



T3.2 Integration of PID practices into FAIR data management T3.2.1 PIDs in workflows

UNIMAN: Carole Goble, Stian Soiland-Reyes, Nick Juty 2022-10-06



This work is licensed under a <u>Creative Commons Attribution 4.0 International License</u>. Cite as: <u>https://doi.org/10.5281/zenodo.7157647</u>



FAIR Workflow registry

Workflow-system agnostic

Search for and discover workflows

Metadata standardization

(CWL, schema.org, custom tags, RO-Crate)

DOI publication, citation & credit

Collections

Teams, Organizations and Communities

Programmatic access: GA4GH TRS API, RO-Crate

Registry, not repository

Workflows can live elsewhere, e.g. GitHub

Integration with execution platforms (incl. usegalaxy.eu)

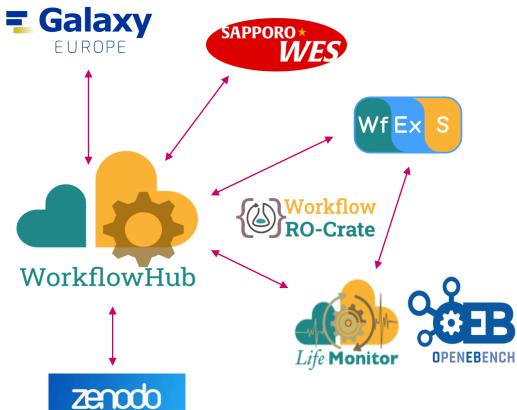
Challenges: PID before workflow registration is ready to be promoted to DataCite?

Encourage workflow citations.



MC COVID19like Ass	embly Reads	isit source 🕼 Download RO-Crate 🗮 Run on usegalaxy eu
Version 1	and the second se	
	access	For the second se
(Werthat an open as)	uilable from https://covid19.galaxyproject.org/genomics/2-assembly/ II ha	Creators and Submitter
the second s	a. Prior to be submitted to INDSC databases, these data need to be clear	
versions & status		Not specified Submitter & CTG
To facilitate the deposition of raw sequencing reads in INDSC di different levels of stringency/filtration.	icycler and the SPAdes assemblers, similar to the orginal WV. atabases, different fastq files are saved during the different steps of the WV. Which i	
	quently, a similarity search is performed against the reference assembly of the SARS	3-CoV-2
genome, to retain only SARS-CoV-2 like reads. (3) Finally, SAR corresponding fastg files) only reads that are completely identic	ral to the text accembly of the gapping	License
Any of the fastq files produced in (1), (2) or (3) are suitable for	liconsin	Apache Software License 2.0
contain more data, including for example genomic sequences of		Parente soninario environte e lo
results. File obtained at (2) include all the reads in the sequence	ing reaction that had some degree of similarity with the reference SARS-CoV-2 geno	one, these
may include subgenomic RNAs, but also polymorphic regions/vi these data is not guarateed to produce identical and consistent	ariants in the case of a coinfection by multiple SARS-CoV-2 strains. Consequently, re I results, depending on the parameters used during the assessment	sanaysis or
more information.	analy	Views: 398 Downloads: 24
Please feel free to comment, ask questions and/or add suggest	tors analy	
Colony ELDOR-ITAL: * Colony ELDOR-ITAL: * +		Last used. 3rd Jul 2021 at 02:14
← → C © 0 # 90.147.102.233/workflow/editor*	18/08/48/115/200c04 876 - 0 Q ± 16 C	D # # E
Colory / FLIXIB-ITALY	and these descent and the last B	Calestonia O Tags
Contract of Contra	THE MAKE BELIEVE	covid-19
Name Carl Same	Provenue as	
United Date United Typestates (20 Date	Carling and the second	() Attributions
Ter and Termine Consult Fernine Fare and Let	Anno Lag and Allen ar origination and an origination of the second sec	
Anno and and annotation annotati	And the second s	other
Alter som at the set of the set o	Alter and a plant in the second	other
rs and Submitter	and functions, and functions of the second s	workflows
	and for a second	WORKHOWS
	and the statement of th	
Brack, 🛔 Oliver Woolland, 🛔	And a first of the second of t	
Livermore .		manhacican
r 1	Partie of a particular data a second data a se Second data a second data	Imphasis on
Woolland	Phase or operanding colors	and a second second
	Version automatic solida, Version automatic	netadata
	Phone is an associated in phone When it is an associated in phone When it is an associated in phone with the second phone is a second phone with the second phone is a second phone is a second phone with the second phone is a sec	neldudld
funciari	[Man - ray damage - p]	
🔏 Сору		
, Woolland, O., & Livermore,	1.0.	//
). De novo digitisation.	http:	s://workflowhub.eu/
). De novo digitisation. vHub. https://doi.org/10.48546	<u>http</u> :	s://workflowhub.eu/
	<u>http</u> :	s://workflowhub.eu/

