



# Towards FAIR Workflows The WorkflowHub and The EOSC-Life Workflow Collaboratory

**Carole Goble,**  
Stian Soiland-Reyes, Stuart Owen  
The University of Manchester / ELIXIR-UK

**Frederik Coppens,**  
VIB / ELIXIR-BE

The rest of the WorkflowHub Club

GO-FAIR FAIR Workflows Webinar, 16<sup>th</sup> June 2021

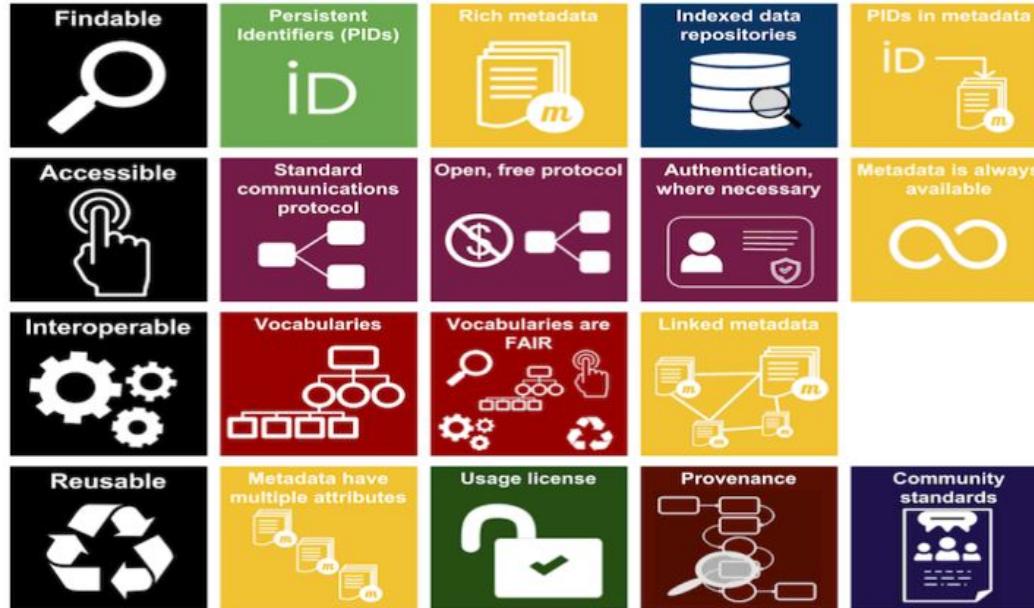


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

# FAIR in a nutshell



tl;dr:



**Persistent machine-readable and actionable metadata**

**Persistent identifiers**

**Clear licensing**

**Protocols for machine accessibility**

**Register / Index**

<https://doi.org/10.1038/sdata.2016.18>



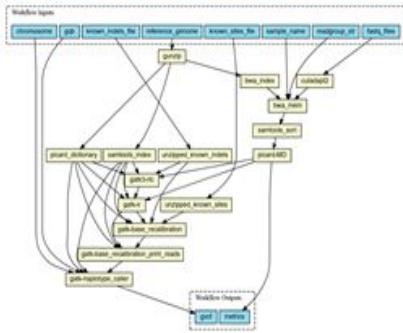
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

# Computational Workflows in a nutshell



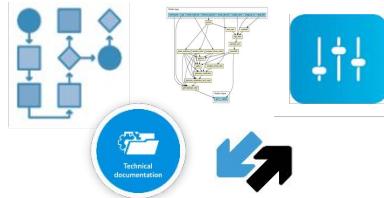
# A special kind of software

## Separation of the workflow specification from its execution



- Multi-step
- Leverage third party codes
- Scalable processing of data
- Transparent research
- Quality control

## Specification description



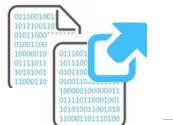
*Precise description of a procedure:  
multi-step process coordinated by  
input/output data relationships (data  
types).*

## Software Execution



*Execution of computational processes (run a code, invoke a service, run a container...). Data is consumed and produced by each step.*

# Associated Objects

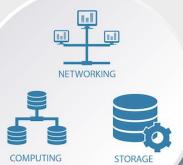


## Test engines





# European Open Science Cloud =



Enable researchers to access data, storage and compute ("cloud") via an Europe wide federation of IT services ("e-Infrastructure")



Drive the transition to Open Science (Open Data, Open Standards, Open Literature) - bring research benefits to European societies at large



Populate EOSC with the scientific data resources and computational tools from research infrastructures – drive usage by to Europe's 1.7 M researchers

**E-Infrastructure consolidation**

+

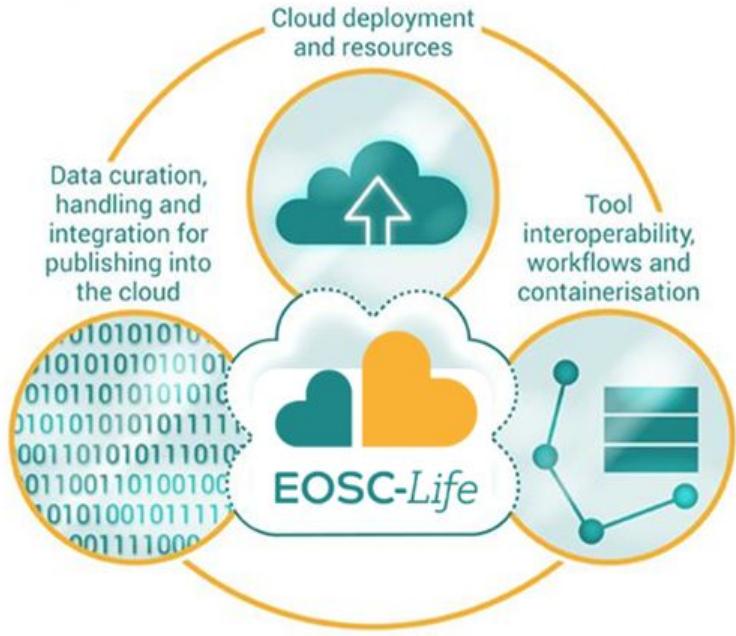
**Open Science**

+

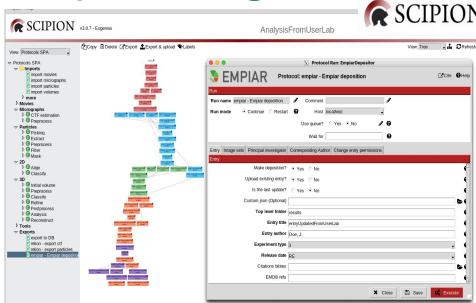
**Scientific Communities' content and users**



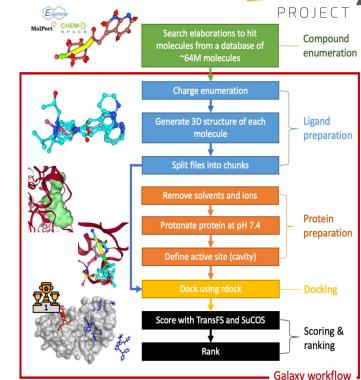
EOSC-Life pan-national thematic commons for bioscience data and methods



