web front-end to allow configuration and execution of workflows through a graphical user interface. Sapporo implements TRS to allow it to import and execute workflows from WorkflowHub and provides an endpoint for exporting results of a workflow execution as an RO-Crate.
- **WfExS**[59] back-end service developed by ELIXIR is another example of a service for executing different workflow formats through a common interface. WfExS has a particular focus on executing workflows that handle sensitive data in a secure way. WfExS automatically handles the installation and execution of containers utilised by the steps of a workflow, allowing workflows to be run without prior configuration. WfExS supports import of workflows from WorkflowHub via TRS. In addition, workflows can be provided as Workflow RO-Crates, or directly from a Git repository. Workflow runs can also be exported as RO-Crates.

WorkflowHub Preservation and End of Life

WorkflowHub's sustainability plan is to ensure the availability of its contributions and metadata up to and beyond 2027[60]. If and when the WorkflowHub reaches its end of service after that, the published contributions and metadata will be archived as RO-Crates and made available through a public repository, such as Zenodo, Figshare or another appropriate resource at that time. DOI registrations[61] will in this case be updated to link to the archived deposits.

---

[49] https://signposting.org/adopters/#workflowhub
[50] http://signposting.org/bibliographic_metadata/
[51] http://signposting.org/publication_boundary/
[52] https://usegalaxy.eu/
[53] https://galaxyproject.org/news/2023-02-23-structured-data-exports-ro-bco/
[54] https://www.researchobject.org/workflow-run-crate/profiles/workflow_run_crate
[55] https://gallantries.github.io/video-library/events/smorgasbord3/
[56] https://gallantries.github.io/video-library/modules/ro-crate
[57] https://github.com/sapporo-wes/sapporo
[58] https://ddbj.nig.ac.jp/
[59] https://github.com/inab/WfExS-backend
[60] https://about.workflowhub.eu/project/#retention-and-end-of-life-policy
[61] https://about.workflowhub.eu/project/citable.md

WorkflowHub contribution to FAIR Computational Workflows

Using a registry like WorkflowHub gets you a long way to FAIR Workflows, both data principles [8] and software principles [9] (Table 1).

	Software	Data
A workflow is assigned globally unique and persistent identifiers	F1.2	F1
A workflow’s versions are assigned globally unique and persistent identifiers	F1.2	–
The steps of a workflow maybe assigned distinct identifiers (if using Abstract CWL)	F1.1	–
A workflow is described and indexed with rich metadata	F3	F3
The metadata are FAIR, searchable and indexable	F4	F4
A workflow is retrievable by its identifier using a standardised communications protocol	A1	A1
A workflow is accessible, even when the workflow is no longer available	A2	A2
The metadata is machine and human processable, and the vocabularies are FAIR	F2, F3	F2, F3
Metadata clearly and explicitly include the identifier of the workflow they describe	F4	F4
A workflow is described with a plurality of accurate and relevant attributes, and includes qualified references to other objects, workflows, and its component codes	R1	R1
A workflow is licensed	R1.1	R1.1
A workflow has provenance	R1.2	R1.2
A workflow meets domain-relevant community standards	R1.2, R3	R1.2, R1.3

Table 1: WorkflowHub contribution to FAIR Computational Workflows

This project has received funding from the *European Union’s Horizon 2020 research and innovation programme* under grant agreement No 824087.

Adoption of WorkflowHub

WorkflowHub is a registered service in the EOSC Service Catalogue and Marketplace[62].

Within EOSC-Life Cluster and the Life Science Research Infrastructures:

- **ELIXIR** has fully adopted WorkflowHub. It is an ELIXIR service sponsored by two ELIXIR Nodes (UK, Belgium), a service of the ELIXIR Tools ecosystem[63] and is part of the European COVID-19 Data Portal[64].
- **EuroBioImaging-ERIC** includes WorkflowHub in the bioimage informatics roadmap and registered workflows for OME[65] and IDR[66].
- **IBISBA** uses WorkflowHub as its registry, linked to its own project results IBISBAHub[67].
- **INFRAFRONTIER** two nodes are using WorkflowHub for registering IMPC data analysis workflows[68], [69].
- **INSTRUCT-ERIC** registered SCIPION workflows for cryo-electron microscopy[70].

WP3 Open Demonstrators:

- **NMR**[71] (demonstrator 1)
- **Cryo electron microscopy**[10] (demonstrator 2)
- **MGnify metagenomics**[72] (demonstrator 3)
- **Nucleolus biology image processing**[73] (demonstrator 6)
- **Rare diseases** [74] (demonstrator 7)

For other EOSC-Life work packages

- WP6[75]

Workflow managers and community repositories include:

- **Galaxy** Intergalactic Workflow Commission[76] have a well-established semi-automated pipeline and integration with Galaxy Europe[77].
- **Nextflow**'s nf-core[78]
- **Snakemake Catalogue**[79]

---

[62] https://marketplace.eosc-portal.eu/datasources/eosc.elixir-uk.5126ffcc8e23f65bbbe219d36128f2c8
[63] https://elixir-europe.org/platforms/tools
[64] https://www.covid19dataportal.org/related-resources
[65] https://workflowhub.eu/projects/69
[66] https://workflowhub.eu/collections/4
[67] https://workflowhub.eu/projects/1
[68] https://workflowhub.eu/projects/68
[69] https://workflowhub.eu/projects/67
[70] https://workflowhub.eu/workflows/73
[71] https://workflowhub.eu/projects/14#workflows
[72] https://workflowhub.eu/projects/9
[73] https://workflowhub.eu/workflows/41
[74] https://workflowhub.eu/projects/31
[75] https://workflowhub.eu/projects/36
[76] https://workflowhub.eu/projects/33
[77] https://workflowhub.eu/projects/24
[78] https://workflowhub.eu/projects/15
[79] https://workflowhub.eu/projects/55

- **Common Workflow Language** [80]

Other European Research Infrastructures and from other disciplines include:

- **BioExcel** Center of Excellence for Computational Biomolecular Research[81] which uses the Hub for its BioBB Building blocks and has registered 49 workflows;
- **DISSCO/Lifewatch Biodiversity** H2020 Synthesys+ Specimen Data Refinery[82]. Horizon Europe Projects BioDT[83] & BGE[84] are set up to use WorkflowHub;
- **Square Kilometer Array (SKA) Astronomy** Snakemake workflows[85];
- **Climate change** Galaxy Workflows[86].

Outside Europe

- Adopted by the Australian BioCommons[87,88] who have registered 42 workflows.
- 8 USA organisations have registered but have yet to register workflows.
- 3 Chinese organisations have registered but have yet to register workflows.

Dissemination Summary

WorkflowHub Club and EOSC-Life have been very active promoting the resource. We nominated the Hub for UNESCO Global Call for Best Practices in Open Science[89]. WorkflowHub is a member of the SciCodes[90] Consortium of scientific software registries and repositories.

Additional disseminations on WorkflowHub and/or its metadata framework include:

- Conference and Workshop presentations
    - EOSC Symposium 2021
    - 2nd Workshop of the Data Citation Community of Practice[91] for Earth, Space, and Environmental Sciences (and Related Sciences) 8 June 2021
    - FAIR Convergence Symposium Oct-Dec 2021
    - FAIR Festival 22 June 2021
    - Workshop on Sustainable Software Sustainability (*WoSSS*) 2021, 6-8 October 2021
    - ELIXIR-DE Workflow Workshop
    - IEEE eScience Conference 22 September 2021
    - FAIReScience satellite workshop, IEEE eScience, 20 September 2021
    - CWL Conference 2021
    - EOSC Symposium 2022
    - 1st FAIR Digital Object Conference (FDO2022) 26-28 October 2022

---

[80] https://workflowhub.eu/projects/44
[81] https://workflowhub.eu/programmes/2
[82] https://workflowhub.eu/projects/72
[83] https://biodt.eu/
[84] https://biodiversitygenomics.eu/
[85] https://workflowhub.eu/workflows/141
[86] https://workflowhub.eu/projects/18
[87] https://australianbiocommons.github.io/2_1_workflows.html
[88] https://workflowhub.eu/programmes/8
[89] https://en.unesco.org/feedback/unescos-global-call-best-practices-open-science
[90] http://scicodes.net/
[91] https://agu-data.github.io/DataCitationCoP/2nd-workshop-data-citation

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

- Conference keynotes featuring WorkflowHub as a FAIR Workflow registry
    - DataVerse User Conference, 15 June 2021
    - JOBIM French National Bioinformatics Conference 2021, 8 July 2021
    - German Bioinformatics Conference 2021, 7 September 2021
    - 16th Workshop on Workflows in Support of Large-Scale Science, 15 November 2021
    - SERC Swedish e-Science Research Center Annual Meeting, 13 May 2022
    - Helmholtz Metadata Collaboration Conference 2022, 05 October 2022
    - Irish Computational Biology and Genomics Symposium, 2 December 2022

- Invited presentations at meetings
    - DataVerse User Conference, 15 June 2021
    - Synthesys+ AGM 2021
    - RDA Data Fabric Interest Group, 25 February 2021
    - ELIXIR All Hands meeting 2021
    - ELIXIR Data Platform + Interoperability Platform Workshop, 1 December 2021
    - BioDT meeting, 13 September 2022
    - ELIXIR All Hands meeting 2022
    - Workflow Community Initiative Summit 2021, 2022
    - SciCodes meeting, 17 March 2022

- Webinars
    - GO-FAIR Webinar 16 June 2021
    - FAIRPoints-AMA event on Software and workflows, 21 February 2023

- Training Schools
    - TS1 “Defragmentation” imaging training school, funded by WP9 open call, online 30 September 2022; TS2 co-located with the NEUBIAS conference, Porto, 8-12 May 2023

- Posters
    - ISMB 2020, FAIR Symposium 2020
    - BOSC 2021
    - ECCB 2022, ELIXIR All Hands 2022
    - ELIXIR All Hands 2023

Three papers in preparation

- Ten Simple Rules for making tools and workflows findable and citable (target: PLoS Comp Bio)
- EOSC-Life Workflow Collaboratory (target: GigaScience)
- WorkflowHub: a FAIR registry for FAIR Computational Workflows (target: Nature Scientific Data)

Other related publications include

- Brack P, Crowther P, Soiland-Reyes S, Owen S, Lowe D, Williams AR, et al. (2022) Ten simple rules for making a software tool workflow-ready. PLoS Comput Biol 18(3): e1009823. https://doi.org/10.1371/journal.pcbi.1009823 [10]

This project has received funding from the *European Union’s Horizon 2020 research and innovation programme* under grant agreement No 824087.