EOSC-Life ecosystem



FAIR-IMPACT

The services in the Workflow Collaboratory exchange

digital objects as Workflow RO-Crates

Packaging workflow files & companion objects

Submission / download

Exchange between services & systems

Reproducibility & Testing

Citation

Challenge: PID propagation and mutability

https://doi.org/10.5281/zenodo.4605654

EOSC-*Life*

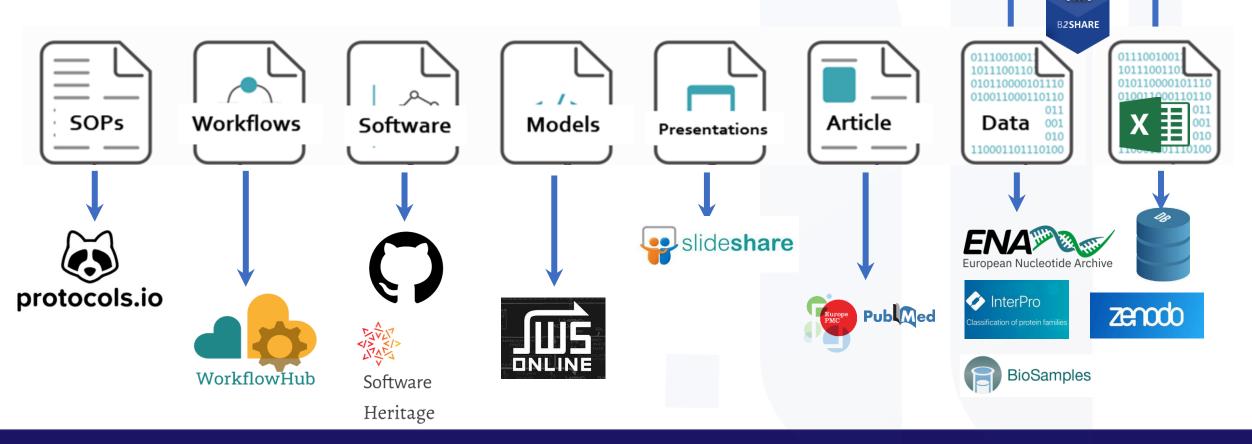


Multiple platforms and repositories

+_//...

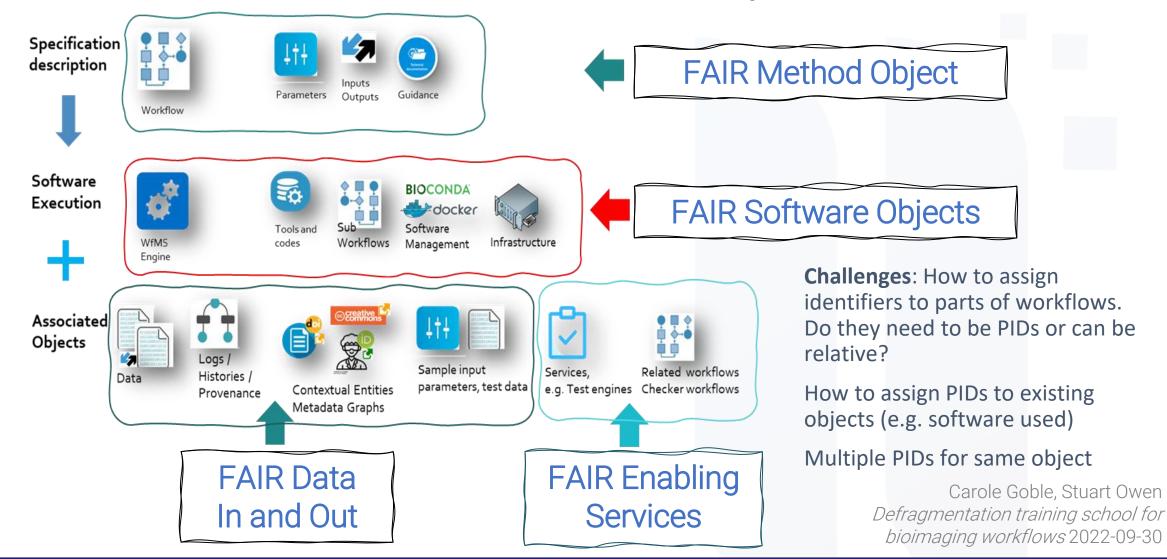
Challenge: Digital Objects are deposited in different repositories, depending on their type and domain