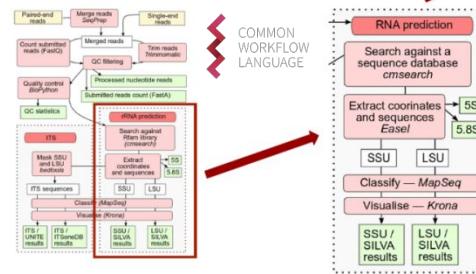
## CryoEM Image Analysis



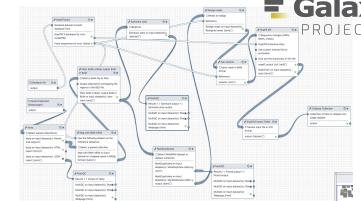
# Drug Discovery



## Metagenomic Pipelines

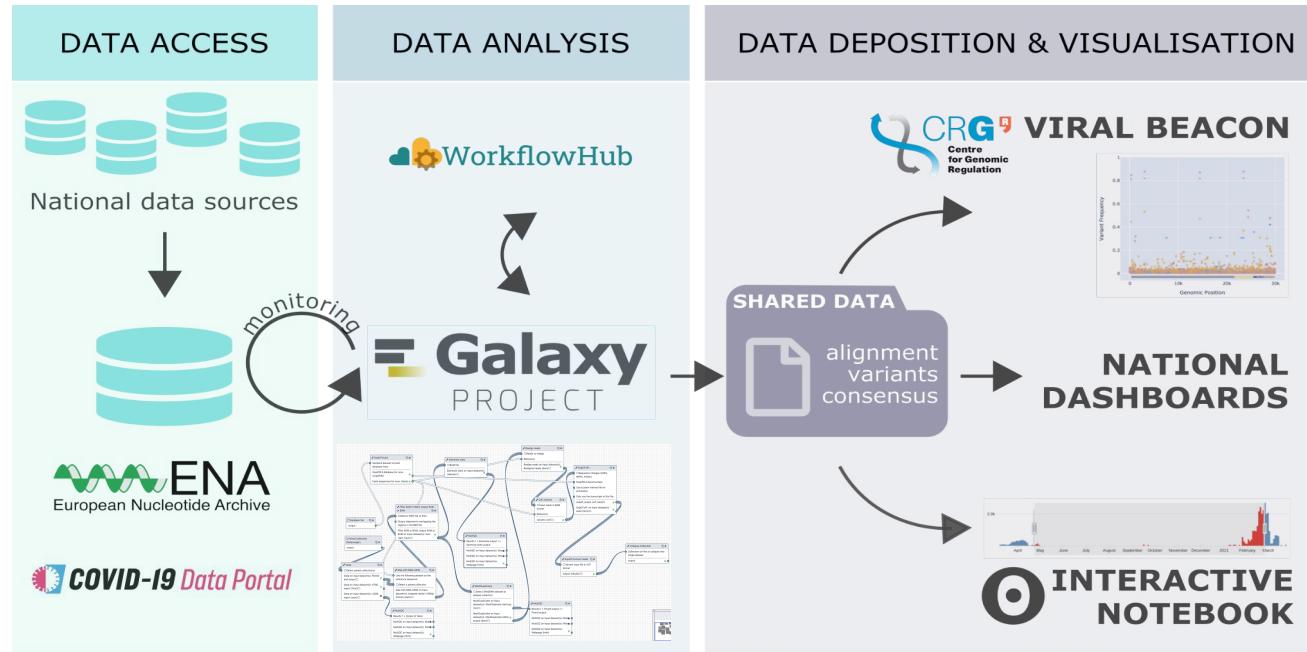


# SARS-CoV-2 Monitoring



## Using and sharing data, tools and workflows in the cloud

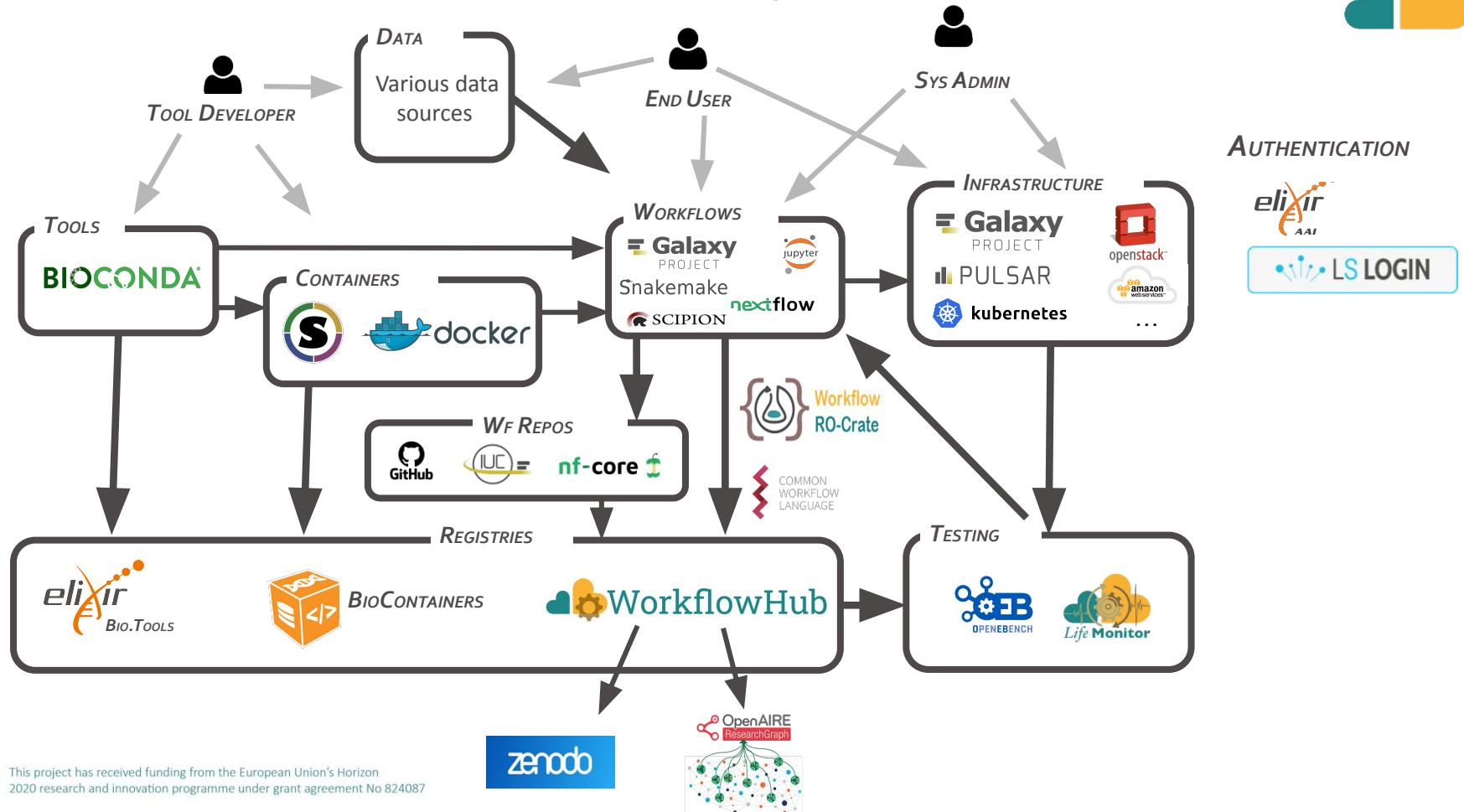
# SARS-CoV-2 pre-processing, monitoring, analysis