Next Steps

Horizon Europe projects committed to using WorkflowHub and contributing to supporting its development and sustainability for the next 4 years include:

- EOSC4Cancer[92] - A European-wide foundation to accelerate data-driven Cancer Research;
- BY-COVID[93] - BeYond-COVID, make data from COVID-19 and other infectious diseases open and accessible to everyone;
- EuroScienceGateway[94] - Improving computational workflows in Galaxy by maturing Pulsar Network and publishing computational workflows as FAIR Digital Objects;
- AgroServ[95] - Integrated SERVices supporting a sustainable AGROecological transition;
- BioDT[96] - Biodiversity Digital Twin;
- Biodiversity Genomics Europe (BGE)[97] - Tackle biodiversity crisis using DNA data.

The service is part of the ELIXIR Tools ecosystem portfolio. ELIXIR National nodes UK and Belgium are seeking national funding to support the Hub. Within the UK the WorkflowHub is the registry for the DARE-UK TRE-FX[98] project - delivering a federated network of TREs to enable safe analytics, and work is underway to drive adoption of the Hub across the UK's Digital Research Infrastructures. We plan to replicate this across ELIXIR Nodes.

Internationally, the Australian BioCommons are committed to support WorkflowHub and handle Pacific-time requests for Team registration and Oak Ridge National Laboratory act as the North American handler of requests.

In 2021-2022 the University of Manchester conducted a User Experience study for WorkflowHub in which 22 people took part. Although satisfaction was high improvements identified included:

- occasional problems of situational awareness;
- some tasks lack simplicity;
- improvements needed in search functionality;
- explanation needed for certain concepts;
- inconsistencies in the user interface.

Work will continue on the UI, adding new features and general improvements. In particular for:

- workflow ensembles - workflows that are groups of workflows;
- tracking workflow cloning;
- richer support for sub-workflows and nested workflows.

Several features for linking data with workflows, and linking workflows to each other, have been underused. Future work, through better guidelines, is needed to get these rich and powerful features used.

---

[92] https://esciencelab.org.uk/projects/eosc4cancer/
[93] https://esciencelab.org.uk/projects/by-covid/
[94] https://esciencelab.org.uk/projects/eurosciencegateway/
[95] https://esciencelab.org.uk/projects/agroserv/
[96] https://esciencelab.org.uk/projects/biodt/
[97] https://esciencelab.org.uk/projects/biodiversitygenomics/
[98] https://esciencelab.org.uk/projects/tre-fx/

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

Widening support for Workflow Management Systems

EOSC-Life particularly focused on custom support for Galaxy, both the WorkflowHub and sister services providing dedicated support for Galaxy and its curated repository (Intergalactic Workflow Constellation), and for Galaxy adopting WorkflowHub and its metadata framework. CWL has also had support.

Next steps are to widen custom support for popular managers, notably Nextflow, Snakemake and Jupyter Notebooks. Such support entails:

- Custom semi-automated metadata extraction for workflow registration;
- Support for established community repositories (nf-core, Snakemake Catalog, BinderHub) including repository layout guidelines, auto-harvesting and release support, leveraging the LifeMonitor GitHub App;
- running on TRS complaint instances (e.g. Nextflow Tower);
- RO-Crate and Abstract CWL support by the workflow managers with tooling support.

This work requires strengthening links with these workflow manager developers and their communities.

Improving support for Workflow Lifecycle

Workflows in WorkflowHub are stored internally as Git repositories, making it possible to freely add, modify and delete files for uploaded “local” workflows, and freeze snapshots of workflow versions. Workflows uploaded by Git import or RO-Crate need to be explicitly versioned and reuploaded. The intention is that WorkflowHub and its sister services should support the living lifecycle of a workflow not just end of life publishing - from “work in progress” to archiving (64 workflows are currently labelled as work in progress). Improvements to git repository support for living workflows is planned, which requires co-developed with workflow communities and their native repositories.

Workflow are inherently compositional and may be made up of sub workflows, or work with workflow ensembles. Workflows may be repurposed, remixed and adopted. Tracking the evolution of workflows (with credit and attribution), and better support for sub-workflows is planned.

Community of Practice and Workflow Organisation

The Spaces/Teams model and registration process is recognised as a barrier to adoption and is being reviewed to be streamlined. In addition, the way collections are organised across Spaces/Teams is challenging. Collections are beginning to pick up momentum, especially as more Teams need to gather and share workflows between them. Although collections can be heterogeneous, mixing Workflows, Datafiles, Documents and any other registrable item in WorkflowHub, to date they have been only used to group Workflows. Better guidelines should improve best practice in the use and power of collections, and how to set up Spaces and Teams.

Policies, guidelines and training

WorkflowHub was set up much faster than intended to service the COVID-19 crisis, and we had few available resources to provide support for organising workflows. As practice emerged it is clear that we need much stronger guidelines on how to best use WorkflowHub and we need to include best practices for FAIR Workflows developed by the community through the Workflow Community Initiative and by workflow communities such as nf-core and snakemake. Guidelines also extend to a few recommended ways of organising a Git repository so that release/version updating can be automatically managed as we do for the Galaxy IWC.

Next steps also include the addition of WorkflowHub into existing workflow training materials, such as:

- Galaxy training network Smörgåsbords[99] and Galaxy training;
- Software Carpentry lessons for Nextflow[100] and Snakemake[101] as developed by the Ed-DASH project[102] - discussions have begun with the Ed-DASH team.

This work will be carried out in the EuroScienceGateway, BioDT and BGE projects, and with the Australian BioCommons.

Workflow publishing

WorkflowHub has features for minting DOIs and harvesting citation metadata from citation.cff files, as well as author and contributor credit. The Hub has begun to be used as a registry for supplementary material for workflows associated with publications. The next steps are to formalise this by making WorkflowHub a recommended registry for Journals with best practice recommendations for how to organise workflows for journals. Discussions have opened with GigaScience and the F1000 Bioinformatics Gateway.

Several journals have moved towards Workflow publications - that is the publications are primarily about the workflow. Here recommended practices are required for peer review of workflows, in so far as they are a specialised form of software.

Curation

WorkflowHub does not have the people resources for curating workflows, reorganising collections, checking links are up to date, tagging etc. LifeMonitor provides facilities for checking the viability of the workflow. More can be done to semi-automate metadata annotation - for example using LLMs. Clearer guidelines will improve the quality of workflow organisation on the Hub. Big projects such as BioDT, BY-COVID, BGE should have strong curation practice and policies from the start.

---

[99] https://training.galaxyproject.org/

[100] https://carpentries-incubator.github.io/workflows-nextflow/

[101] https://carpentries-incubator.github.io/snakemake-novice-bioinformatics/

[102] https://edcarp.github.io/Ed-DaSH/

Integration with other services

The metadata framework, with RO-Crates, provides the means to exchange workflows and their components. Alongside closer integration with more Workflow Managers, other services were planned to be integrated and are still planned:

- **OpenEBench[103]** for bench marking tools, has benchmarking workflows[104] but has yet to be fully integrated with WorkflowHub using RO-Crates.
- **OpenAIRE[105]** is a scholarly communication infrastructure (CRIS) that provides a catalogue of services[106] including an OpenAIRE Research Graph[107], open citations and enriching content. Through DataCite we have a means to incorporate those few workflows with DOIs, (DataCite MDSchema is used by OpenAire and others for OAI-harvesting MD from other MD Aggregators)[108], but there is much more to do and greater benefits to be had.
- **Research Organisation Registry[109]** for organisation identifiers.

Community engagement

WorkflowHub is a member of the SciCodes[110] Consortium of scientific software registries and repositories and seeks to comply to their nine best practices[111] [11].

The Manchester WorkflowHub team are members of Horizon Europe FAIR-IMPACT[112] project which includes work on the implementation of FAIR Principles for Research Software (FAIR4RS) and the specification of CodeMeta metadata for software. Workflows are a hybrid of FAIR principles for data and software [3].

