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

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

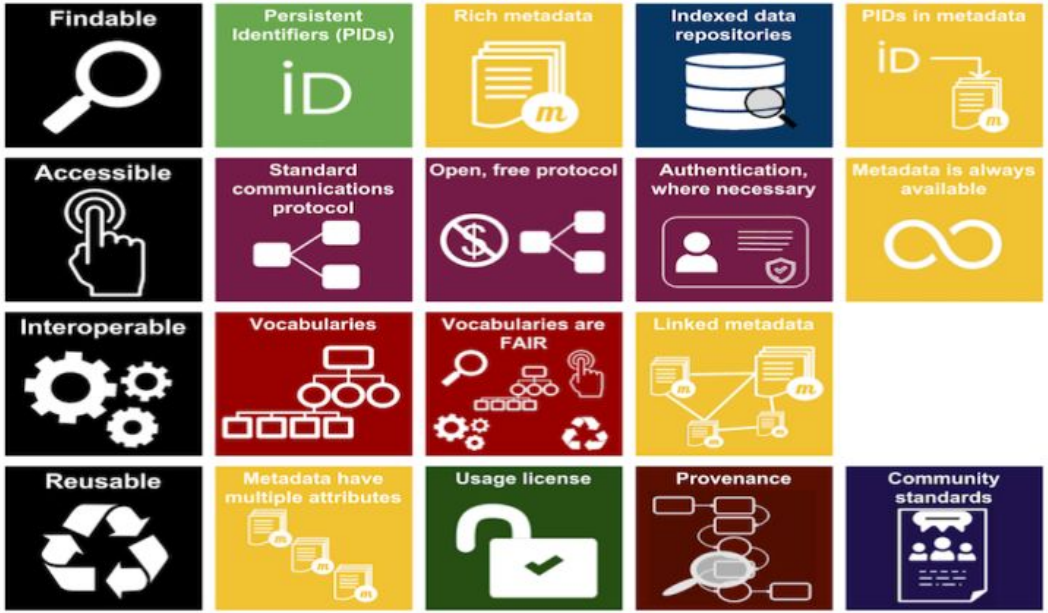
Frederik Coppens,
VIB / ELIXIR-BE

The rest of the WorkflowHub Club

GO-FAIR FAIR Workflows Webinar, 16th 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



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

tl;dr:

Persistent machine-readable and actionable metadata

Persistent identifiers

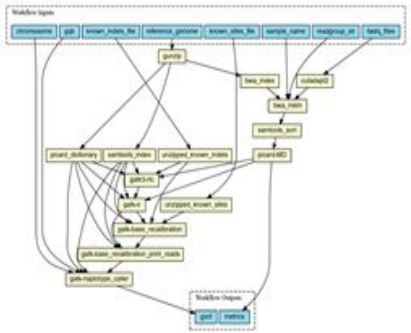
Clear licensing

Protocols for machine accessibility

Register / Index

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

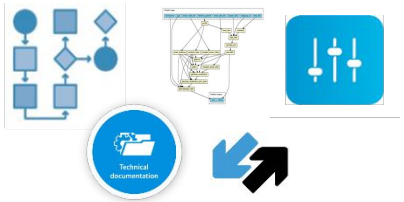
Computational Workflows in a nutshell



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

A special kind of software
Separation of the **workflow specification** from its **execution**

Specification description

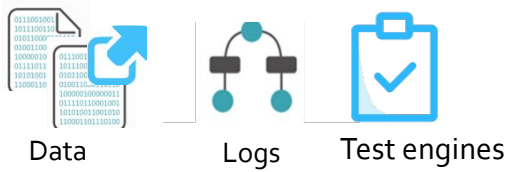


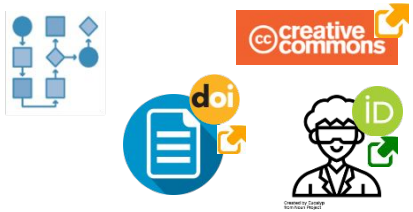
Precise description of a procedure: multi-step process coordinated by input/output data relationships (data types).

Software Execution

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

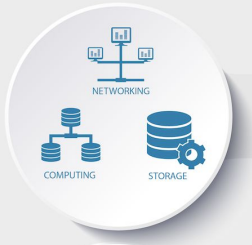
Associated Objects





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

European Open Science Cloud =



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

E-Infrastructure consolidation

+

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

Open Science

+

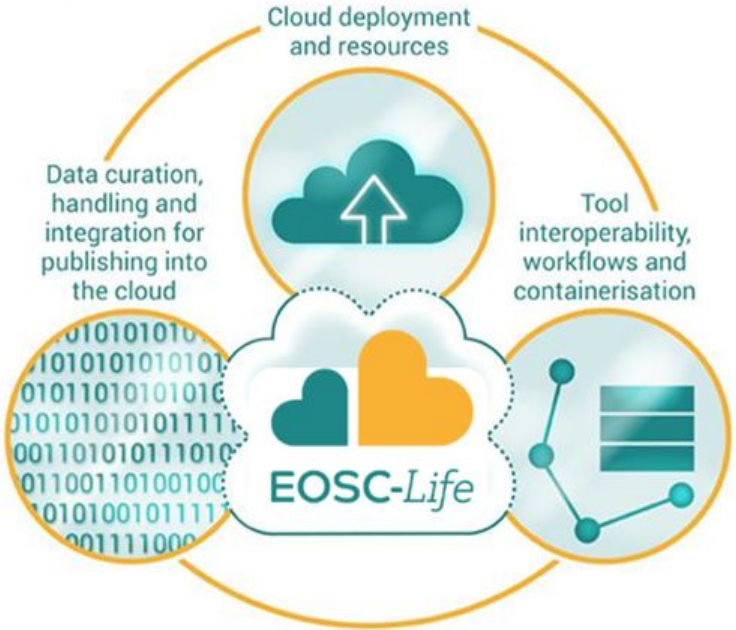
Populate EOSC with the scientific data resources and computational tools from research infrastructures – drive usage by to Europe's 1.7 M researchers