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

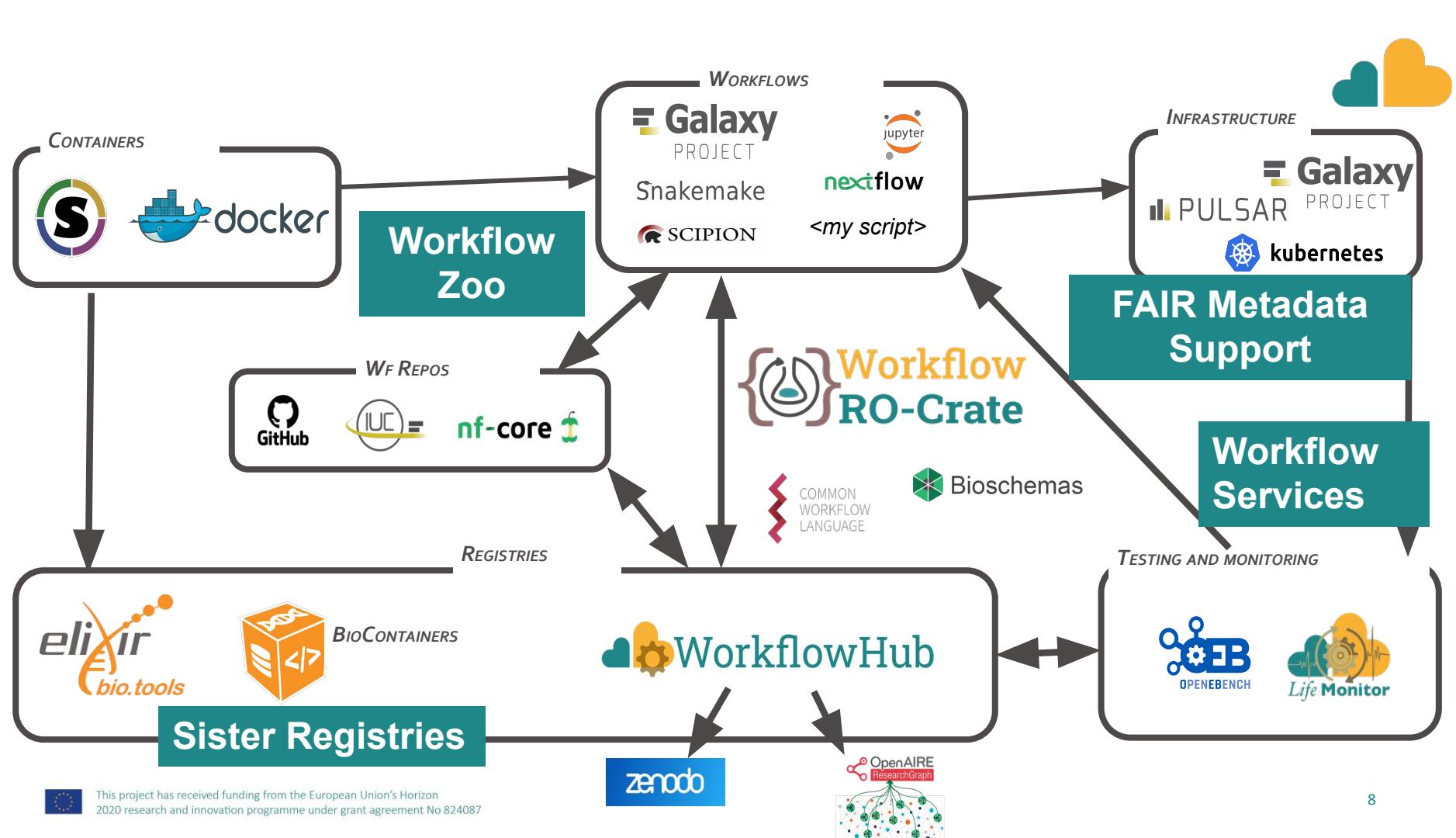
<https://elixir-europe.org/news/covid-19-variants-galaxy>

# EOSC-Life Workflow Collaboratory



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

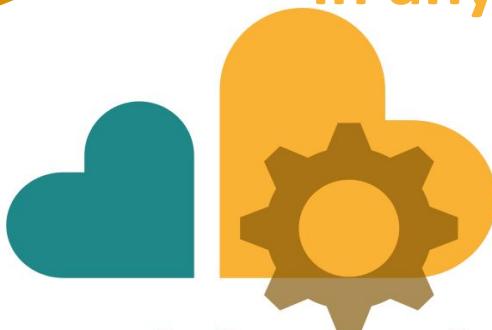




This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

Beta Sept 2020

# Registry for Workflows Open to All in any country!



## WorkflowHub

<https://workflowhub.eu>

Open to any workflow platform, any subject, any person in any country

### WorkflowHub Club



Towards FAIR workflows and a FAIR registry  
Workflow Management System Agnostic

#### Find and access Workflows

- Registry & repository functionality
- Workflows may remain in their **native repositories** in their native form. Or can be deposited.
- Register (push) / Harvest (pull)

#### Workflows interoperability and reusability

- Using a **metadata standards framework**



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



**COVID-19 Data Portal**

1/3<sup>rd</sup> workflows COVID-19

WorkflowHub interface showing the VIRify pipeline details. It includes sections for Overview, Creators and Submitter, License, Activity, Tags, and Attributions.

**VIRify**  
**Inputs**

ID	Name	Description	Type
input_fasta_file	input_fasta	Set this parameter if the input fasta is mostly viral. See: <a href="https://github.com/mgnt/virify/issues/50">https://github.com/mgnt/virify/issues/50</a>	File
viroblast_virus_db	viroblast_virus	Set this parameter if the input virus database is not included in the viroblast database. Default value: 'viroblast_virus'.	boolean
viroblast_data_dir	viroblast_data	Viroblast supporting database files.	Directory
add_hmmdb_dir	add_hmmdb	Additional metadata file.	File
hmmscan_database_dir	hmmscan_database	Hmmscan HMM database (databases/virify/viridit_database). NOTE: It needs to be a full path.	Directory
ncbi_taxonomy_db	ncbi_taxonomy	etc/NCITaxa.db (https://github.com/mgnt/virify/blob/main/etc/NCITaxa.db) This file was manually built and placed in the corresponding path (in databases).	File
img_bias_database_dir	img_bias_database	Downloaded from: <a href="https://igenome.org/doeportal/viridit_VIRIFY_HM.hml">https://igenome.org/doeportal/viridit_VIRIFY_HM.hml</a>	Directory
mashmap_reference_file	mashmap_reference	MashMap Reference file. Use MashMap to generate a reference file.	File?
pprmeta_sing	PPR-Meta singularity sing file	PPR-Meta singularity sing file	File

**Steps**

ID	Name	Description
fastq_rename	Filter contigs	Default weight 14b: https://github.com/EBI-Metagenomics/virify-script-dockerfile
viroblast	Filter contigs	Default weight 14b: https://github.com/EBI-Metagenomics/virify-script-dockerfile

# Workflow Management System Agnostic, Degrees of support



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



## COVID-19-PE-ARTIC-ILLUMINA (v0.3)

Version 3 +

### COVID-19: variation analysis on ARTIC PE data

The workflow for Illumina sequenced ARTIC 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 ARTIC primer sequences off reads with the ivar package. In addition, this workflow uses ivar also to identify amplicons affected by ARTIC primer-leaving site mutations and, if possible, uses reads derived from such "barred" amplicons when calculating allele-frequencies of other variants.