FAIR Computational Workflows is a working group of the Workflow Community Initiative[113] co-chaired by Carole Goble, the PI of WorkflowHub, where further work will be developed on the FAIR principles for workflows and the features needed for a FAIR Workflow Registry.

---

[103] https://openebench.bsc.es/
[104] https://workflowhub.eu/projects/71#workflows
[105] https://www.openaire.eu/
[106] https://catalogue.openaire.eu/home
[107] https://graph.openaire.eu/
[108] https://guidelines.openaire.eu/en/latest/data/index.html
[109] https://ror.org/
[110] http://scicodes.net/
[111] https://scicodes.net/best-practices-for-software-registries-and-repositories/
[112] https://fair-impact.eu/
[113] https://workflows.community/about

Further details

Communities of Practice

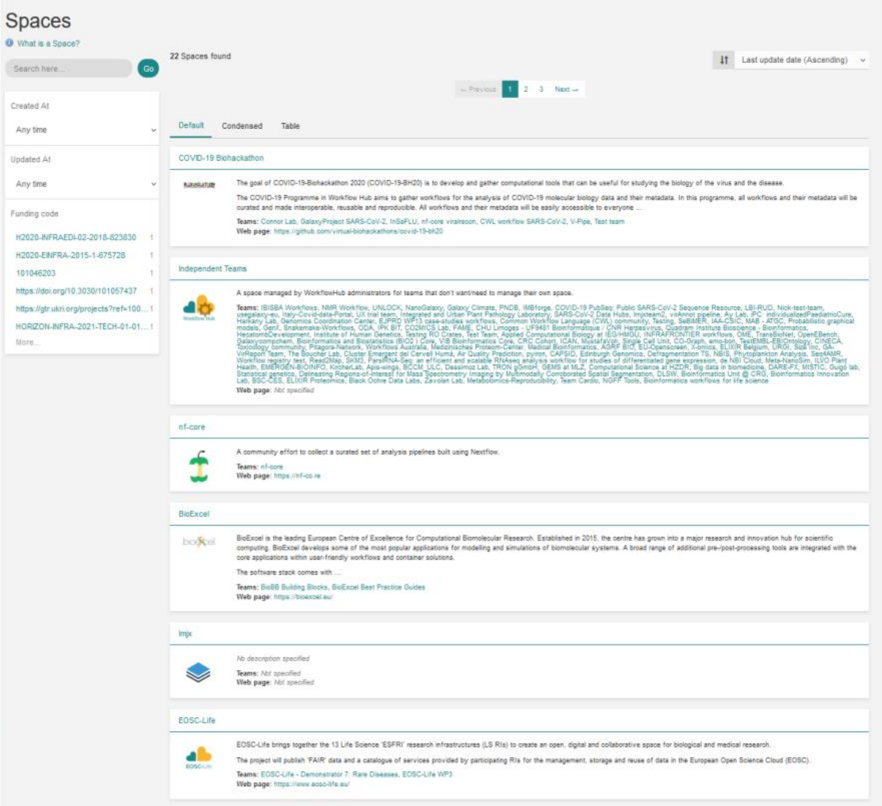
Spaces and Teams

Workflows are developed by individuals but also in Communities of Practice chiefly organised as Teams and Spaces. Workflows are available for browsing, downloading, and launching without login in. To add content contributors must register with WorkflowHub.

A Workflow is associated with at least one Team (Figure 7). **A Team** represents a group of people collaborating on a particular activity and provides the ability to describe and promote the activity, and also give author credit as well as being an anchor point for workflows and other WorkflowHub registered content. A Team supports default sharing permissions and licence, and the ability to easily set sharing permissions across all members. A member of the Team is related to an Organisation for the context of that Team, allowing users to be related to different Organisations for different Teams.

To register a Workflow, a contributor must be a member of the Team that will own the Workflow. To request membership the Team must already exist with an administrator who can grant membership. If a new Team is needed a request to make one needs to be made to a Spaces administrator, because Teams are grouped into Spaces. **A Space** (Figure 6) is a self-administered umbrella organisation for consortia or other long-running activities that may cover multiple Teams. The default Space is “Independent Teams” and the administrator is WorkflowHub. All other Spaces are set up and administered by independent operators who have been given authority by WorkflowHub to create the Teams within their Space.

Figure 8 presents the Space/Team model. This setup is to avoid spam and malicious content, but it is recognised that it is confusing and a barrier to onboarding and is under review.



COVID-19 Biohackathon

The goal of COVID-19-Biohackathon 2020 (COVID-19-BH20) is to develop and gather computational tools that can be useful for studying the biology of the virus and the disease.

The COVID-19 Programme in Workflow Hub aims to gather workflows for the analysis of COVID-19 molecular biology data and their metadata. In this programme, all workflows and their metadata will be curated and made interoperable, reusable and reproducible. All workflows and their metadata will be easily accessible to everyone ...

Teams: Connor Lab, GalaxyProject SARS-CoV-2, InSaFLU, nf-core viralrecon, CWL workflow SARS-CoV-2, V-Pipe, Test team
Web page: https://github.com/virtual-biohackathons/covid-19-bh20

Independent Teams

A space managed by WorkflowHub administrators for teams that don't want/need to manage their own space.

Teams: IBISBA Workflows, NMR Workflow, UNLOCK, NanoGalaxy, Galaxy Climate, PNDB, IMBforge, COVID-19 PubSeq: Public SARS-CoV-2 Sequence Resource, LBI-RUD, Nick-test-team, usegalaxy-eu, Italy-Covid-data-Portal, UX trial team, Integrated and Urban Plant Pathology Laboratory, SARS-CoV-2 Data Hubs, Imjxteam2, virAnnot pipeline, Ay Lab, iPC: individualizedPaediatricCure, Harkany Lab, Genomics Coordination Center, EJPRD WP13 case-studies workflows, Common Workflow Language (CWL) community, Testing, SeBiMER, IAA-CSIC, MAB - ATGC, Probabilistic graphical models, GenX, Snakemake-Workflows, ODA, IPK BIT, CO2MICS Lab, FAME, CHU Limoges - UF9481 Bioinformatique / CNR Herpesvirus, Quadram Institute Bioscience - Bioinformatics, HecatombDevelopment, Institute of Human Genetics, Testing RO Crates, Test Team, Applied Computational Biology at IEG/HMGU, INFRAFRONTIER workflows, OME, TransBioNet, OpenEBench, Galaxycompchem, Bioinformatics and Biostatistics (BIO2) Core, VIB Bioinformatics Core, CRC Cohort, ICAN, MustafaVoh, Single Cell Unit, CO-Graph, emo-bon, TestEMBL-EBIOntology, CINECA, Toxicology community, Pitagora-Network, Workflows Australia, Medizinisches Proteom-Center, Medical Bioinformatics, AGRF BIO, EU-Openscreen, X-omics, ELIXIR Belgium, URGI, Size Inc, GA-VirReport Team, The Boucher Lab, Cluster Emergent del Cervell Humà, Air Quality Prediction, pyiron, CAPSID, Edinburgh Genomics, Defragmentation TS, NBIS, Phytoplankton Analysis, Seq4AMR, Workflow registry test, Read2Map, SKM3, ParslRNA-Seq: an efficient and scalable RNAseq analysis workflow for studies of differentiated gene expression, de.NBI Cloud, Meta-NanoSim, ILVO Plant Health, EMERGEN-BIOINFO, KircherLab, Apis-wings, BCCM_ULC, Dessimoz Lab, TRON gGmbH, GEMS at MLZ, Computational Science at HZDR, Big data in biomedicine, DARE-FX, MISTIC, Guigó lab, Statistical genetics, Delineating Regions-of-interest for Mass Spectrometry Imaging by Multimodally Corroborated Spatial Segmentation, DLSW, Bioinformatics Unit @ CRG, Bioinformatics Innovation Lab, BSC-CES, ELIXIR Proteomics, Black Ochre Data Labs, Zavolan Lab, Metabolomics-Reproducibility, Team Cardio, NGFF Tools, Bioinformatics workflows for life science
Web page: *Not specified*

nf-core

A community effort to collect a curated set of analysis pipelines built using Nextflow.

Teams: nf-core
Web page: https://nf-co.re

BioExcel

BioExcel is the leading European Centre of Excellence for Computational Biomolecular Research. Established in 2015, the centre has grown into a major research and innovation hub for scientific computing. BioExcel develops some of the most popular applications for modelling and simulations of biomolecular systems. A broad range of additional pre-/post-processing tools are integrated with the core applications within user-friendly workflows and container solutions.

The software stack comes with ...