Scientific **C**ommunities' content and users

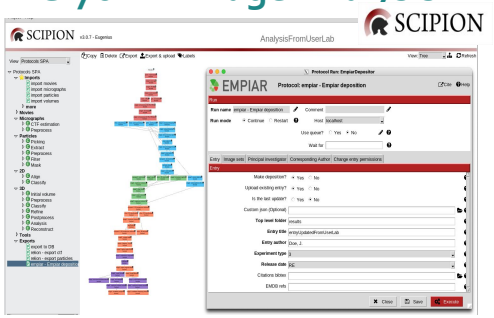
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

https://doi.org/10.5281/zenodo.3631247

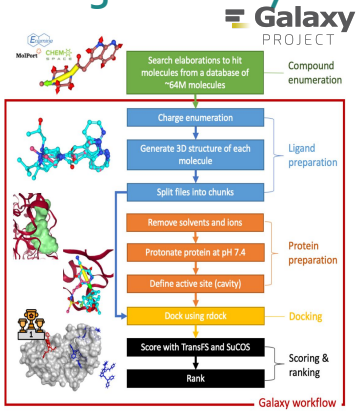
EOSC-Life pan-national thematic commons for bioscience data and methods



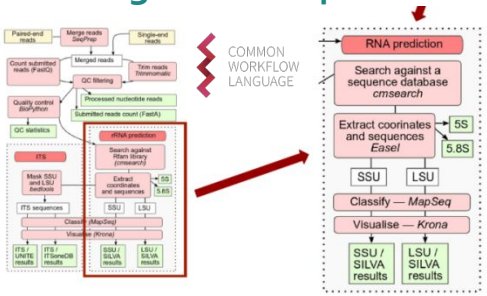
CryoEM Image Analysis



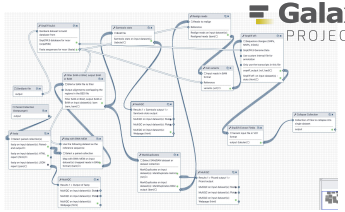
Drug Discovery



Metagenomic Pipelines



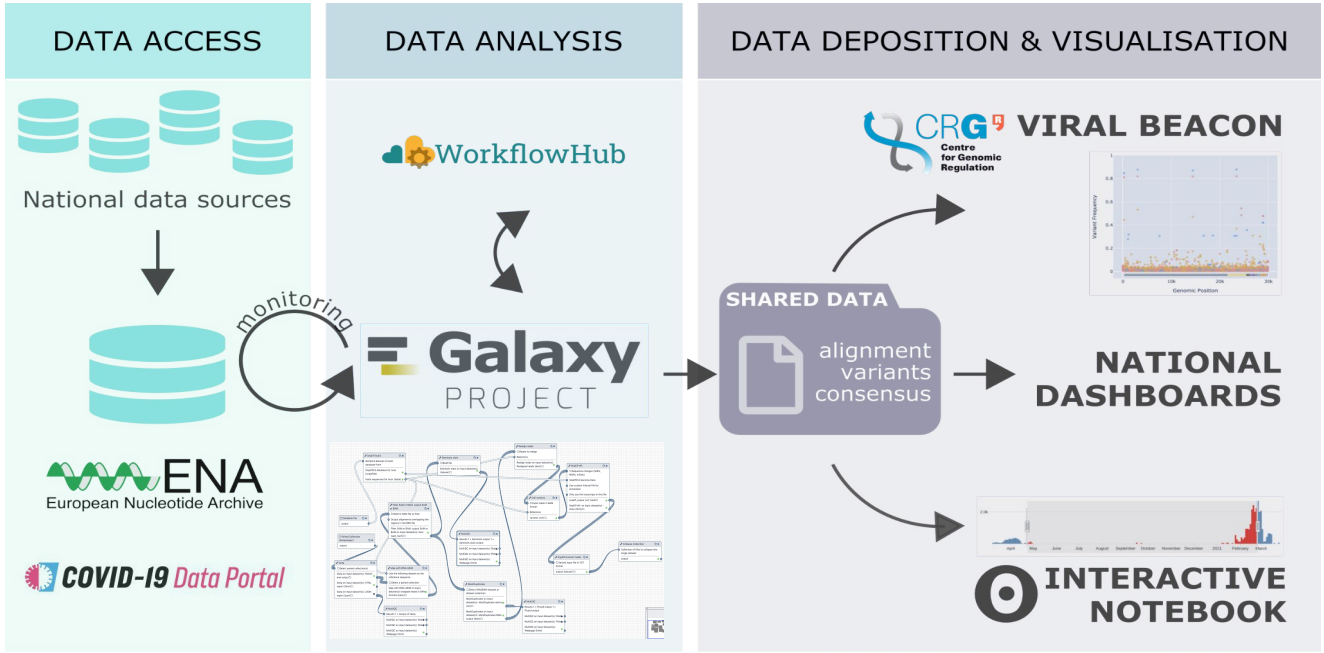