SEEK ID: <https://workflowhub.eu/workflows/110?version=3>

#### Inputs

ID	Name	Description	Type
Paired Collection	Paired Collection	Illumina reads from ARTIC assay with fastqparser encoding	raw
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 2 isolate Wuhan-Hu-1, complete genome	raw
ARTIC primer BED	ARTIC primer BED	BED file containing ARTIC primer positions. Can be retrieved from: <a href="https://usegalaxy.eu/u/worfogt/master/covid19-resources">https://usegalaxy.eu/u/worfogt/master/covid19-resources</a>	raw
ARTIC primers to amplicon assignments	ARTIC primers to amplicon assignments	Used by ivar trim and ivar removeseq for assigning primers to amplicons. Should have one line of tab-separated primer names per amplicon. Can be retrieved from: <a href="https://usegalaxy.eu/u/worfogt/master/covid19-resources">https://usegalaxy.eu/u/worfogt/master/covid19-resources</a>	raw
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.	raw
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	raw
Minimum DP for amplicon bias correction	Minimum DP for amplicon bias correction	Minimum DP for amplicon bias correction	raw

At any given variant site use the amplicon bias-corrected read only if the depth of coverage of the site retains at least this value after amplicon removal.

Workflow visibility



← access mechanisms

← authors & credit

← licensing

← analytics



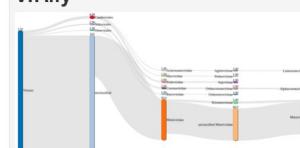
# nextflow

## VIRify Version 1

State

CWL 1.2.0-dev nextflow 20.01.0 uses docker uses conda build passing

### VIRify



VIRify is a recently developed pipeline for the detection, annotation, and taxonomic classification of viral contigs in metagenomic and metatranscriptomic assemblies. The pipeline is part of the repertoire of analysis services offered by MGNT. VIRify's taxonomic classification relies on the detection of taxon-specific profile hidden Markov models (HMMs), built upon a set of 22,014 orthologous protein domains and referred to as VPHOGs.

VIRify was implemented in CWL.

#### What do I need?

The current implementation uses CWL version 1.2 dev-2. It was tested using Toil version 4.10 as the workflow engine and conda to manage the software dependencies.

Tags: [SARS-CoV-2](#) [viroblast](#)

SEEK ID: <https://workflowhub.eu/workflows/27?version=1>

[View on GitHub](#) [Download RO-Crate](#) [Run on usegalaxy.eu](#)

Creators and Submitter

Creator

✓ Martin Beracochea, Martin Beracochea, Martin Höller, Alexandre Almeida, Guillermo Rangel-Pineros and Ekaterina Sakharkova

Submitter

✓ Laura Rodriguez-Navas

License

Apache Software License 2.0

Activity

Views: 563 Downloads: 24 Created: 8th Jun 2020 at 11:21 Last updated: 8th Jun 2020 at 13:56 Last used: 27th Nov 2020 at 15:04

Tags

ARTIC covid-19 covid19.galaxyproject.org

Attributions

None

[View on GitHub](#) [Download RO-Crate](#)

Creators and Submitter

Creator

✓ Martin Beracochea, Martin Höller, Alexandre Almeida, Guillermo Rangel-Pineros and Ekaterina Sakharkova

Submitter

✓ Laura Rodriguez-Navas

License

Apache Software License 2.0

Activity

Views: 798 Downloads: 55

Created: 8th Jun 2020 at 11:29

Last updated: 8th Mar 2021 at 21:57

Last used: 1st Jun 2021 at 02:51

Tags

covid-19

# GalaxyProject SARS-CoV-2

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

Space: COVID-19 Biohackathon

SEEK ID: <https://workflowhub.eu/projects/3>

Public web page: <https://github.com/galaxyproject/SARS-CoV-2>

Organisms: Homo sapiens, Sars-cov-2

## Related items

People (13)   Organizations (9)   Workflows (21)

Frederik Coppens



Teams: GalaxyProject SARS-CoV-2  
Organizations: ELIXIR Belgium  
[ID](https://orcid.org/0000-0001-8565-5145)

Flora D'Anna



Teams: GalaxyProject SARS-CoV-2  
Organizations: ELIXIR Belgium  
[ID](https://orcid.org/0000-0003-4665-6673)

Bert Droebeke



Teams: GalaxyProject SARS-CoV-2, Galaxy Training Network  
Organizations: ELIXIR Belgium  
[ID](https://orcid.org/0000-0003-0522-5674)

Dan Fornika



Teams: GalaxyProject SARS-CoV-2  
Organizations: BC Centre for Disease Control  
[ID](https://orcid.org/0000-0002-6178-3585)

WorkflowHub PALS: No PALS for this Team

Team created: 8th Apr 2020

Overview



## Workflows organized by:

Spaces  
Teams  
Collections  
Properties

- Tags
- Type
- Status
- Dates....etc

Makers are custodians of their own workflows

Preserve personal attribution, affiliations and contribution credit

## Search & Browsing

Expertise: Bioinformatics, Data management, Molecular biology  
Tools: Databases, PCR, Workflows, Web services



## Workflows

Query

Search here... [Go](#)

Created At

Any time

Workflow Type

Galaxy

Common Workflow Language

Nextflow

Snakemake

Tag

covid-19

Alignment

CWL

Assembly

INDELS

RNASeq

More...

Submitter

Bert Droebeke

Ambarish Kumar

Laura Rodriguez-Navas

Douglas Lowe

Sergi Sayols

WorkflowHub Bot

More...

Team

GalaxyProject SARS-CoV-2

IBBSA Workflows

BioBB Building Blocks

CWL workflow SARS-CoV-2

UNLOCK

IMBforge

More...

Space

Independent Teams

COVID-19 BioHackathon

BioExcel

nf-core

EOSC-Life

EOSC-Life-WP6

Creator

Jasper Koehorst

Genis Bayarri

Sergi Sayols

Bart Nijssse

Melchior du Lac

Adam Hospital

More...

Maturity

94 Workflows visible to you, out of a total of 97

Default

Condensed

Table

Protein Ligand Complex MD Setup tutorial using BioExcel Building Blocks (biobb)

Last update date (Descending)

Perpetual Development

# Ambition to add value



## Improve visibility and sharing

### Associated supplementary materials

- test, example data, documentation, publications

### Adding and enriching metadata & packaging

- Showing connections between workflows and sub-workflows
- Tracking workflows

### Linking to Monitoring services



### Supporting teams and collections

### Lifecycle support - versioning, snapshots, sub-workflow blocks

- Yellow pages
- Spaces
- Teams
- Organizations
- People
- Assets
- Data files
- SOPs
- Workflows
- Publications
- Documents
- Collections
- Activities
- Presentations
- Events
- Samples
- Organisms



# A registry Coupled to execution deploys and native repositories

Galaxy  
PROJECT



Global Alliance  
for Genomics & Health  
Collaborate. Innovate. Accelerate.

Tool Registry Service API

Genomics - PE Variation Version 1

Analysis of variation within individual COVID-19 samples using Illumina Paired End data. More info can be found at <https://covid19.galaxyproject.org/genomics/>

Workflow Inputs:

- Paired Collection (fastqparser)
- GenBank file

Workflow Steps:

1. fastqparser
2. SnpEff build
3. SnpEff
4. Map with BWA-MEM
5. MultiQC
6. Filter SAM or BAM, output SAM or BAM
7. Samtools stats
8. MarkDuplicates
9. MuRQC
10. Realoads
11. MuRQC
12. Call variants
13. SnpEff
14. Snpeff Extract Fields
15. Collapse Collection

SEEK ID: <https://workflowhub.eu/workflows/?version=1>

Inputs

ID	Name	Description	Type
#main/GenBank file	n/a	n/a	File
#main/Paired Collection (fastqparser)	n/a	n/a	File

Outputs

ID	Name	Description	Type
#main/GenBank file	n/a	n/a	File
#main/Paired Collection (fastqparser)	n/a	n/a	File



Galaxy Europe

Analyze Data Workflow Visualize Shared Data Help User

Workflow: COVID-19 - Genomics [4] PE Variation (imported from uploaded file)

Run Workflow

History

search datasets

Unnamed history (empty)

This history is empty. You can load your own data or get data from an external source

Tools

search tools

Upload Data

Get Data

Send Data

Collection Operations

GENERAL TEXT TOOLS

Text Manipulation

Filter and Sort

Join, Subtract and Group

GENOMIC FILE MANIPULATION

Convert Formats

FASTA/FASTQ

FASTQ Quality Control

Quality Control

SAM/BAM

BED

VCF/BCF

Nanopore

COMMON GENOMICS TOOLS

Operate on Genomic Intervals

Fetch Sequences / Alignments

History Options

Send results to a new history

No

D 1: GenBank file

No genbank or genbank.gz dataset available.