 \rightarrow Many PID providers, but some repositories don't assign PIDs



Workflows are composite objects

FAIR-IMPACT

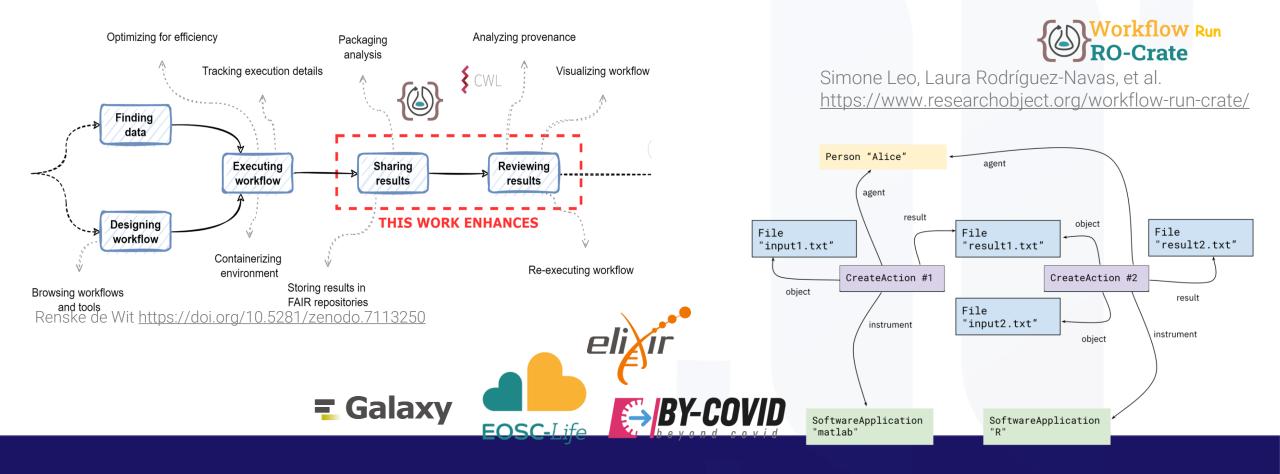




Using standards for workflow execution provenance

Challenges: Assigning PIDs (or not) to data in workflow runs.

Workflow Engine that writes provenance does not know where data will go – can it mint PIDs?





Resolving WorkflowHub PIDs to RO-Crate FDOs

FAIR Signposting – HTTP headers for PID, metadata, ++

Common types from schema.org

Multiple metadata files incl. JSON-LD, DataCite XML

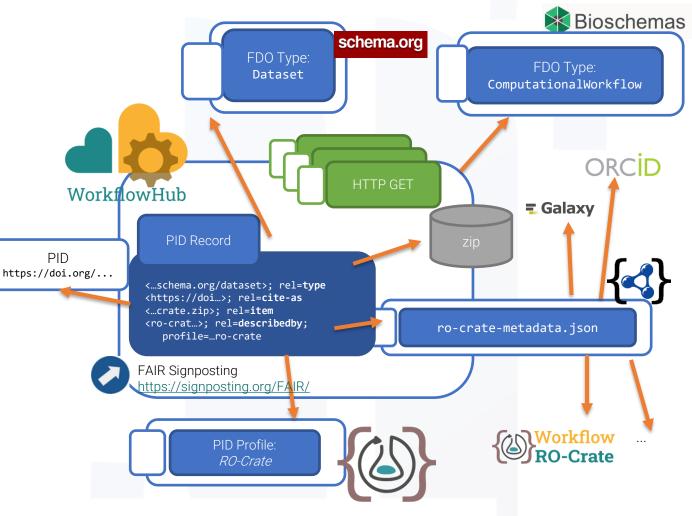
Domain-specific profiles (e.g. Workflow Crate)

Conforming to existing Community APIs (e.g. GA4GH)

Challenges: Exposing individual parts of RO-Crate while also in ZIP file, git repo etc? Multiple PIDs.

Formalizing this convention, more implementations

<u>Updating Linked Data practices for FAIR Digital Object principles</u> & <u>Creating lightweight FAIR Digital Objects with RO-Crate</u>. *Research Ideas and Outcomes* **8**:e93937 , <u>1st Intl Conf on FAIR Digital Objects</u>





T3.2.1 contributions

- Formalize/document FAIR RO-Crate PID conventions for workflows
- Capture experiences in BY-COVID on workflow provenance PIDs
- Augment RO-Crate: Retrospectively assign PIDs
- Explore decentralised PIDs:
 - <u>arcp</u> generate global identifiers for parts of a package
 - Content-based addressing including Naming Things With Hashes (RFC6920)
 - RO-Crate as glue between local and global identifiers



FAIR-IMPACT integrations

- WP3
 - Contribute practical PID experiences from FAIR workflows (T3.3, T3.4)
 - Lobby for more workflow details into DataCite graph (e.g. workflow language, software dependencies)
- WP4
 - "Stress-test" FAIR Signposting convention/tooling (T4.3, T4.4)
 - Ensure PID practices are part of RO-Crate tooling (T4.2)
- WP5
 - Contribute FAIR Signposting benchmarks (T5.3)
 - WorkflowHub as exemplar for metrics testing (T5.1, T5.3)
- WP6
 - PID + FAIR Digital Object + RO-Crate (T6.1)
 - EOSC integration of FDOs (w/ EuroScienceGateway) (T6.3)