SARS-CoV-2 Monitoring



Using and sharing data, tools and workflows in the cloud

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

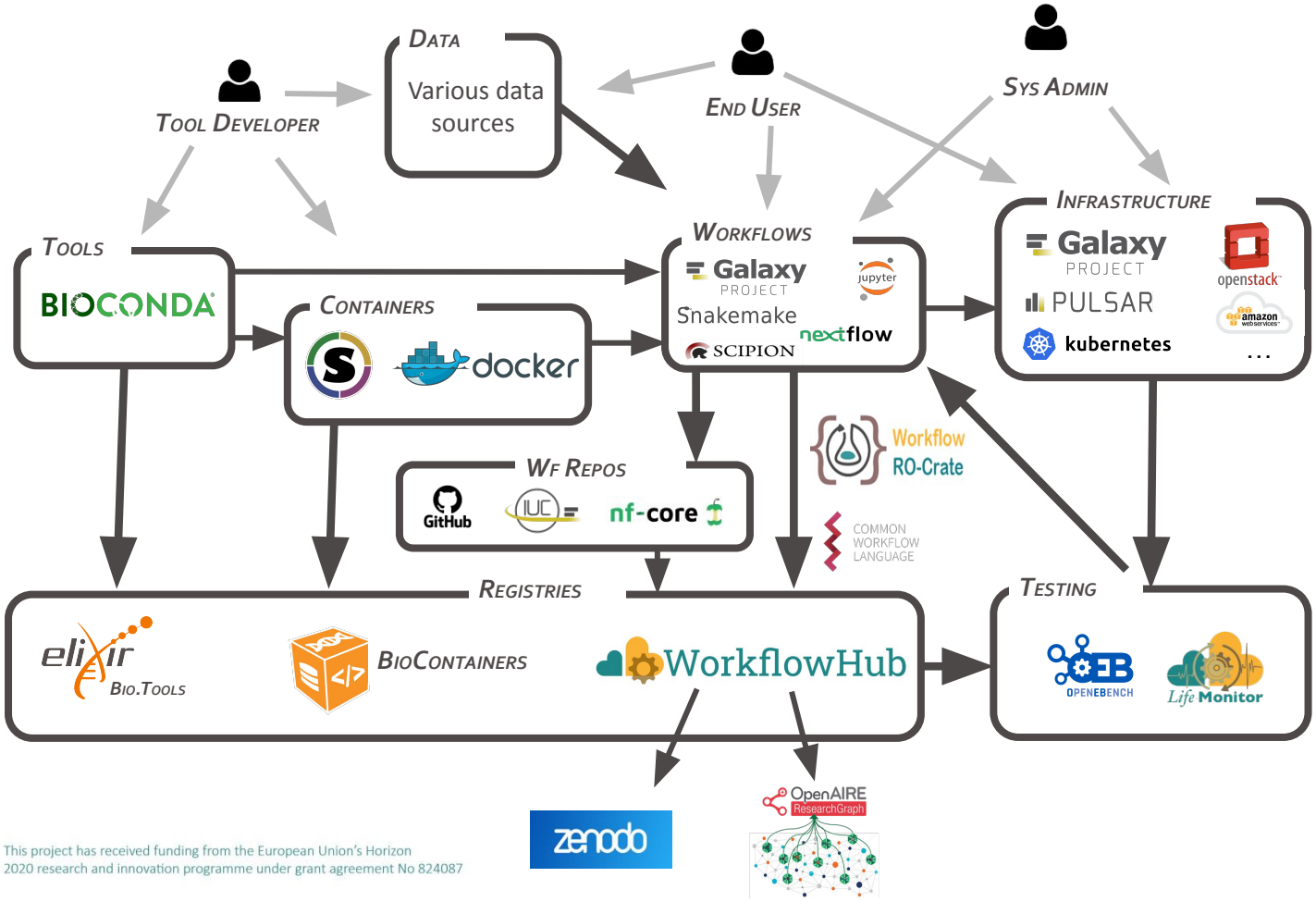
SARS-CoV-2 pre-processing, monitoring, analysis



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

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

EOSC-Life Workflow Collaboratory

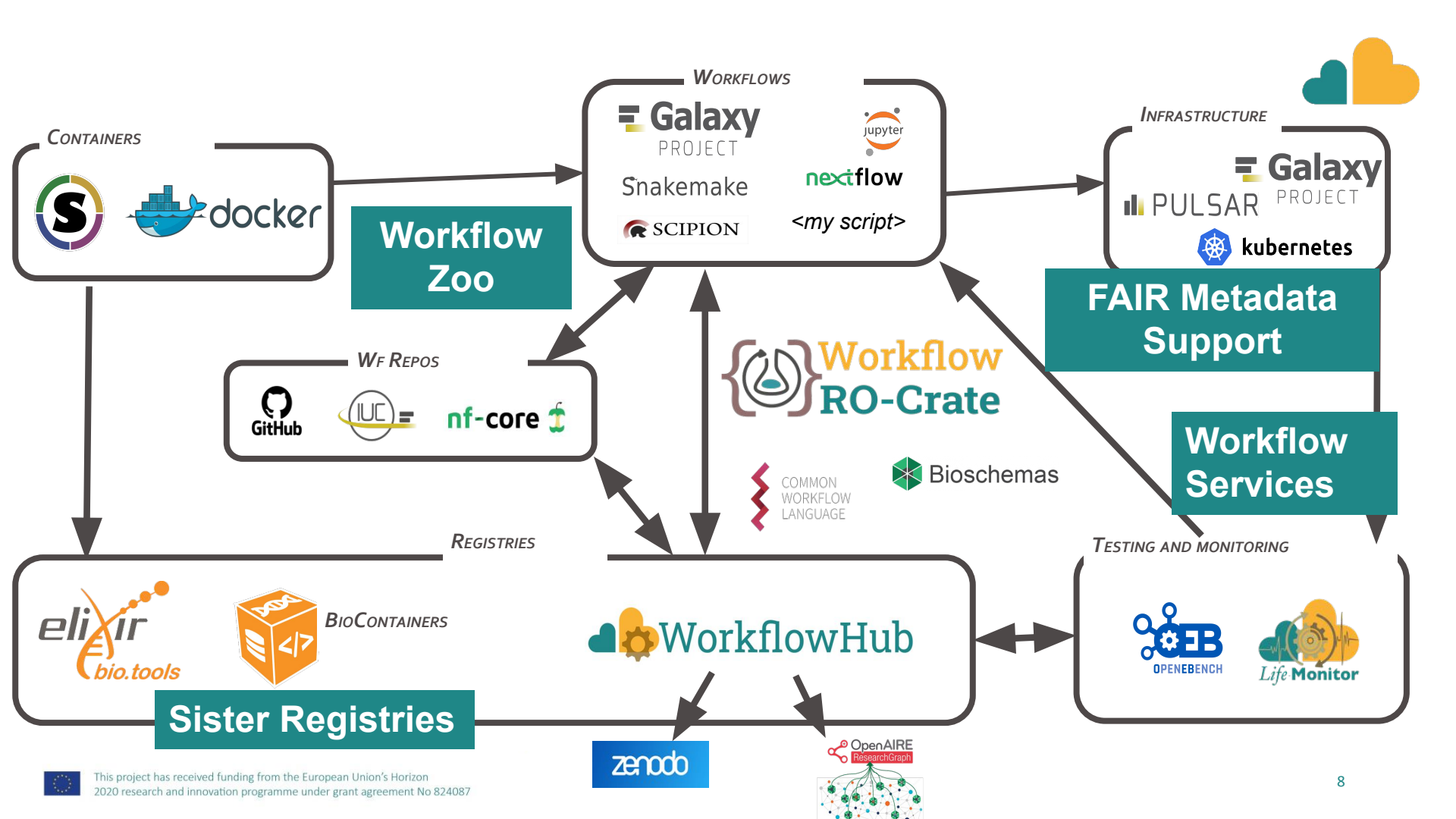


Authentication

elixir AAI

LS LOGIN

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!