C 2: Paired Collection (fastqparser)

No data dataset collection available.

F 3: SnpEff build: (Galaxy Version 4.3+galaxy4)

F 4: fastp (Galaxy Version 0.19.5+galaxy1)

F 5: Map with BWA-MEM (Galaxy Version 0.7.17.1)

F 6: MultiQC (Galaxy Version 1.7.1)

F 7: Filter SAM or BAM, output SAM or BAM (Galaxy Version 1.8+galaxy1)

SAM or BAM file to filter

Header in output

Include header

View on GitHub

Download RO-Crate

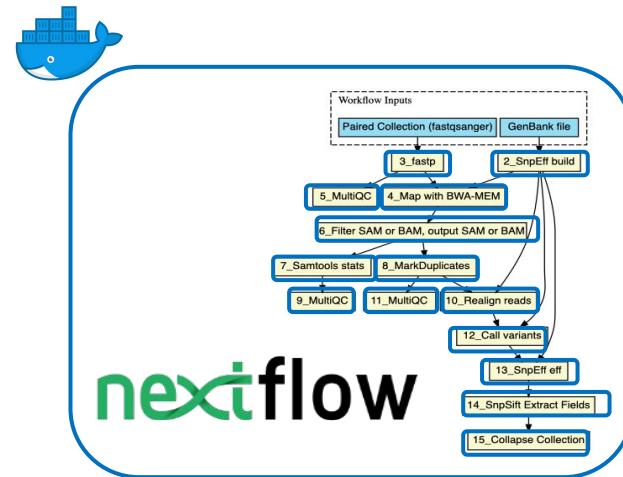
Run on usegalaxy.eu

# A Generous Registry



**Support for workflows in many forms** from scripts, over notebooks to Workflow Management Systems

**Stimulating best practices** from base FAIR to enriched towards reuse & reproducibility  
**Enabling diverse use cases** from native, over containers as building blocks to fully encapsulated in a container



# Snakemake



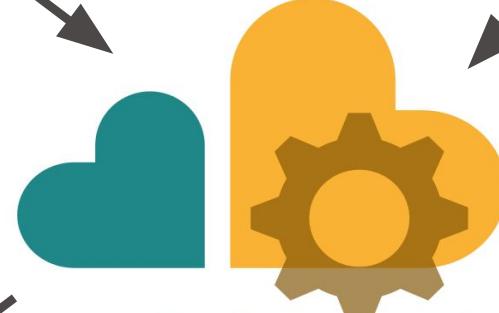
# Workflows as entry point for the community



EOSC-Life



Australian  
**BioCommons**



# WorkflowHub



COMMON  
WORKFLOW  
LANGUAGE

<your favorite>

 Galaxy  
PROJECT

# Snakemake



BioCONTAINERS



## Dockstore

DockerHub, Quay.io, ...



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

15

# FAIR is about machine processable metadata



COMMON  
WORKFLOW  
LANGUAGE



ED  
AM



Bioschemas  
[schema.org](http://schema.org)

EOSC-Life PROV

## Canonical workflow description

Native executable linked to containers  
Abstract CWL



{} **Workflow**  
**RO-Crate**

## Type the input and outputs of the steps

Ontology of types of data and data identifiers, data formats, operations in life sciences



## Common Metadata for registration and discovery

Schema.org profile and types  
ComputationalWorkflow, FormalParameter  
ComputationalTool

Package a workflow, its documentation and its components, with associated metadata into a citable object.

## Provenance

CWLProv  
EOSC-Life Common Provenance Model



**Reporting, Exchange and Archive format**  
**Carrier of metadata**

**Run Record format**



# FAIR is about Machine Processable Metadata

Working with WfMS for auto-extracting metadata, abstract CWL generation

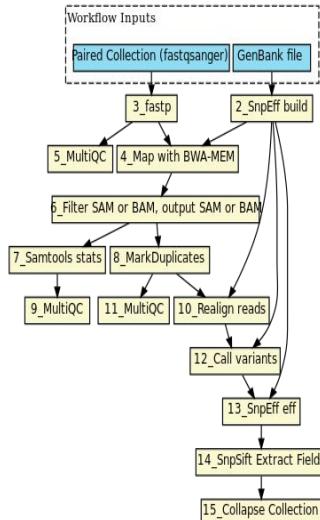


COMMON  
WORKFLOW  
LANGUAGE

Canonical description  
of the workflow  
Links to containerised tools

Abstract CWL & “Full fat”  
CWL alongside the native  
description

Abstract CWL describes just  
the structure and steps as a  
canonical description  
Links out to bio.tools and  
biocontainers



Bioschemas

Metadata about a workflow for registering it and  
finding it  
Used by Registries and Workflow-RO-Crate

Name	Cardinality	Type			
creator	Many	Organization or Person			
dateCreated	One	Date or DateTime			
input	Many	FormalParameter			
license	Many	CreativeWork or URL			
name	One	Text	Name	Cardinality	Type
output	Many	FormalPara	additionalType	One	URL to subclasses of http://edamontology.org/data_0006
programmingLang	Many	ComputerL	encodingFormat	Many	Text or URL to subclasses of http://edamontology.org/format_1915
sdPublisher	One	Organizatio			
url	One	URL	name	One	Text
version	One	Number or	description	One	Text
defaultValue	One	Thing			
Identifier	Many	PropertyValue or Text or URL			
valueRequired	One	Boolean			

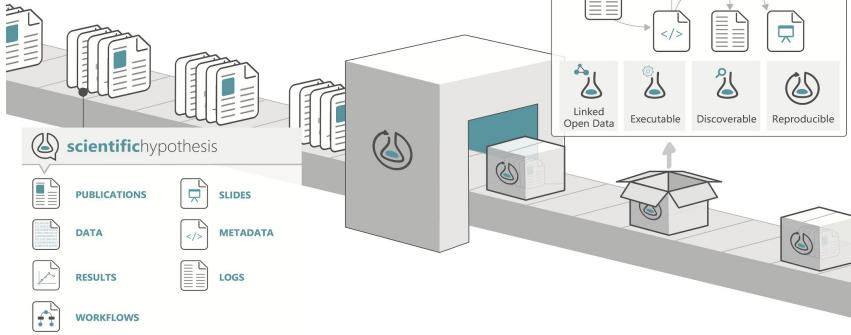


# FAIR is about Machine Processable Metadata

## Working with WfMS and services to exchange metadata objects



 Enabling reproducible, transparent research.