Teams: BioBB Building Blocks, BioExcel Best Practice Guides
Web page: https://bioexcel.eu/

lmjx

No description specified

Teams: *Not specified*
Web page: *Not specified*

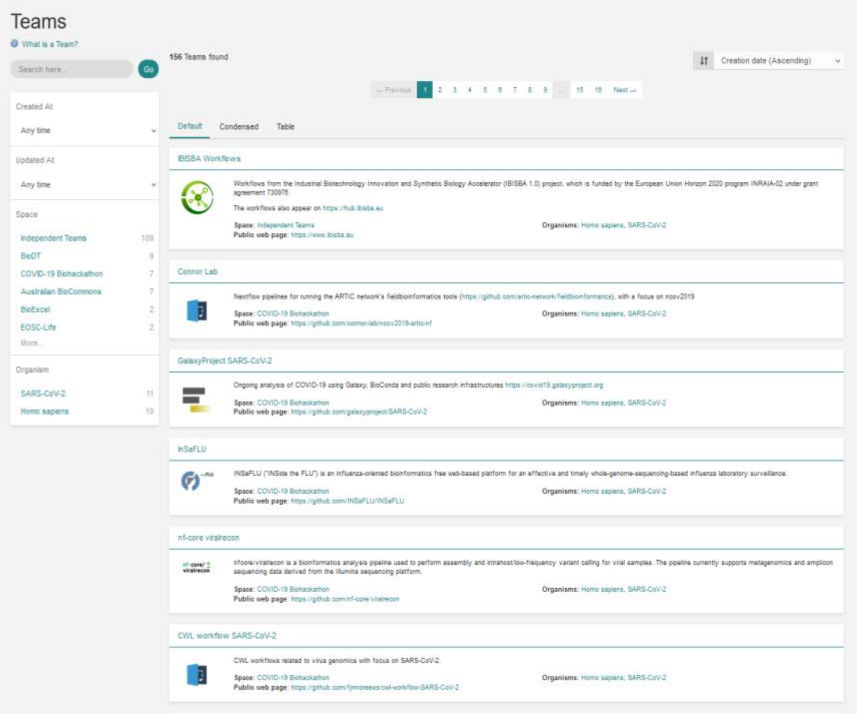
EOSC-Life

EOSC-Life brings together the 13 Life Science 'ESFRI' research infrastructures (LS RIs) to create an open, digital and collaborative space for biological and medical research.

The project will publish 'FAIR' data and a catalogue of services provided by participating RIs for the management, storage and reuse of data in the European Open Science Cloud (EOSC).

Teams: EOSC-Life - Demonstrator 7: Rare Diseases, EOSC-Life WP3
Web page: https://www.eosc-life.eu/

Figure 6: Spaces including Independent Teams and EOSC-Life (06/04/2023)



IBISBA Workflows

Workflows from the Industrial Biotechnology Innovation and Synthetic Biology Accelerator (IBISBA 1.0) project, which is funded by the European Union Horizon 2020 program INRAIA-02 under grant agreement 730976.

The workflows also appear on https://hub.ibisba.eu

Space: Independent Teams
Public web page: https://www.ibisba.eu
Organisms: Homo sapiens, SARS-CoV-2

Connor Lab

Nextflow pipelines for running the ARTIC network's fieldbioinformatics tools (https://github.com/artic-network/fieldbioinformatics), with a focus on ncov2019

Space: COVID-19 Biohackathon
Public web page: https://github.com/connor-lab/ncov2019-artic-nf
Organisms: Homo sapiens, SARS-CoV-2

GalaxyProject SARS-CoV-2

Ongoing analysis of COVID-19 using Galaxy, BioConda and public research infrastructures https://covid19.galaxyproject.org

Space: COVID-19 Biohackathon
Public web page: https://github.com/galaxyproject/SARS-CoV-2
Organisms: Homo sapiens, SARS-CoV-2

InSaFLU

INSaFLU ("INSide the FLU") is an influenza-oriented bioinformatics free web-based platform for an effective and timely whole-genome-sequencing-based influenza laboratory surveillance.

Space: COVID-19 Biohackathon
Public web page: https://github.com/INSaFLU/INSaFLU
Organisms: Homo sapiens, SARS-CoV-2

nf-core viralrecon

nfcore/viralrecon is a bioinformatics analysis pipeline used to perform assembly and intrahost/low-frequency variant calling for viral samples. The pipeline currently supports metagenomics and amplicon sequencing data derived from the Illumina sequencing platform.

Space: COVID-19 Biohackathon
Public web page: https://github.com/nf-core/viralrecon
Organisms: Homo sapiens, SARS-CoV-2

CWL workflow SARS-CoV-2

CWL workflows related to virus genomics with focus on SARS-CoV-2.

Space: COVID-19 Biohackathon
Public web page: https://github.com/fjrmoreews/cwl-workflow-SARS-CoV-2
Organisms: Homo sapiens, SARS-CoV-2

Figure 7: Teams (06/04/2023)

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

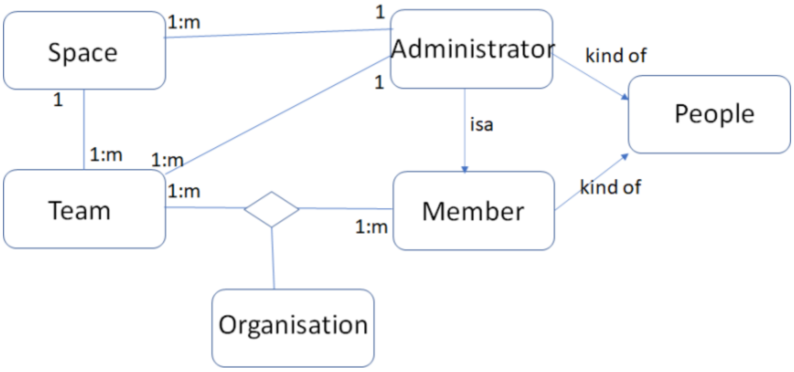
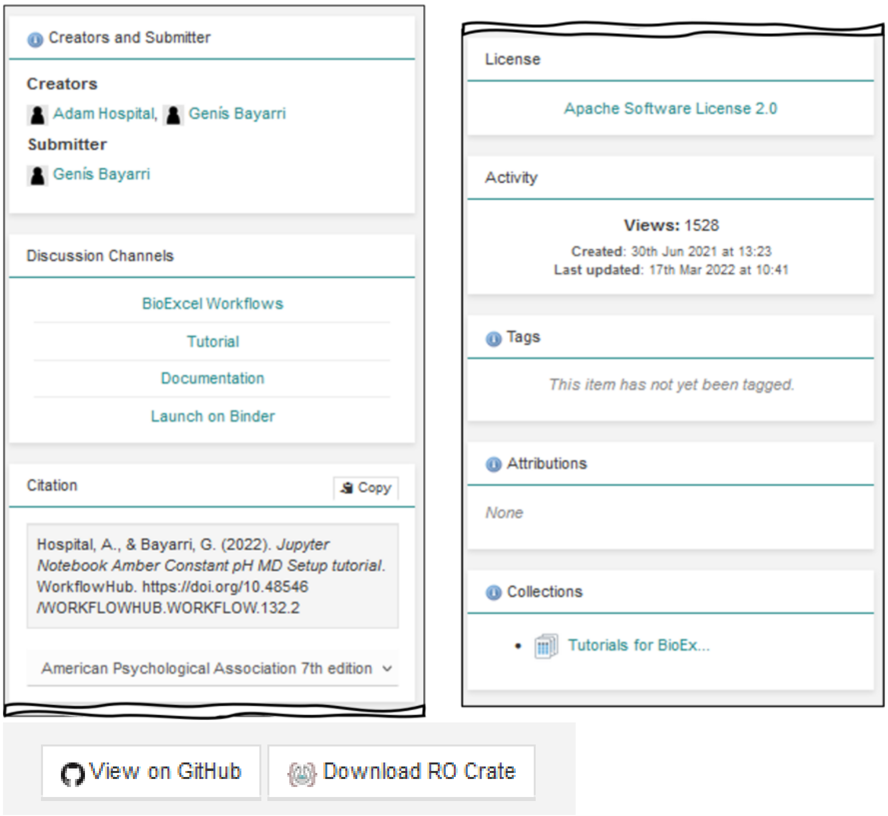


Figure 8: Relationships between Spaces, Teams, Organisations and People.

Connections to community channels

As well as Teams enabling Communities of Practice to link together contributors, other features link the developers of the workflows with their potential user base. Figure 9 shows that discussion channels are directly linked to the information bar associated with a workflow, alongside a direct link to the workflow's community repository on GitHub.



Creators and Submitter

Creators

Adam Hospital, Genís Bayarri

Submitter

Genís Bayarri

Discussion Channels

BioExcel Workflows

Tutorial

Documentation

Launch on Binder

Citation

Hospital, A., & Bayarri, G. (2022). *Jupyter Notebook Amber Constant pH MD Setup tutorial.* WorkflowHub. https://doi.org/10.48546/WORKFLOWHUB.WORKFLOW.132.2

License

Apache Software License 2.0

Activity

Views: 1528

Created: 30th Jun 2021 at 13:23
Last updated: 17th Mar 2022 at 10:41

Tags

This item has not yet been tagged.

Attributions

None

Collections

- Tutorials for BioEx...

Figure 9: Information bar for Workflow https://workflowhub.eu/workflows/132 and link to its GitHub repository. Workflow registered by the BioExcel European Centre of Excellence for Computational Biomolecular Research.