EOSC-Life

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)

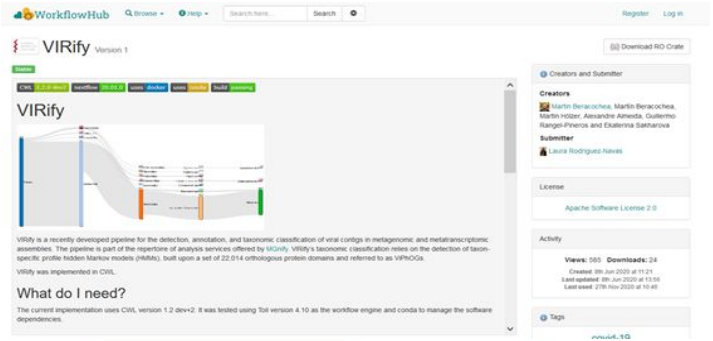
Workflows **interoperability** and **reusability**

- 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

1/3rd workflows COVID-19



SEEK ID: https://workflowhub.eu/workflows/26?version=1

Inputs

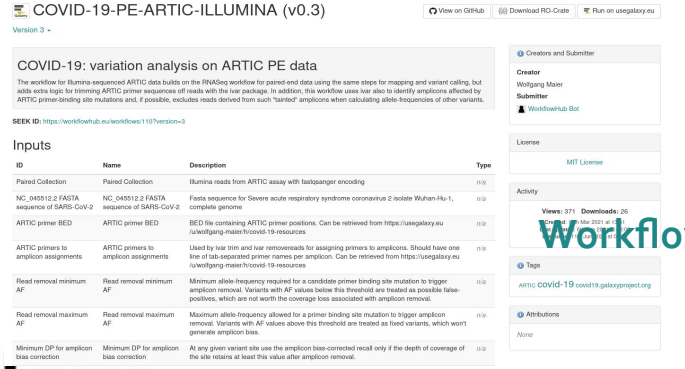
ID	Name	Description	Type
input_fasta_file	n/a	n/a	File
virsorter_virome	n/a	Set this parameter if the input fasta is mostly viral. See: https://github.com/simroux/VirSorter/issues/50	boolean
virsorter_data_dir	n/a	VirSorter supporting database files.	Directory
add_hmms_tsv	n/a	Additonal metadata tsv	File
hmmscan_database_dir	n/a	HMMScan Viral HMM (databases/vpHMM/vpHMM_database). NOTE: it needs to be a full path.	Directory
ncbi_tax_db_file	n/a	ete3 NCBITaxa db https://github.com/etetoolkit/ete/blob/master/ete3/ncbi_taxonomy/ncbiquery.py http://etetoolkit.org/docs/latest/tutorial/tutorial_ncbitaxonomy.html This file was manually built and placed in the corresponding path (on databases)	File
img_blast_database_dir	n/a	Downloaded from: https://genome.jgi.doe.gov/portal/IMG_VR/IMG_VR.home.html	Directory
mashmap_reference_file	n/a	MashMap Reference file. Use MashMap to	File?
ppr-meta_simg	n/a	PPR-Meta singularity simg file	File

Steps

ID	Name	Description
fasta_rename	Filter contigs	
[illegible]	[illegible]	Default lenght 1kb https://github.com/EBI-Metagenomics/emg-virify-scripts/issues/6
virfinder	VirFinder	

COMMON WORKFLOW LANGUAGE

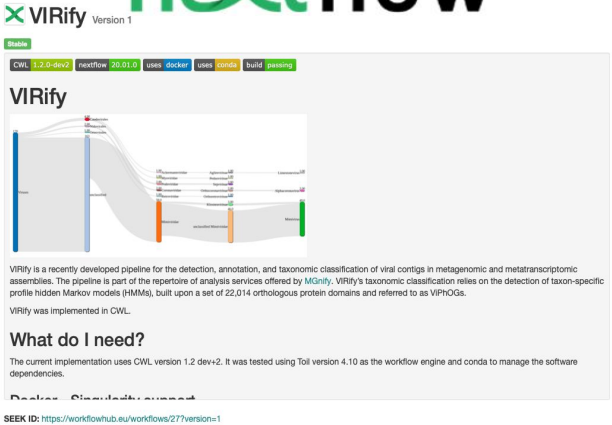
Galaxy PROJECT



WorkflowHub

Workflow visibility

nextflow



← access mechanisms

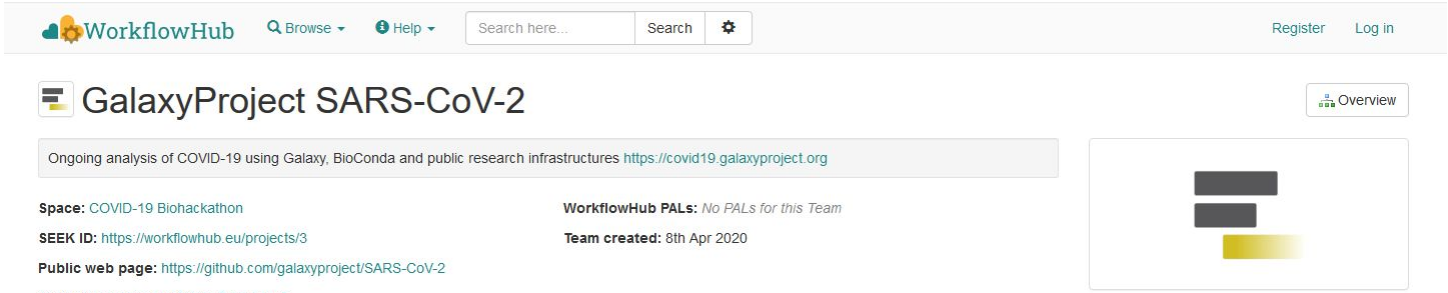
← authors & credit

← licensing

← analytics

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



Workflows organized by:

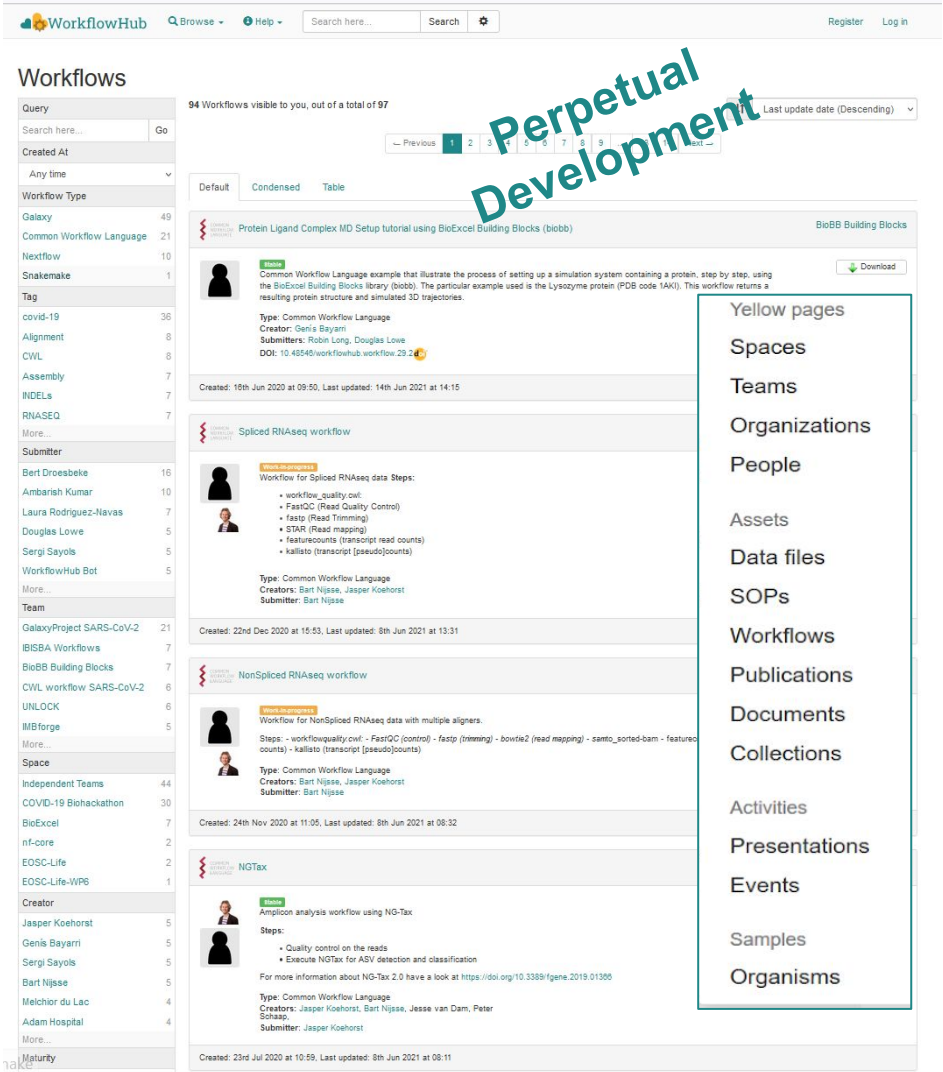
Spaces
Teams
Collections
Properties

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

Search & Browsing

Makers are custodians of their own workflows

Preserve personal attribution, affiliations and contribution credit



Ambition to add value

WorkflowHub

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

Life Monitor

OPENEBENCH

Supporting teams and collections

Lifecycle support - versioning, snapshots, sub-workflow blocks

A registry

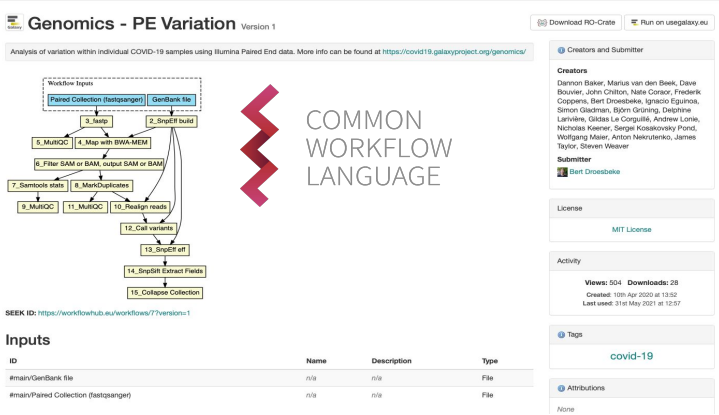
Coupled to execution deploys and native repositories

Galaxy PROJECT

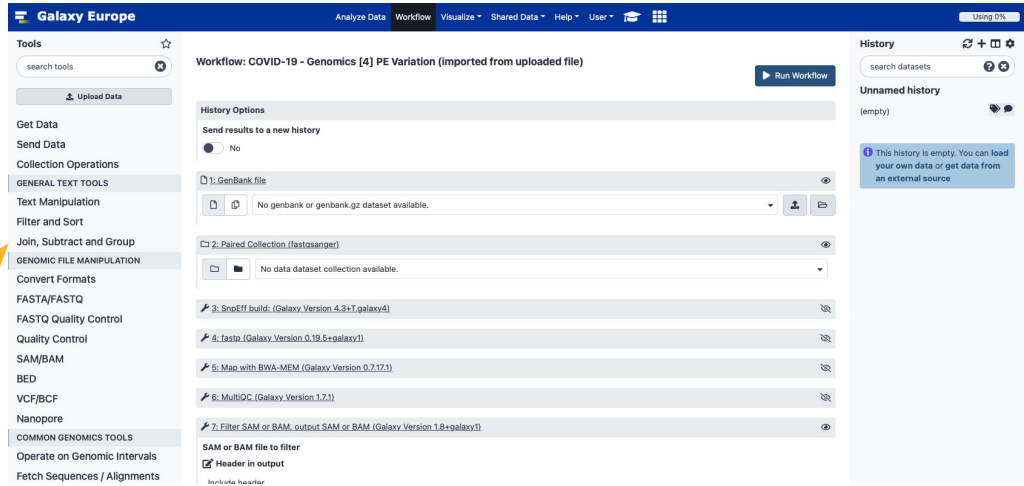
Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.

Workflow RO-Crate

Tool Registry Service API



COMMON WORKFLOW LANGUAGE



View on GitHub | Download RO-Crate | Run on usegalaxy.eu

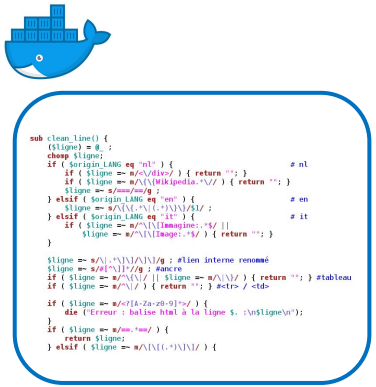
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