### Metadata Objects

Aggregate **files**, any **URI-addressable content**,  
**another RO-Crate**, along with contextual  
information, into an RO-Crate.  
*RO-Crate has its own pid and metadata.*

**Developer friendly**  
**Standard Web Native PIDs + JSON-LD +**  
**Schema.org, off the shelf archiving formats**

**Self-describing and open-ended**  
*Typed by **profiles** + add more schema.org  
and domain ontologies*

**Anything referenceable anywhere** using **PIDs**  
and **metadata held alongside heterogeneous**  
**data**

**Infrastructure independent** exchange between  
repositories, registries and services, and avoid  
vendor lock-in



## Aside: RO-Crate in Earth Science



### Earth Science RO-Crate profile (ongoing work)

Data Cubes (Collections and Products) are very common in Earth Science research

We create the concept of **Data Cube Data Entity**

Relevant **metadata** for a Data Cube

- Simple **geolocation** (sch:contentLocation) and **temporal coverage** (sch:temporalCoverage)
- **Spatial coverage** (extent, resolution, crs)
- **Temporal coverage** (startTime, endTime, resolution-step)
- **Vertical coverage** (highestElevation, lowestElevation, resolution-step)
- Other (processing level, processor, instrument, platform, etc.)

[http://www.rohub.org/rodetails/InSAR\\_GPS\\_Campi\\_Fregrei\\_2011\\_2013-release/](http://www.rohub.org/rodetails/InSAR_GPS_Campi_Fregrei_2011_2013-release/)



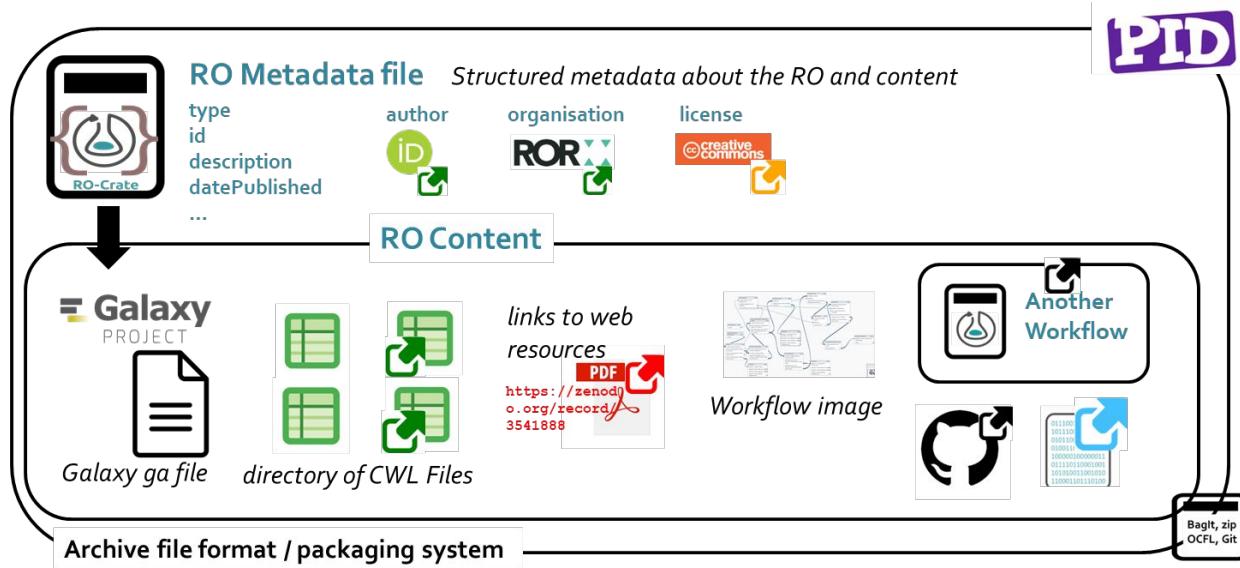
A valid RELIANCE RO-Crate JSON-LD graph MUST describe:

1. The [RO-Crate Metadata File Descriptor](#)
2. The [Root Data Entity](#)
3. One or more **Data Cube Data Entities**
4. Zero or more [Data Entities](#)
5. Zero or more [Contextual Entities](#)



# FAIR is about Machine Processable Metadata

## Working with WfMS and services to exchange metadata objects



**Bundles** descriptions, references, files  
**Adds** context, provenance, examples, data  
**Relates** data collections, SOPs, lab protocols  
**Links** descriptions with native workflows

Metadata packaging framework

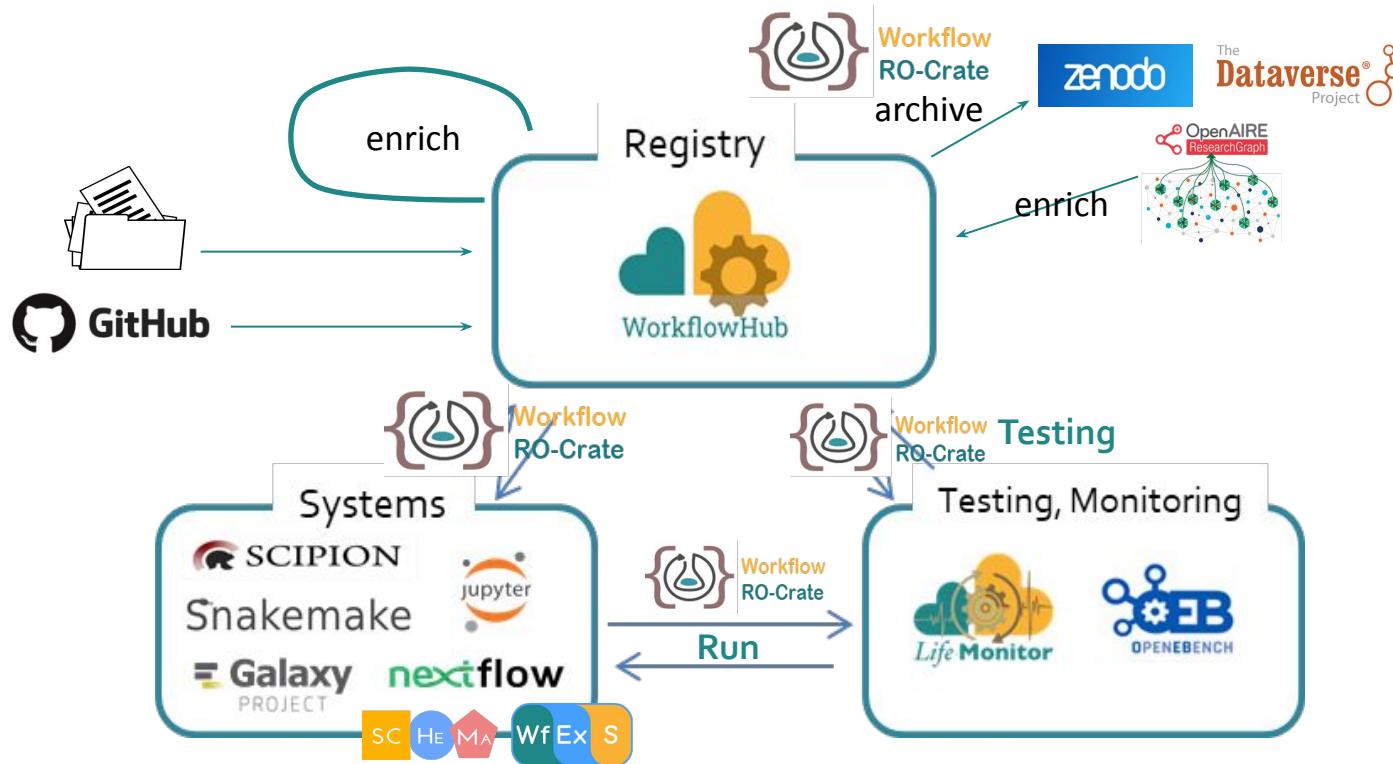
References, files and directories

Used for  
Exchange, Archiving, Reporting,  
Citation ...