Credit and Attribution

Figure 8 demonstrates features for building credit and visibility for workflows and their developers:

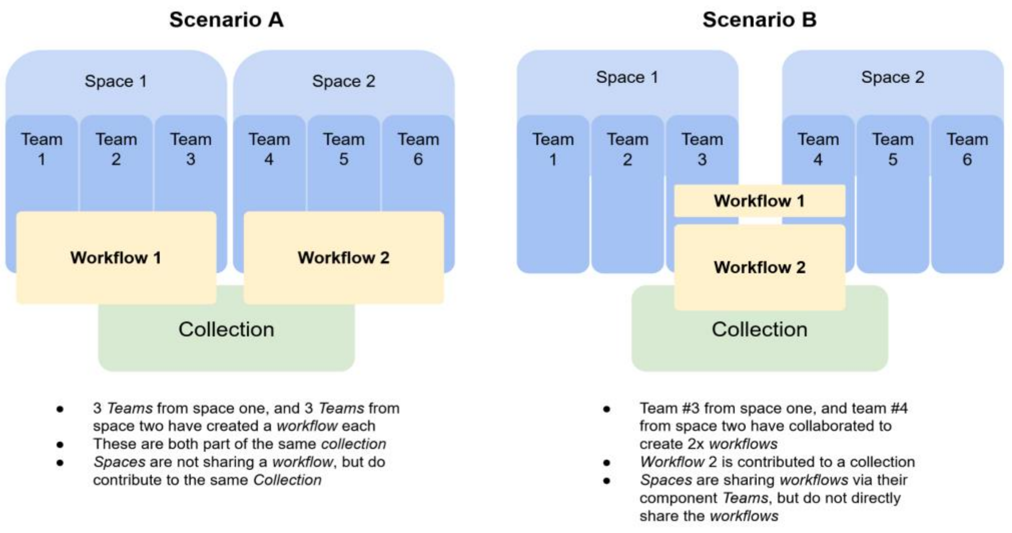
- Both the creators of the workflow and the submitters of the entry are acknowledged.
- A DOI snapshot of a workflow, using DataCite, enables the workflow to be cited in the scholarly literature and feeds into the DataCite PID Graph[114].
- Activity analytics give both potential users and its developers information on the popularity and currency of the workflow.
- Attributions are links to workflows that have influenced this one - for example it is derived from or associated with. This is underused currently and an area of further development.
- License informs the conditions under which this workflow can be reused.
- Tags are free text words and phrases used for search and organisation. Here the workflow has not been tagged - ideas for auto tagging are in the pipeline.
- Curated collections are another form of building visibility for workflows.

Sharing permissions

Teams have detailed control on workflow visibility to team members, members outside the team and to the public. Although open access is encouraged, it is recognised that workflows in development may be registered privately. Of the 324 workflows registered in WorkflowHub, 15 are private.

Organising workflows

Workflows can be organised using the Spaces & Teams (as above), through Collections and using tags. Workflows can appear in multiple Teams and multiple Collections.



- 3 *Teams* from space one, and 3 *Teams* from space two have created a *workflow* each
- These are both part of the same *collection*
- *Spaces* are not sharing a *workflow*, but do contribute to the same *Collection*

- Team #3 from space one, and team #4 from space two have collaborated to create 2x *workflows*
- *Workflow* 2 is contributed to a collection
- *Spaces* are sharing *workflows* via their component *Teams*, but do not directly share the *workflows*

Figure 10: Different ways workflows can be organised using Spaces, Teams and Collections

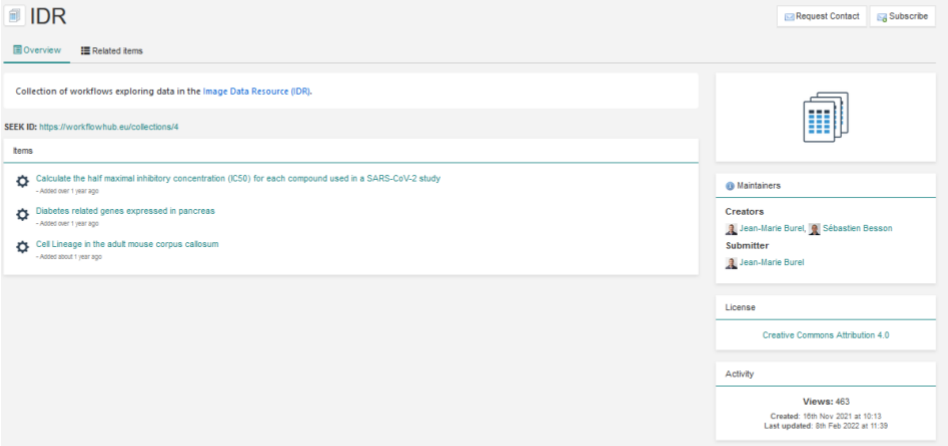
---

[114] https://blog.datacite.org/tag/pid-graph/

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

Collections

Collections are explicitly created curated groupings of any kind of content in WorkflowHub that cut across Teams, and Collections can be associated with any number of Teams. Items within a collection can be ordered. They are not nested, so you cannot have collections of collections. Collections have “maintainers” with the ability to add and remove items from the Collection. Figure 11 shows a collection of Jupyter Notebooks from the Euro-BioImaging-ERIC.



IDR

Collection of workflows exploring data in the Image Data Resource (IDR).

SEEK ID: https://workflowhub.eu/collections/4

Items

- Calculate the half maximal inhibitory concentration (IC50) for each compound used in a SARS-CoV-2 study
  - Added over 1 year ago
- Diabetes related genes expressed in pancreas
  - Added over 1 year ago
- Cell Lineage in the adult mouse corpus callosum
  - Added about 1 year ago

Maintainers

Creators
Jean-Marie Burel, Sébastien Besson
Submitter
Jean-Marie Burel

License

Creative Commons Attribution 4.0

Activity

Views: 463
Created: 16th Nov 2021 at 10:13
Last updated: 8th Feb 2022 at 11:39

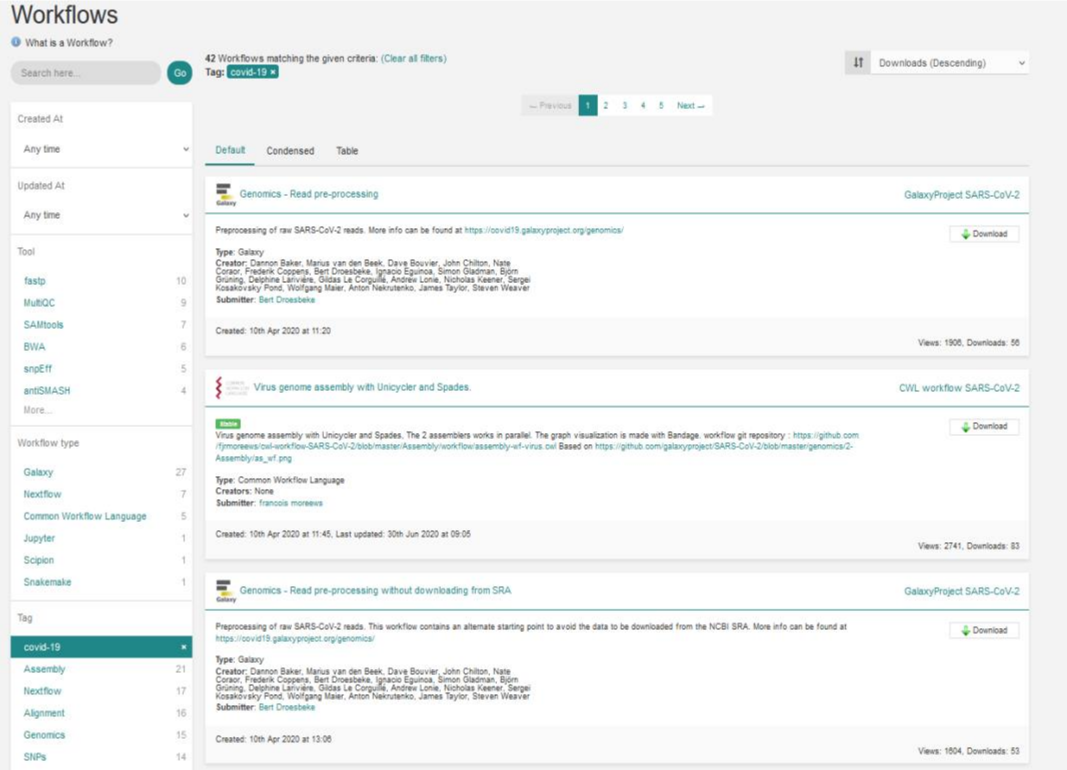
Figure 11: Collection of Jupyter Notebooks from the Euro-BioImaging-ERIC.

After a slow start, collections are beginning to pick up momentum, especially as more Teams need to gather and share workflows between them. Although collections can be heterogeneous, mixing Workflows, Datafiles, Documents and any other registrable item in WorkflowHub, to date they have been only used to group Workflows. Better guidelines should improve best practice in the use and power of collections.

Tagging

A third mechanism for organising workflows (and other content) is by tagging using keywords. Contributors are encouraged to use tags that are already available and are promoted. Faceted browsing using the tags enables dynamic sub-collections to be generated (Figure 12). This, however, is only as good as the tagging. Covid-19 is the most popular tag, unsurprisingly.

This project has received funding from the *European Union’s Horizon 2020 research and innovation programme* under grant agreement No 824087.