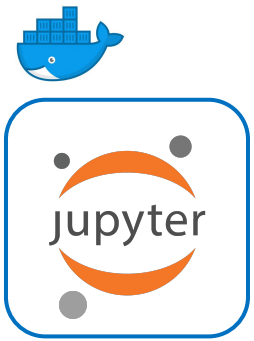
A Generous Registry

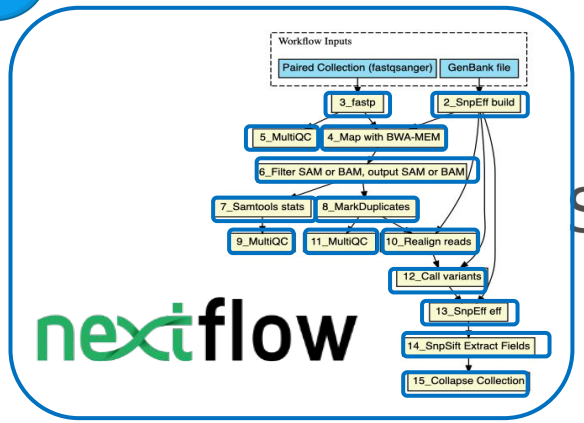
Support for workflows in many forms from scripts, over notebooks to Workflow Management Systems

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







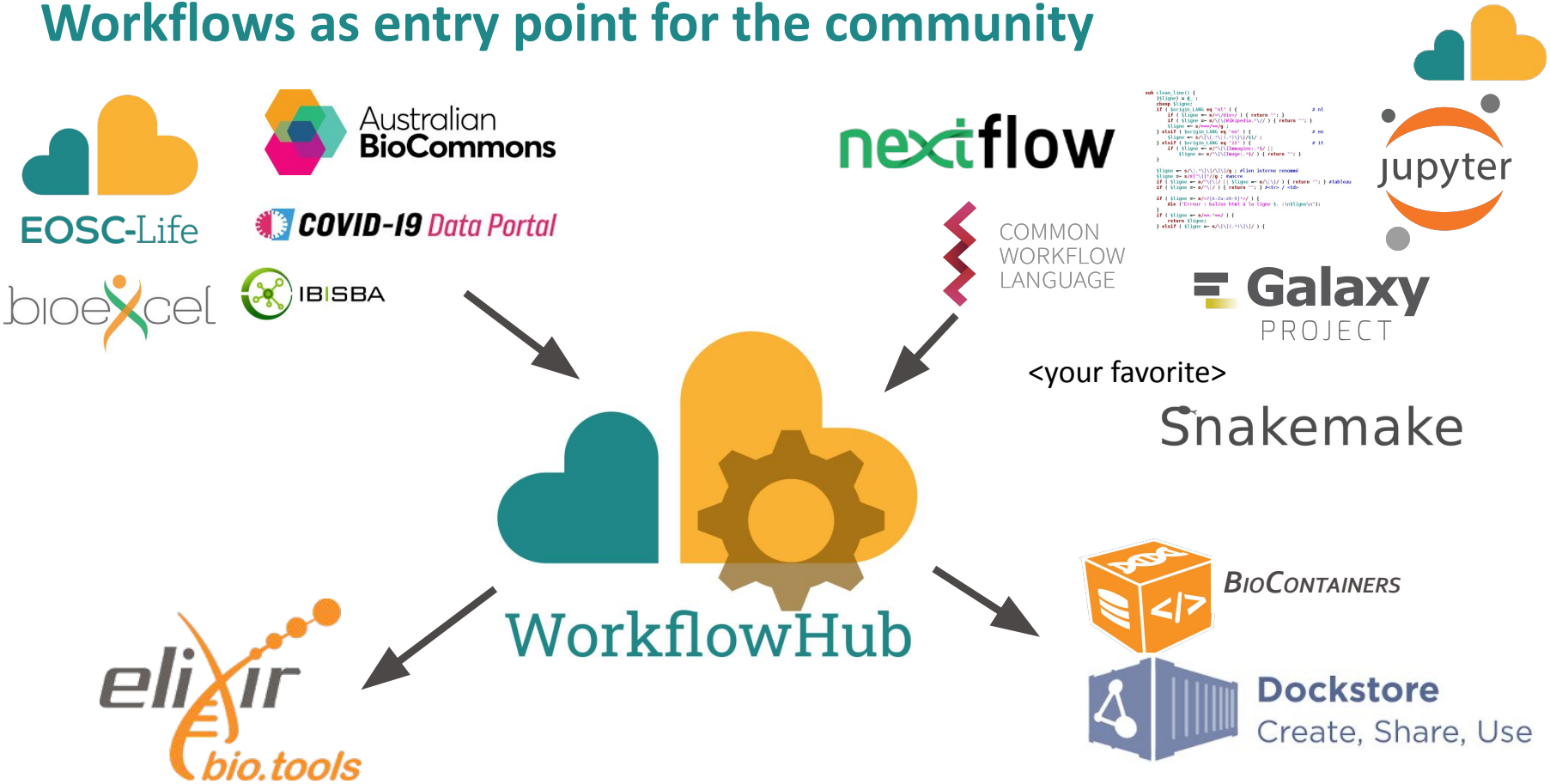
Galaxy PROJECT

Snakemake

COMMON WORKFLOW LANGUAGE

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

Workflows as entry point for the community



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

FAIR is about machine processable metadata

COMMON WORKFLOW LANGUAGE

EDAM

Bioschemas

schema.org

Canonical workflow description

Native executable linked to containers
Abstract CWL

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

Workflow RO-Crate

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

Reporting, Exchange and Archive format Carrier of metadata

EOSC-Life

W3C PROV

Provenance

CWLProv
EOSC-Life Common Provenance Model

Run Record format

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

FAIR is about Machine Processable Metadata

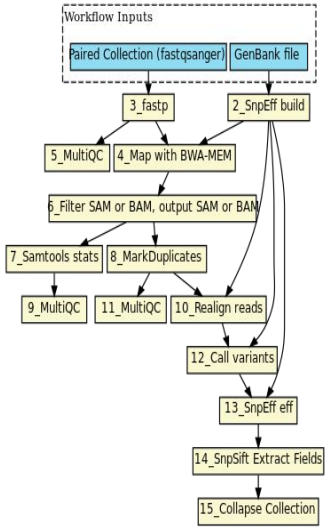
Working with WfMS for auto-extracting metadata, abstract CWL generation

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
output	Many	FormalPara
programmingLanguage	Many	ComputerL
sdPublisher	One	Organizatic
url	One	URL
version	One	Number or

Name	Cardinality	Type
additionalType	One	URL to subclasses of http://edamontology.org/data_0006
encodingFormat	Many	Text or URL to subclasses of http://edamontology.org/format_1915
name	One	Text
description	One	Text
defaultValue	One	Thing
Identifier	Many	PropertyValue or Text or URL
valueRequired	One	Boolean