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

# The Crates are specialised, added to, extended, enriched to get FAIRer. WorkflowHub is a RO-Crate factory



# FAIR workflows are processual digital objects properties of data and software



**Composite structure - FAIR all the way down**

**Living remixable artefacts - with limited lifespans**

**Have agency – the things they call FAIR?**

**Usable not just reusable**

**Forms** specification, implementation, instantiation,  
run result

**Workflows as FAIR Digital Objects:** Data-like properties as method objects.  
The principles can be adapted.

**Workflows as FAIR Software:**  
FAIR+R & FAIR+.  
The principles can be revised.

RDA/ReSA FAIR4RS WG

[First Draft of FAIR4RS principles](#)

Lamprecht, Anna-Lena et al. DOI: 10.3233/DS-190026  
Goble et al doi: 10.1162/dint\_a\_00033



# FAIR workflows as processual digital objects

## properties of data and software



### Composite structure: FAIR all the way down

tools, sub-workflows, ensembles  
granularity, dependencies  
FAIR inheritance and compatibilities

### Living artefacts with limited lifespans

versioned, forked, cloned  
recycled, repurposed, remixed

### Usable: Role of quality, maintainability, maturity, reproducibility, testing

### Agency

call on software, tools and other workflows, that are FAIR?

### FAIR and Share which forms and parts?

specification with test or exemplar data

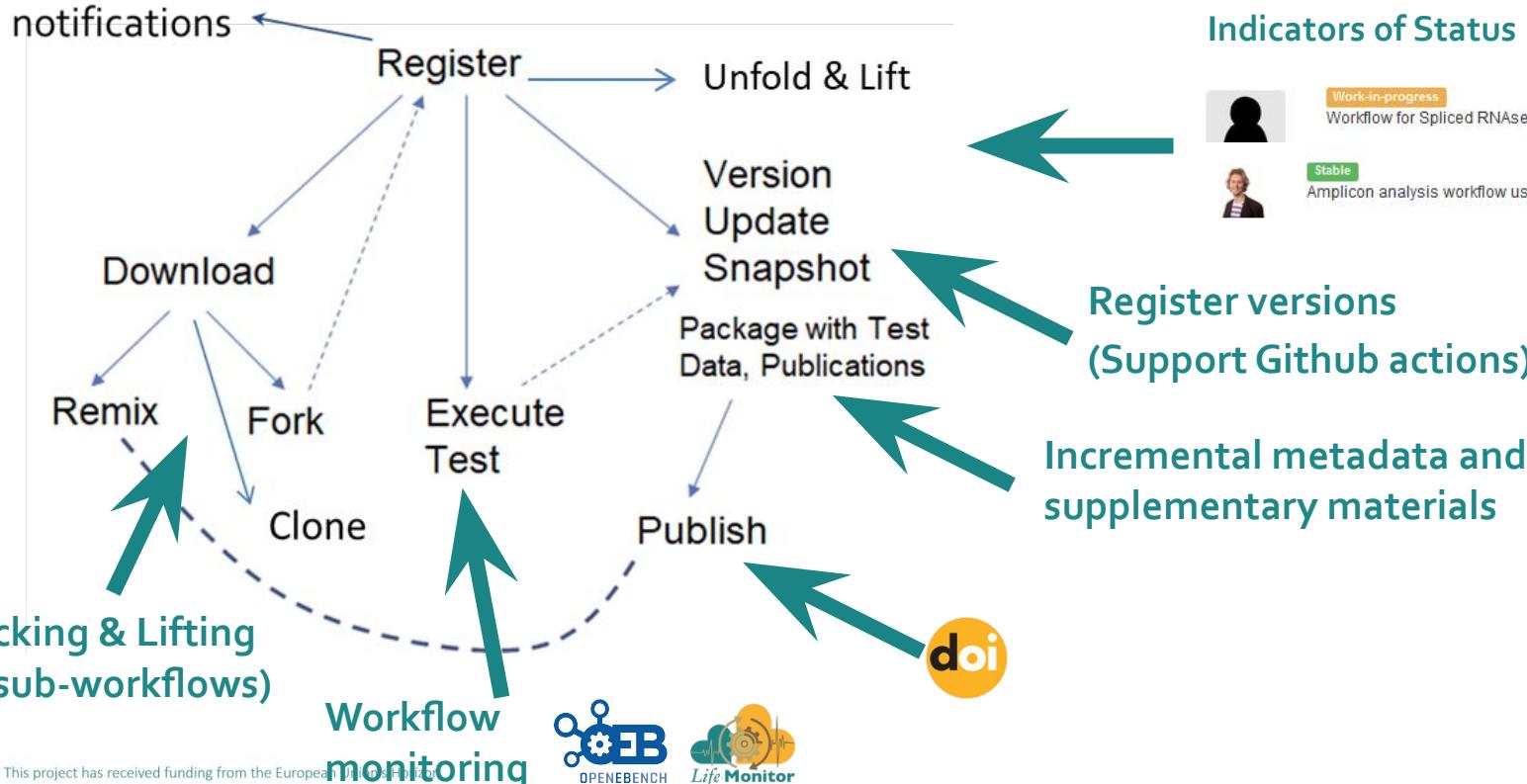
implementation in a WfMS

instantiation with input data, parameters set, containers deployed, computation ready

run result, intermediate/final data products, provenance logs?



# FAIR Workflow are FAIR Software living and with dependencies...workflow history/provenance



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

# Findable in WorkflowHub



## Identifiers

- Versioned URLs per workflow (incl. drafts)
- DOI assignment for workflows  
<https://doi.org/10.48546/workflowhub.workflow.29.2>

## Self-describing metadata

- Bioschemas mark up
- RO-Crate Archive of workflow with metadata
- Datacite DOI registration

## Facetted browsing & search

- Different organisational views and tags

**F1. metadata and workflows are assigned a globally unique and persistent identifier**

**F2. workflows are described with rich metadata**

**F3. metadata clearly and explicitly include the identifier of the workflow it describes**

**F4. metadata and workflow are registered or indexed in a searchable resource**



# Accessible from WorkflowHub

A1. **metadata and workflows** are **retrievable** by their identifier using a standardized comms protocol

## Retrievable Data & Metadata

- Direct Web access, View on GitHub
- **RO-Crate archive**
- DOI metadata
- GA4GH TRS API and WFF API

## Open protocols

- WorkflowHub API, OpenAPI, GitHub API, TRS API

## Optional: Federation and authorisation

- Most workflows are public
- Rich sharing/visibility permissions
- Restricted access possible via WFHub API

A2. **metadata** are accessible, even when the workflow is no longer available

## Metadata Preservation

- **RO-Crate archive** preserves metadata **and** workflow
- (Core metadata is preserved in DataCite)

## Preservation of metadata independent of WorkflowHub

- **RO-Crate archive** & republish in a long-term archive



# Workflow Interoperability Object or Software?



- I1. metadata and workflows use a formal language for knowledge representation.
- I2. metadata and workflows use vocabularies that follow FAIR principles
- I3. metadata and workflows include qualified references to other metadata and workflows

- I1. The workflow interoperates with other workflows & tools through (i) exchanging (meta)data, (ii) APIs.
- I2. Workflows read, write and exchange data meeting domain community standards
- I3. Workflows include qualified references to other objects\*



Bioschemas.org  
schema.org



W3C PROV



JSON-LD



Command-line utilities to assist in developing Galaxy and Common Workflow Language tools



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

\*FAIR4RS Proposed Principles for FAIR Software

# Workflow Reusability\*



**Reusable (can be understood, modified, built upon or incorporated into other workflows)**

**Workflows** are described in detail

- Lifting metadata out of systems
- Enriching metadata by other services

Track **versions** of workflows

- Each version archived as separate DOI/RO-Crate
- Not everyone uses GitHub...