42 Workflows matching the given criteria: (Clear all filters)
Tag: covid-19 ×

Genomics - Read pre-processing — GalaxyProject SARS-CoV-2

Preprocessing of raw SARS-CoV-2 reads. More info can be found at https://covid19.galaxyproject.org/genomics/

Type: Galaxy
Creator: Dannon Baker, Marius van den Beek, Dave Bouvier, John Chilton, Nate Coraor, Frederik Coppens, Bert Droesbeke, Ignacio Eguinoa, Simon Gladman, Björn Grüning, Delphine Larivière, Gildas Le Corguillé, Andrew Lonie, Nicholas Keener, Sergei Kosakovsky Pond, Wolfgang Maier, Anton Nekrutenko, James Taylor, Steven Weaver
Submitter: Bert Droesbeke

Created: 10th Apr 2020 at 11:20
Views: 1906, Downloads: 56

Virus genome assembly with Unicycler and Spades. — CWL workflow SARS-CoV-2

Virus genome assembly with Unicycler and Spades, The 2 assemblers works in parallel. The graph visualization is made with Bandage. workflow git repository : https://github.com/fjmoreews/cwl-workflow-SARS-CoV-2/blob/master/Assembly/workflow/assembly-wf-virus.cwl Based on https://github.com/galaxyproject/SARS-CoV-2/blob/master/genomics/2-Assembly/as_wf.png

Type: Common Workflow Language
Creators: None
Submitter: francois moreews

Created: 10th Apr 2020 at 11:45, Last updated: 30th Jun 2020 at 09:05
Views: 2741, Downloads: 83

Genomics - Read pre-processing without downloading from SRA — GalaxyProject SARS-CoV-2

Preprocessing of raw SARS-CoV-2 reads. This workflow contains an alternate starting point to avoid the data to be downloaded from the NCBI SRA. More info can be found at https://covid19.galaxyproject.org/genomics/

Type: Galaxy
Creator: Dannon Baker, Marius van den Beek, Dave Bouvier, John Chilton, Nate Coraor, Frederik Coppens, Bert Droesbeke, Ignacio Eguinoa, Simon Gladman, Björn Grüning, Delphine Larivière, Gildas Le Corguillé, Andrew Lonie, Nicholas Keener, Sergei Kosakovsky Pond, Wolfgang Maier, Anton Nekrutenko, James Taylor, Steven Weaver
Submitter: Bert Droesbeke

Created: 10th Apr 2020 at 13:06
Views: 1604, Downloads: 53

Figure 12: a subcollection of workflows generated by selection of a tag. The results can be sorted by dates, titles and views/downloads.

LifeMonitor

The LifeMonitor[115] is a service that complements the WorkflowHub, as an environment to publish and share computational workflows, by supporting the workflows' improved reusability – i.e., the "R" in FAIR. More specifically, LifeMonitor supports ensuring the correct operation of the workflow over time through monitoring and triggering automated workflow tests; the application of workflow best practices, during and after publication, by providing an extensible framework for automated checks on metadata and adherence to best practices on the workflow's source code Git repository. In addition to being open for public usage, the service is supporting the maintenance of the Galaxy Intergalactic Workflow Collection (IWC) – the curated collection of Galaxy workflows – by monitoring the test status of the workflows and, in the future, by automatically publishing new or updated IWC workflow releases to WorkflowHub and LifeMonitor itself. The LifeMonitor service is hosted by CSC on the Rathi container cloud platform.

The following subsections provide a high-level overview of the system's main features. A more detailed and practical description can be found in the LifeMonitor documentation[116], while additional details about how the LifeMonitor fits within the WP2 ecosystem can be found in the previously submitted D2.2.

---

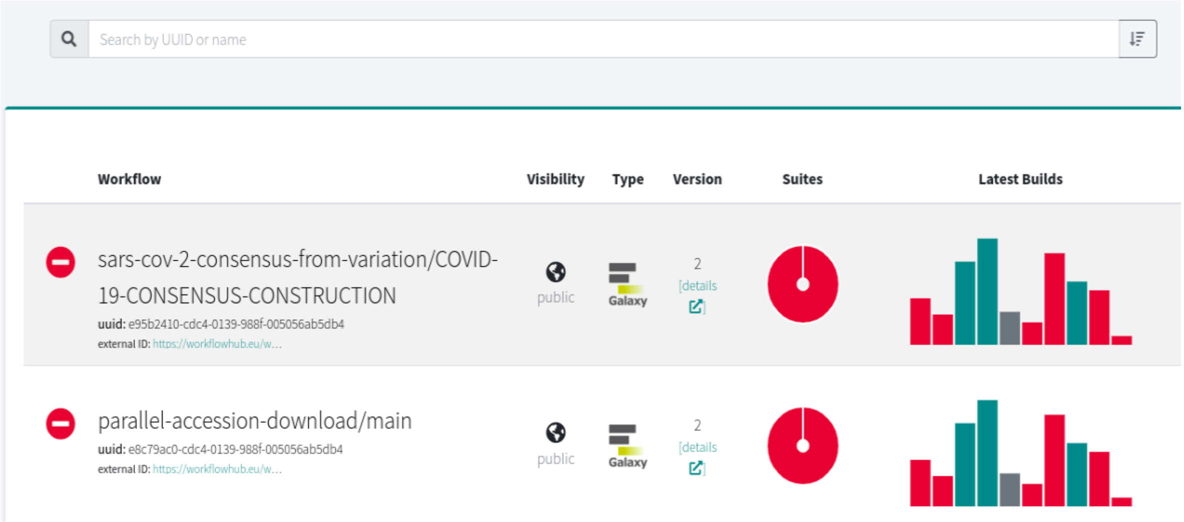
[115] https://www.lifemonitor.eu/
[116] https://www.lifemonitor.eu

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

Workflow Test Monitoring

Published workflows can be registered with LM, whereby the service acquires essential workflow metadata including a pointer to one or more instances of the workflow's tests running on a supported software automation service. LM will then periodically query the automation service to determine whether the workflow is operating correctly – i.e., that its automated tests have been executed recently and that they passed. LifeMonitor currently supports the GitHub Actions, Jenkins, and Travis CI automation services, and provides a single unified API through which the test result data from all these services can be queried and retrieved. The testing status of registered workflows along with the results of the latest workflow test executions are also reported through a convenient web-based dashboard (see Figure 13). LM has the concept of public and private visibility, so users can start using the service even before they are ready to share the workflow with the public. Moreover, users can add to their own personal workflow dashboard by subscribing to see the status of other users' publicly shared workflows registered on the system.

Further, a key feature of LifeMonitor is the availability of opt-in email notifications of workflow failures. Interested users can thus be notified when the workflows that are important to them start failing to operate properly – and, when they are fixed. This feature can help ensure that attention is brought to *software collapse* [Hinsen 2019[117]] events when they arise and encourage resolvent action by the workflow maintainers and the wider user community.



Workflow	Visibility	Type	Version	Suites	Latest Builds
sars-cov-2-consensus-from-variation/COVID-19-CONSENSUS-CONSTRUCTION **uuid:** e95b2410-cdc4-0139-988f-005056ab5db4 external ID: https://workflowhub.eu/w...	public	Galaxy	2 [details]		
parallel-accession-download/main **uuid:** e8c79ac0-cdc4-0139-988f-005056ab5db4 external ID: https://workflowhub.eu/w...	public	Galaxy	2 [details]		

Figure 13: The LifeMonitor dashboard shows testing information and other metadata for the collection of workflows to which the user "subscribes", plus the user's own workflows. The dashboard reports test success/failure and running times for the last several test execution attempts. In addition, links to the WorkflowHub entry for the workflow are provided, if available, and data for specific workflow versions of interest can be selected.