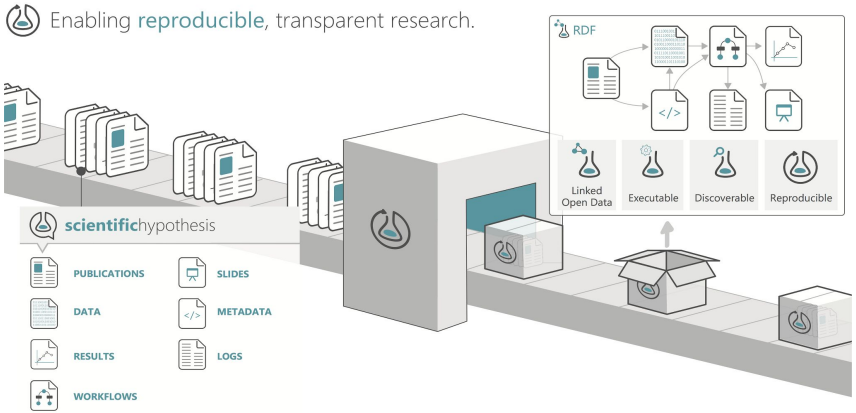
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

researchobject.org

FAIR is about Machine Processable Metadata

Working with WfMS and services to exchange metadata objects

RO-Crate



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*

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

https://www.researchobject.org/ro-crate/

Aside: RO-Crate in Earth Science

Earth Science RO-Crate profile (ongoing work)

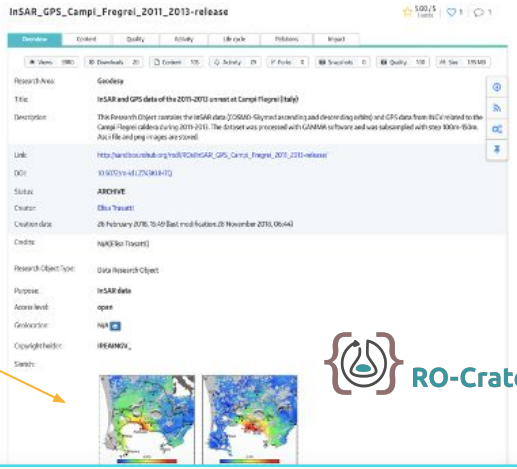
Data Cubes (**Collections and Products**) are very common in Earth Science research

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/



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

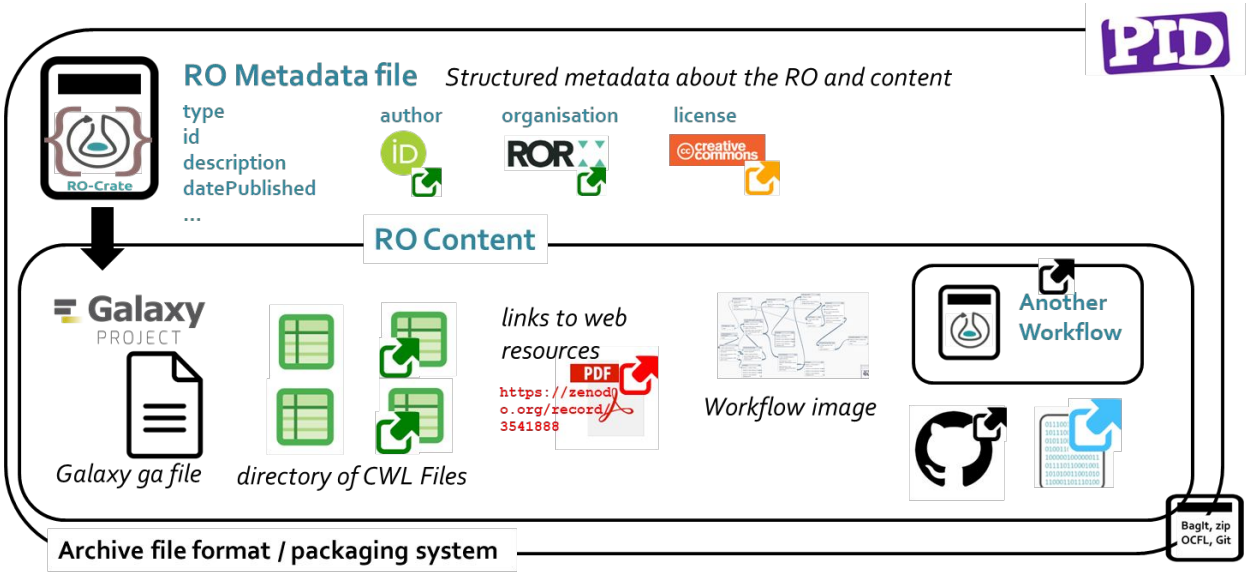
1. The RO-Crate Metadata File Descriptor
2. The Root Data Entity
3. One or more Data Cube Data Entities
4. Zero or more Data Entities
5. Zero or more Contextual Entities