How to **design for reuse**?

- Libraries of tested “canonical” workflows
- MGnify, BioExcel BioBuilding Blocks...

**Usable (can be executed)**

RO-Crate download, link to containers

TRS API integration with **execution** platforms

- UseGalaxy.\*
- WfExS fetches RO-Crate for general execution

Testing and monitoring



Multiple wf/test backends: Galaxy, Pandemo, CWL, Jenkins ....



Check workflow performance  
Provenance on containers,  
memory usage etc

\*FAIR4RS Proposed Principles for FAIR Software



# FAIR Data for/from workflows

## FAIRification of Workflows



Can we indicate how a workflow follows FAIR Data principles?

- To what extent does the Workflow consume, propagate and generate FAIR **identifiers** for data?
- Does the workflow use proprietary formats?
- What is the **license** of data outputs from the workflow?
- What are the **usage restrictions** on the reference data it needs?
- How does the workflow **access** FAIR data?
- Are **parameters validated** to preclude workflow failure and faulty/unsafe results ?
- Are the **processing steps black boxes**?
- Can we **fully track data provenance** through the workflow?

Challenge of diverse API & AAI landscape, formats and packaging

Practical guides and checklists

Open-up to scrutiny and peer review

Build reviewed libraries of workflows, sub-workflows & tools



# Conclusion

## Workflows as Data and Software Objects



- Revised principles, developer friendly guides and best practice

## FAIRification of workflows

It takes a village of services, WfMS, repositories, people

- Lightweight extensible metadata framework
- Metadata collection and enrichment from all sources



## One size does not fit all

- From scripts to the fully fledged systems
- Not all use GitHub or Containers
- FAIR spectrum

Moving up the FAIR ladder





# <https://about.workflowhub.eu/community/>

- Alan R Williams (The University of Manchester)
- Alexander Vasilenko (VKM IBPM RAS, MIRRI)
- Alexander Kanitz
- Alban Gaignard
- Ambarish Kumar (Jawaharlal Nehru University, New Delhi, India)
- Antonio Rosato
- Bert Droebeke (ELIXIR-BE, VIB-UGent Center for Plant Systems Biology)
- Björn Grüning (University of Freiburg, ELIXIR-DE, Galaxy Project)
- Carole Goble (The University of Manchester, ELIXIR-UK)
- Carlos Oscar Sorzano (CNB CSIC)
- Castrense Savojardo (ELIXIR-IT)
- Dan Fornika
- Decruw Cedric
- Djura Smits (Netherlands eScience Center)
- Emidio Capriotti (UNIBO, Italy)
- Emmy Tsang
- Finn Bacall (The University of Manchester, ELIXIR-UK)
- Flora D'Anna
- Frédéric Lemoine (Institut Pasteur, Paris)
- Frederik Coppens (ELIXIR-BE, VIB-UGent Center for Plant Systems Biology, EOSC-Life)
- Georg Peiter
- Giacomo Tartari (ELIXIR-IT)
- Hervé Ménager (Institut Pasteur, ELIXIR-FR, bio.tools)
- Ignacio Eguinoa (ELIXIR-BE, VIB-UGent Center for Plant Systems Biology)
- Jennifer Harrow (ELIXIR-Hub, tools platform coordinator)
- Jon Ison (Institut Pasteur, ELIXIR-FR, bio.tools)
- José Mª Fernández (BSC, ELIXIR-ES)
- Kiran K Telukunta
- Lars Ridder (Netherlands eScience Center)
- Laura del Caño (CNB-CSIC, INSTRUCT, EOSC-Life)
- Laura Rodriguez-Navas (BSC, ELIXIR-ES)

- Jennifer Harrow (ELIXIR-Hub, tools platform coordinator)
- Jon Ison (Institut Pasteur, ELIXIR-FR, bio.tools)
- José Mª Fernández (BSC, ELIXIR-ES)
- Kiran K Telukunta
- Lars Ridder (Netherlands eScience Center)
- Laura del Caño (CNB-CSIC, INSTRUCT, EOSC-Life)
- Laura Rodriguez-Navas (BSC, ELIXIR-ES)
- Leyla Garcia
- Luca Pireddu (CRS4/BBMRI)
- Magnus Palmblad (LUMC, ELIXIR-NL)
- Marco Tangaro (ELIXIR-IT)
- Michael R. Crusoe (ELIXIR-NL, CWL, standards developer, interop cheerleader)
- Miguel Vazquez (BSC)
- Miriam Payá Milans (CBGP, Madrid)
- Munazah Andrabí
- Nick Juty (The University of Manchester, ELIXIR-UK)
- Paolo Romano
- Pier Luigi Martelli (ELIXIR-IT)
- Philipp G.
- Rena Bakhshi (Netherlands eScience Center)
- Rob Hooft
- Robin Richardson (Netherlands eScience Center)
- Romain Dallet (EMBC)
- Salvador Capella-Gutierrez (BSC, ELIXIR-ES)
- Simone Leo (CRS4/BBMRI)
- Sirarat Sarntivijai (ELIXIR-Hub)
- Stian Soiland-Reyes (The University of Manchester, ELIXIR-UK, BioExcel, CWL)
- Stuart Owen (The University of Manchester, FAIRDOM, ELIXIR-UK)
- Vahid Kiani
- Vincenzo Laveglia (CIRMMT Florence, EOSC-Life)
- Wolfgang Müller (HITS GmbH)
- Xiaoming Hu (HITS GmbH)



## WorkflowHub community

While WorkflowHub is largely developed as a collaboration between [several projects](#), any contributors are welcome to join our [open community](#).

## WorkflowHub Club

The weekly [WorkflowHub Club](#) is chaired by [Frederik Coppens](#).

- **Schedule:** Bi-weekly, Wednesday 10:00 BST / 11:00 CEST
- **Agenda / call details:** <https://s.apache.org/workflowhub-minutes>

Anyone can [join the WorkflowHub club](#)! Either sign up on GitHub issue #1 or join the [next call](#) and introduce yourself.

See [acknowledgements](#) for a complete list of participants in the WorkflowHub Club.

For asynchronous communication, see also:

- **Mailing list:** [workflowhub@elixir-europe.org](mailto:workflowhub@elixir-europe.org)  
([subscribe](#)/[archive](#))
- **Slack chat:** [#workflows on seek4science.slack.com](#) ([join](#))
- **Google Drive** (to request write-access, ask in Slack channel)

## Code of conduct

This project has a [Code of Conduct](#) to ensure interactions are friendly, respectful and inclusive. You can contact [info@escielab.org.uk](mailto:info@escielab.org.uk) if you have any concerns or questions.

# Links

EOSC-Life <https://www.eosc-life.eu/>

RO-Crate <https://www.researchobject.org/ro-crate/>

WorkflowHub <https://workflowhub.eu/>

Galaxy Europe <https://galaxyproject.eu/>

Bioschemas <https://bioschemas.org/>

Common Workflow Language <https://www.commonwl.org/>

# Bio.Tools and BioContainers



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