---

[117] https://doi.org/10.1109/MCSE.2019.2900945

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

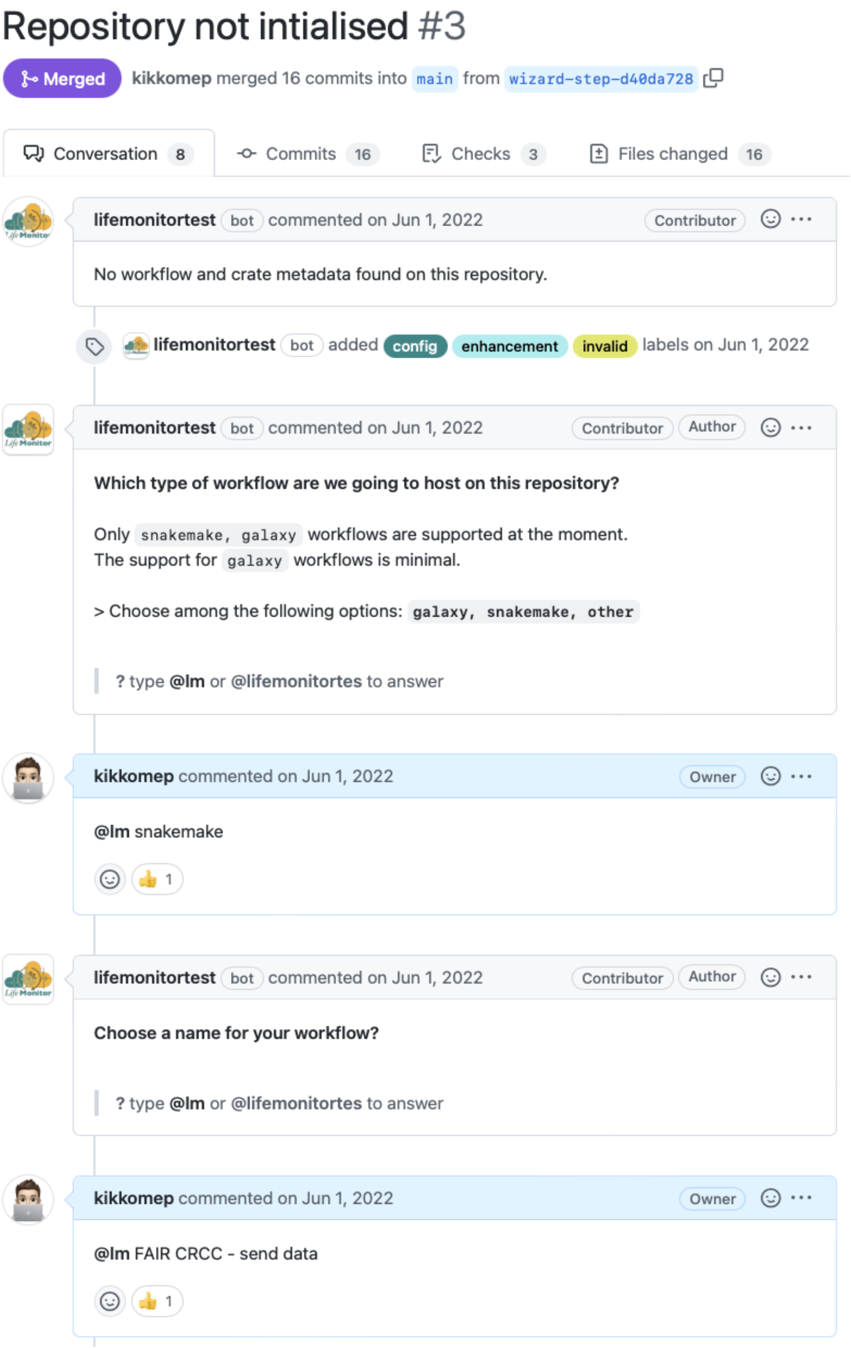
LifeMonitor GitHub App

A GitHub application for LifeMonitor has been developed to provide advanced workflow maintenance functionality facilitated by the GitHub framework for interaction with workflow source code repositories and GitHub project development features, such as *Pull Requests* and *Issues*. The LifeMonitor GitHub app implements the following main features:

- examines the repositories on which it is installed and applies a series of checks.
- suggests pull requests to make changes or additions to bring the workflow repository closer to conforming to community-accepted conventions or best practices.
- opens issues to let the workflow maintainer know about problems detected by the checks.
- automatically registers new releases/versions of the workflow with both the LifeMonitor service and the WorkflowHub workflow registry.

The app provides a bot-like interface which allows users to interact with it through the GitHub issues to almost interactively provide information about the workflow or issue commands for the bot, providing a great framework to guide workflow maintainers in the adoption of workflow sustainability best practices. For instance, when initializing a new Git repository for a workflow, the bot will open an issue on the repository and ask the maintainer for essential metadata through a wizard-like conversation (see Figure 14). The metadata is then used to propose a Workflow RO-Crate for the workflow. If deemed necessary, the metadata is also used to propose the use of community-recommended source code repository structure for the workflow by opening a GitHub Pull Request (see Figure 15). Several repository checks are already implemented to verify various aspects of the workflow metadata and new checks are still being created – e.g., to propose a GitHub Actions workflow template to perform the periodic tests.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.



Repository not intialised #3

Merged **kikkomep** merged 16 commits into `main` from `wizard-step-d40da728`

lifemonitortest bot commented on Jun 1, 2022 Contributor

No workflow and crate metadata found on this repository.

lifemonitortest bot added config enhancement invalid labels on Jun 1, 2022

lifemonitortest bot commented on Jun 1, 2022 Contributor Author

Which type of workflow are we going to host on this repository?

Only `snakemake, galaxy` workflows are supported at the moment.
The support for `galaxy` workflows is minimal.

> Choose among the following options: `galaxy, snakemake, other`

> ? type **@lm** or **@lifemonitortes** to answer

kikkomep commented on Jun 1, 2022 Owner

@lm snakemake

👍 1

lifemonitortest bot commented on Jun 1, 2022 Contributor Author

Choose a name for your workflow?

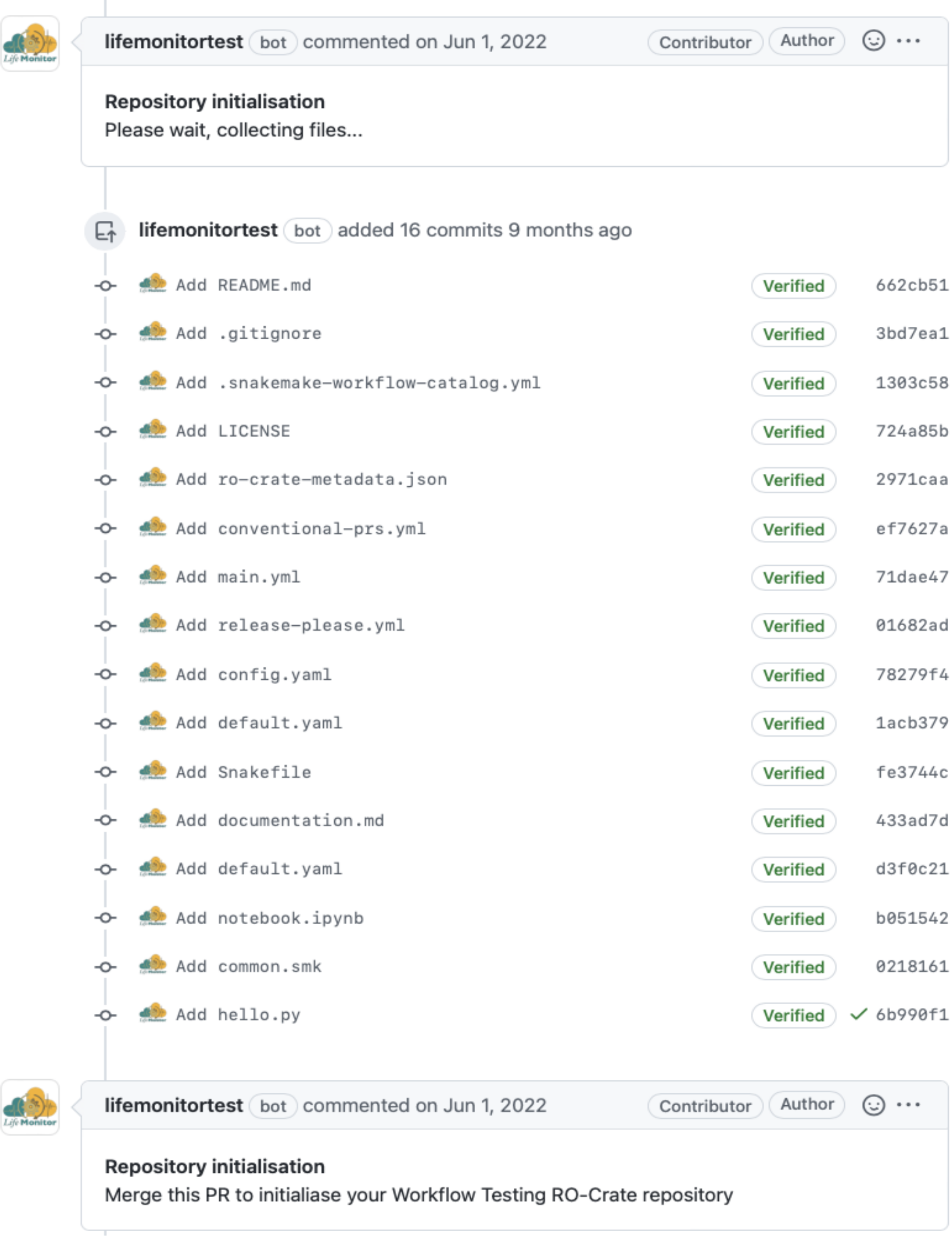
> ? type **@lm** or **@lifemonitortes** to answer

kikkomep commented on Jun 1, 2022 Owner

@lm FAIR CRCC - send data

👍 1

Figure 14: The LifeMonitor GitHub App implements a bot for almost interactive scripted interaction with the user through GitHub conversations. The App can interact with the user to request workflow metadata to be used to create the Workflow RO-Crate FAIR data object, register the workflow with the WorkflowHub, and initialise a Git repository for the workflow following best practices.



lifemonitortest bot commented on Jun 1, 2022 · Contributor · Author

Repository initialisation
Please wait, collecting files...

lifemonitortest bot added 16 commits 9 months ago

```
Add README.md                        Verified   662cb51
Add .gitignore                       Verified   3bd7ea1
Add .snakemake-workflow-catalog.yml  Verified   1303c58
Add LICENSE                          Verified   724a85b
Add ro-crate-metadata.json           Verified   2971caa
Add conventional-prs.yml             Verified   ef7627a
Add main.yml                         Verified   71dae47
Add release-please.yml               Verified   01682ad
Add config.yaml                      Verified   78279f4
Add default.yaml                     Verified   1acb379
Add Snakefile                        Verified   fe3744c
Add documentation.md                 Verified   433ad7d
Add default.yaml                     Verified   d3f0c21
Add notebook.ipynb                   Verified   b051542
Add common.smk                       Verified   0218161
Add hello.py                         Verified ✓ 6b990f1
```

lifemonitortest bot commented on Jun 1, 2022 · Contributor · Author

Repository initialisation
Merge this PR to initialiase your Workflow Testing RO-Crate repository

Figure 15: The LifeMonitor GitHub App can execute checks on workflow source code repositories to validate metadata and adherence to best practices. Checks can provide feedback to the user as GitHub Issues or Pull Requests providing a suggested fix/improvement. This image shows a Pull Request opened the App to propose a repository template that adheres to best practices. The checks framework is extensible, making it straightforward to add checks for new issues or for new workflow types.