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

FAIR is about Machine Processable Metadata

Working with WfMS and services to exchange metadata objects

Workflow RO-Crate



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

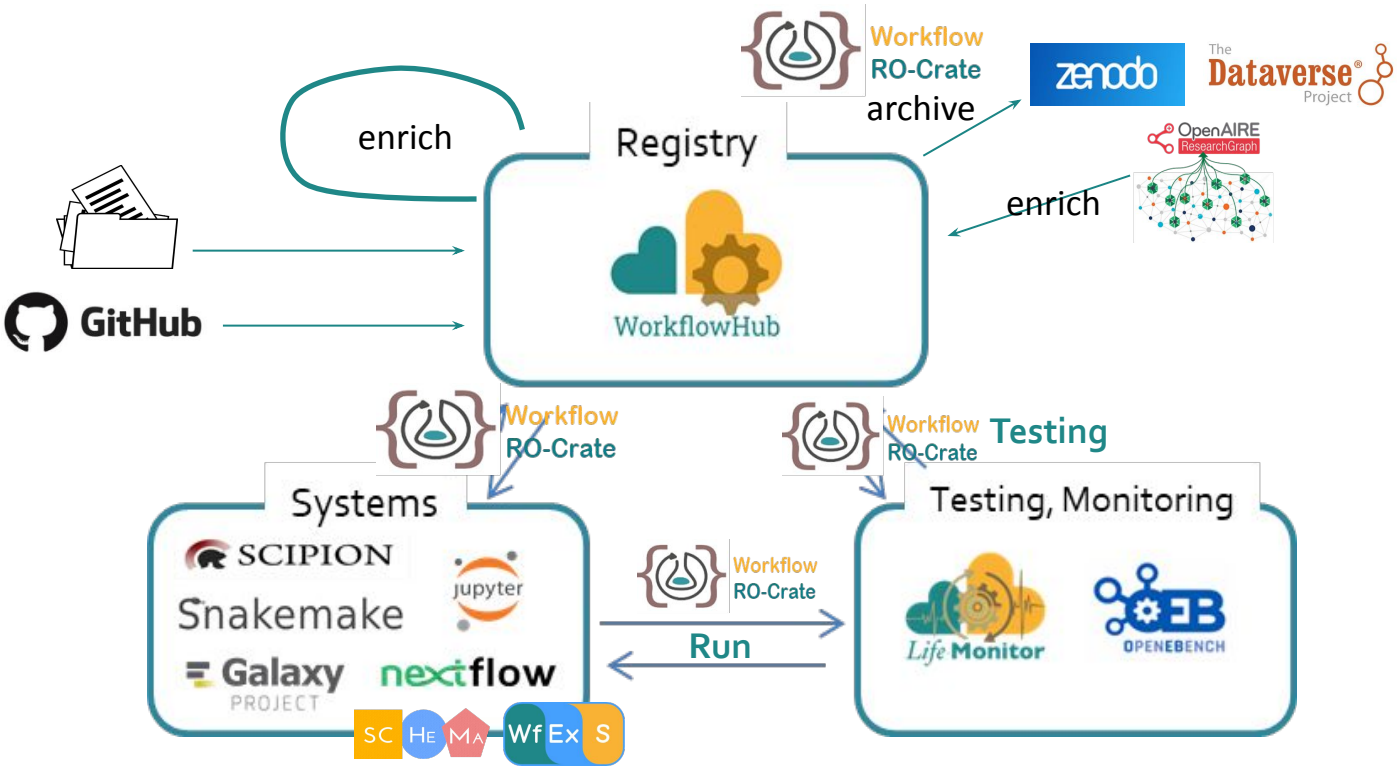
Used for
Exchange, Archiving, Reporting, Citation ...

COMMON WORKFLOW LANGUAGE

Bioschemas

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



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

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

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

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

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

Living artefacts with limited lifespans
versioned, forked, cloned
recycled, repurposed, remixed

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

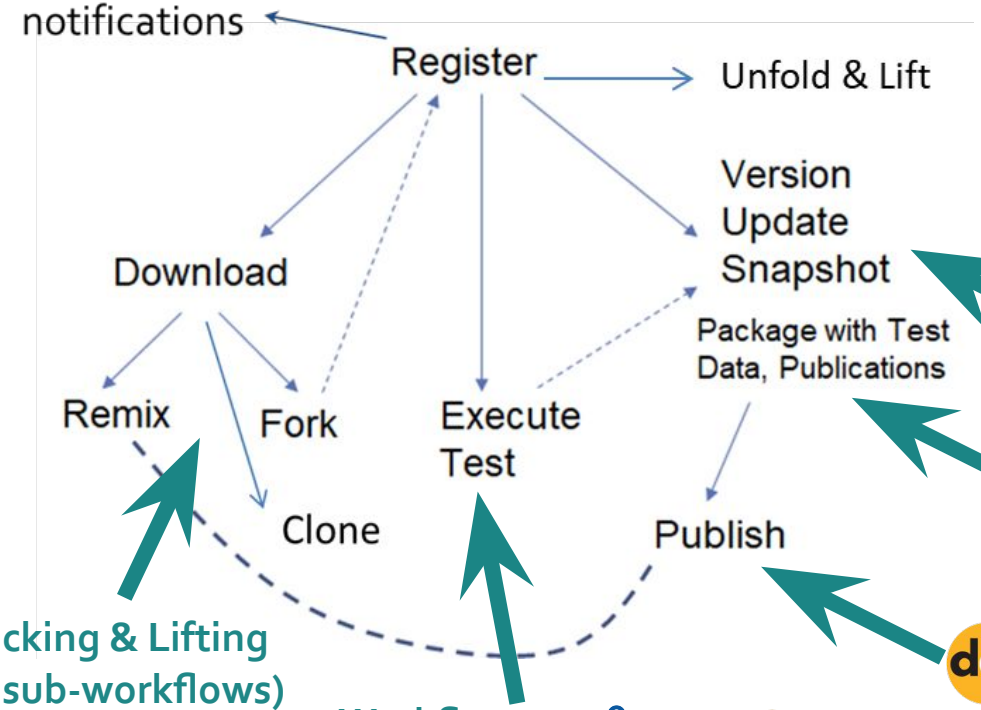
FAIR and Share which forms and parts?
specification with test or exemplar data

implementation in a WfMS → instantiation with input data, parameters set, containers deployed, computation ready

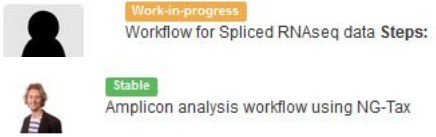
run result, intermediate/final data products, provenance logs?

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

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



Indicators of Status



Register versions
(Support Github actions)

Incremental metadata and
supplementary materials

(Tracking & Lifting
out sub-workflows)

Workflow
monitoring

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

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

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 GitH… …ys
- **RO-Crate archive**
- DOI metadata
- GA4GH TRS API and WFH…

Open… rotocols

- … enAPI, … RS API

Optiona… …ation and **authorisation**

- Most work… …blic
- Rich sharing/v… …bility permissions
- Restricted access possible via WFHub API

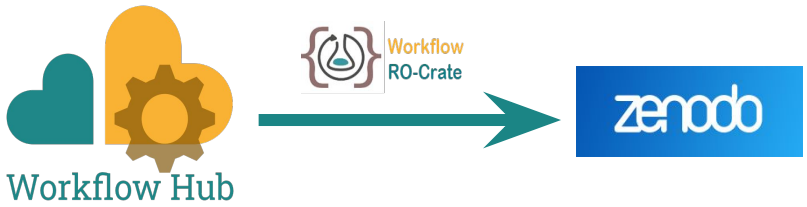
A2. **metadata** are accessible, even when the workflow is no longer available

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



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

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

Workflow RO-Crate

COMMON WORKFLOW LANGUAGE

W3C PROV

EDAM

JSON-LD

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

Galaxy PROJECT

GitHub

creative commons

elixir bio.tools

ROR

doi

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

Life Monitor

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

OPENEBENCH

Check workflow performance Provenance on containers, memory usage etc

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

https://doi.org/10.1162/dint_a_00033

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

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

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

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

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 Droesbeke (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 Andrabi
- 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 (EMBRC)
- 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 (CIRMMP 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 (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@esciencelab.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