In addition to applying metadata and convention validations to the workflow repository, the LifeMonitor GitHub App can also automate the registration of the workflow with the

WorkflowHub registry and with the LifeMonitor service itself. In fact, the App can use the workflow metadata provided by the user to register a proper workflow entry through the respective WorkflowHub and LifeMonitor service APIs, and also to register new workflow versions following a release. This functionality helps maintain the registry up to date, reducing the chances that its information goes stale. The behaviour of the App in this case can be easily configured, for instance defining how to recognize and name a new workflow release, or completely disabling this automatic functionality should the maintainer prefer to manage the registry entry manually.

Because the GitHub App is implemented on the GitHub application framework, its full functionality is only available to workflows stored in a GitHub source code repository. However, work is in progress to provide a command-line interface to manually execute workflow checks on local source code repositories, thus providing a way to leverage the functionality for those users that prefer not to host their repositories on GitHub.

Integration with WorkflowHub

The LifeMonitor service and its GitHub App interact bidirectionally with the WorkflowHub to enrich the overall workflow publishing ecosystem. The LifeMonitor GitHub App facilitates the registration of new workflows with both WorkflowHub and LifeMonitor services as well as maintaining workflow records up to date by registering new releases. The process allows the two services to exchange identifiers for workflows, creating a single workflow namespace that enables additional integrated functionality. In particular, the WorkflowHub fetches data from the LifeMonitor to enrich the workflow view with the test results, communicating to the viewer the current state of the workflow. For this functionality, the WorkflowHub leverages the LifeMonitor service’s REST API, which provides a single access point for the diverse set of automation services on which the workflow tests may run. On the other hand, the LifeMonitor fetches workflow metadata from the WorkflowHub to enhance user experience. The LM can also fetch workflow metadata from the WorkflowHub to facilitate the registration of new workflows, if these are registered manually instead of with the GitHub App.

The exchange of workflow metadata between the two services is based on the use of RO-Crate objects, in particular using the Workflow RO-Crate and Workflow Testing RO-Crate profiles.

Life Science AAI integration

The LifeMonitor service supports the state-of-the-art authentication and authorization protocols OAuth2 and OpenID Connect, and recently added support for user authentication via the Life Science AAI service (see Figure 16). In addition, LM also supports user identities provided by GitHub and WorkflowHub. The support for an advanced AAI protocol is a crucial enabler for LifeMonitor’s features that require interaction with external services – e.g., WorkflowHub, GitHub, etc. In fact, through OAuth2 LifeMonitor can be authorised by the user to perform specific actions on his/her behalf, such as registering or updating workflow entries on WorkflowHub, or triggering periodic test runs on GitHub Actions or Jenkins. Of note, the LifeMonitor does not require that the same user identity be used for all supported external services (e.g., that the workflow maintainer have an LS AAI supported account on GitHub – an impossibility, since GitHub does not support LS AAI). Instead, various external identities can be connected to the internal LifeMonitor user identity, allowing LifeMonitor to switch to the appropriate identity as it interacts with the various external services.

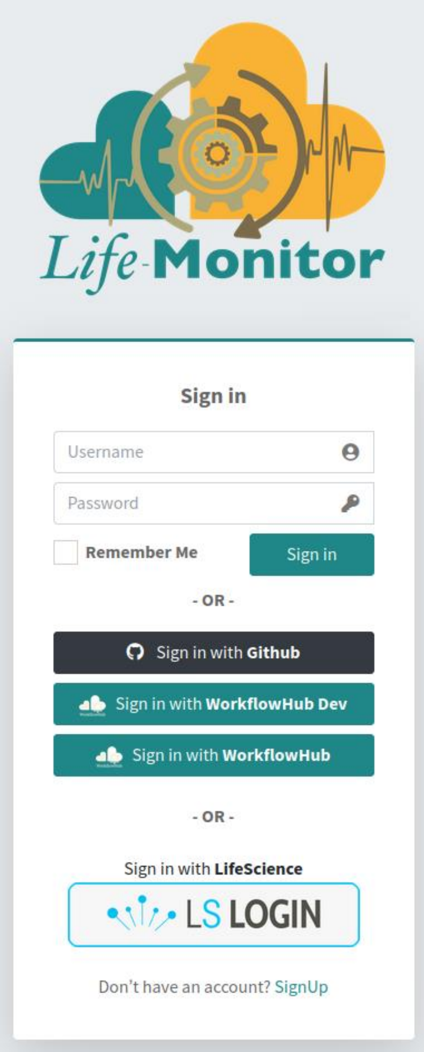


Figure 16: The LifeMonitor service supports authentication with user accounts provided by the LifeScience AAI as well as WorkflowHub and Github. The support for these different identity providers allows users to connect their accounts with the respective system so that the LifeMonitor can automatically execute operations on each service on the user's behalf.

References

1. Goble, C, Bacall, F, Soiland-Reyes, S, et al., 2022, 'WorkflowHub – a FAIR registry for workflows'. F1000Research, 11. DOI: https://doi.org/10.7490/f1000research.1118984.1
2. Goble, C, Soiland-Reyes, S, Bacall, F, et al., 2021, 'Implementing FAIR Digital Objects in the EOSC-Life Workflow Collaboratory'. Zenodo, DOI: https://doi.org/10.5281/zenodo.4605654
3. 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):108–121 https://doi.org/10.1162/dint_a_00033
4. Michael R. Crusoe, Sanne Abeln, Alexandru Iosup, Peter Amstutz, John Chilton, Nebojša Tijanić, Hervé Ménager, Stian Soiland-Reyes, Bogdan Gavrilović, Carole Goble, The CWL Community Methods Included: Standardizing Computational Reuse and Portability with the Common Workflow Language Communications of the ACM, June 2022, Vol. 65 No. 6, Pages 54-63 10.1145/3486897
5. Eguinoa, I, Suchánek, M, Knaisl, V, et al., 2022, 'BioHackEU22 Report: Enhancing Research Data Management in Galaxy and Data Stewardship Wizard by utilising RO-Crates.' BiorXiv (in prep)
6. Soiland-Reyes, S, Sefton, P, Crosas, M, et al., 2022a, 'Packaging research artefacts with RO-Crate'. Data Science 5 (2). DOI: https://doi.org/10.3233/ds-210053
7. Soiland-Reyes, S, Sefton, P, Jael Castro, L, et al., 2022, 'Creating lightweight FAIR Digital Objects with RO-Crate', 1st International Conference on FAIR Digital Objects (FDO 2022) Research Ideas and Outcomes 8:e93937. DOI: https://doi.org/10.3897/rio.8.e93937
8. Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
9. Barker, M., Chue Hong, N.P., Katz, D.S. et al. Introducing the FAIR Principles for research software. Sci Data 9, 622 (2022). https://doi.org/10.1038/s41597-022-01710-x
10. Paul Brack, Peter Crowther, Stian Soiland-Reyes, Stuart Owen, Douglas Lowe, Alan R Williams, Quentin Groom, Mathias Dillen, Frederik Coppens, Björn Grüning, Ignacio Eguinoa, Philip Ewels, Carole Goble (2022): **Ten Simple Rules for making a software tool workflow-ready**. *PLOS Computational Biology* **18**(3):e1009823 https://doi.org/10.1371/journal.pcbi.1009823
11. Garijo D, Ménager H, Hwang L, Trisovic A, Hucka M, Morrell T, Allen A, Task Force on Best Practices for Software Registries, SciCodes Consortium. 2022. Nine best practices for research software registries and repositories. PeerJ Computer Science 8:e1023 https://doi.org/10.7717/peerj-cs.1023
12. Soiland-Reyes, Stian, Sefton, Peter, Crosas, Mercè, Castro, Leyla Jael, Coppens, Frederik, Fernández, José M, Garijo, Daniel, Grüning, Björn, La Rosa, Marco, Leo, Simone, Ó Carragáin, Eoghan, Portier, Marc, Trisovic, Ana, RO-Crate Community, Groth, Paul, Goble, Carole 'Packaging Research Artefacts with RO-Crate'. 2022 Data Science 5(2): 97 – 138. DOI: 10.3233/DS-210053
13. Ison J, Kalas M, Jonassen I, Bolser D, Uludag M, McWilliam H, Malone J, Lopez R, Pettifer S, Rice P. EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. Bioinformatics. 2013 May 15;29(10):1325-32. doi: 10.1093/bioinformatics/btt113. Epub 2013 Mar 11. PMID: 23479348; PMCID: PMC3654706.

14. Alasdair Gray, Leyla J. Castro, Nick Juty, and Carole Goble Schema.org for Scientific Data in (eds Alok Choudhary, Geoffrey Fox, Tony Hey) Artificial Intelligence for Science, pp. 495-514 (2023) https://doi.org/10.1142/9789811265679_0027

Delivery and Schedule

Minor delay to submission of the deliverable due to workload issues caused by the latest round of Infrastructure calls.

WorkflowHub delivered 12 months ahead of schedule.

Adjustments made

None