https://eurosciencegateway.eu/

This project was funded by the European Union's HORIZON-INFRA-2021-EOSC-01 under Grant Agreement number 101057388.

D2.1 Reproducible FAIR Digital Objects for workflows

Work Package 2

2024-08

Technical References

Project Acronym	*ESG*
Project Title	*EuroScienceGateway*
Project Coordinator	*Albert Ludwig University*
Project Duration	*September 2022 / August 2025*

Document	*D2.1*
Work Package	*WP2*
Task	*T2.2, T2.3*
*Dissemination Level**	*PU Public*
Lead Beneficiary	*UNIMAN / VIB*
Contributing Beneficiaries	*VIB, Freiburg, BSC, EPFL*
Due Date of Deliverable	*31 August 2024*
Actual Submission Date	*28 August 2024*

Funded by
the European Union

PU = Public

PP = Restricted to other programme participants (including the Commission Services)

RE = Restricted to a group specified by the consortium (including the Commission Services)

CO = Confidential, only for members of the consortium (including the Commission Services)

Version	Date	Beneficiary	Author	Approved
#1	2024-07-31	UNIMAN	Stian Soiland-Reyes	
#2	2024-08-07	UNIMAN, BSC	Stian Soiland-Reyes, Eli Chadwick, Finn Bacall, José M. Fernández	
#3	2024-08-28	UNIMAN, Freiburg	Stian Soiland-Reyes, Björn Grüning, Hakan Bayındır, Eli Chadwick, Alexander Hambley	

Acknowledgements

This project was funded by the European Union's HORIZON-INFRA-2021-EOSC-01 under Grant Agreement number 101057388.

Cite as

Stian Soiland-Reyes, Eli Chadwick, Finn Bacall, José M. Fernández (2024):
EuroScienceGateway D2.1: Reproducible FAIR Digital Objects for Workflows.
Zenodo
https://doi.org/10.5281/zenodo.13225792

**Funded by
the European Union**

Executive Summary

This EuroScienceGateway report gives an overview of FAIR Digital Objects (FDO), considering their use for computational workflows as scholarly objects. EuroScienceGateway has progressed the technologies Signposting and RO-Crate for implementing Workflow FDOs with the registry WorkflowHub and the workflow system Galaxy, and initiated work with academic publishers to encourage workflow citation practices.

Here we document how WorkflowHub supports research software best practices for workflows, and assist building FAIR Computational Workflows. Provenance of workflow executions has been made possible in an interoperable way across many workflow systems using Workflow Run Crate profiles, including from Galaxy.

Finally this report explores how Workflow FDOs are exposed and can be utilised, e.g. gathered in knowledge graphs and having tighter workflow system integration.

List of Abbreviations

- **ARC:** Annotated Research Contexts
  (not to be confused with Advanced Resource Connector)
- **BYOD:** Bring Your Own Data
- **CI:** continuous integration
- **EOSC**: European Open Science Cloud
- **FAIR**: Findability, Accessibility, Interoperability, and Reusability
- **FDO**: FAIR Digital Object
- **FDO-D:** FDO Data requirements
- **IWC**: Galaxy's Intergalactic Workflow Commission
- **PID**: Persistent Identifier
- **RO**: Research Object
- **TRE**: Trusted Research Environment
- **WfMS**: Workflow Management System
- **WRROC**: Workflow Run RO-Crate

Funded by
the European Union

Table of contents

Reproducible and reusable FAIR Digital Objects

Evolving the FAIR Digital Objects concept

The concept *FAIR Digital Objects* (FDO) has been proposed with a set of principles and recommendations for implementing machine-actionable scholarly outputs with predefined types, attributes and methods [Anders 2023]. The aim is to build ecosystems of structured data and detailed operations that can be predictably combined in an interoperable way, enhancing the FAIR principles beyond static data publishing.

Development of FDO is governed by the FAIR Digital Object Forum, through working groups, and from 2023 through a Technical Advisory Committee, both of which ESG partner UNIMAN are participating in.

In addition, FDO is the main topic of the Research Data Alliance (RDA)'s FAIR Digital Object Fabric interest group; and from 2024 the EOSC Task force FAIR Metrics and Digital Objects Task Force, which include members from UNIMAN, BSC, CESNET. EOSC has highlighted FDO as part of its updated Interoperability Framework [Nyberg Åkerström 2024], along with the need for semantic mappings.

EuroScienceGateway and ELIXIR Europe participated strongly in the EOSC Winter School 2024 [Erxleben 2024], across the Opportunity Areas for *PIDs*, *Metadata, Ontologies & Interoperability*, *FAIR Assessment & Alignment*, *User & Resource Environments, Skills, Training, Rewards, Recognition, & Upscaling* and *Open Scholarly Communication*. Work in EOSC Opportunity Areas continue in parallel with the task forces, with a wider participation mechanism.

FDO specifications

A series of specification documents [FDO-Specs] detail the principles of FDO and its different components[1] such as identifiers, attributes and operations.

In 2024, the FDO Forum drafted a simplified set of FDO Data requirements (**FDO-D**) [Strawn 2024], based on the existing specifications, focusing on the main principles for data accessibility:

1. **Data FAIR Digital Objects** (FDO-D) are machine actionable units of information bundling all information that is needed to enable FAIR processing of any included bit-sequence.
2. A **PID**, standing for a globally unique, persistent and resolvable identifier, is assumed to be at the basis for FDOs.
3. A PID resolves to a structured **FDO-Record** compliant with a specified **FDO-Profile** which leads to predictive resolution results.

---

[1] For a summary of FDO specifications, see https://peerj.com/articles/cs-1781/#an-overview-of-upcoming-fdo-specifications

Funded by the European Union

4. The FDO-Record needs to contain **Mandatory FDO (kernel) Attributes**, may contain **Optional FDO attributes** and attributes agreed upon and defined by recognized communities.
5. **Mandatory-FDO-D Attributes** are: (1) the **FDO-Content-Type**, (2) the reference to the **FDO-Profile**, (3) the reference to the **bit-sequence**(s) encoding data, (4) the references to the different **metadata** resources.

Reproduced from [Strawn 2024]

An overview of FDO-D is shown in Figure 1.

In addition, a new FDO task force is establishing the *Machine Rules for accessing FDOs* (TSIG-TF 02), where UNIMAN is contributing to specify the algorithmic approach for consistent access to the FDO-D concepts across implementations, based on our practical experiences in the EuroScienceGateway project.

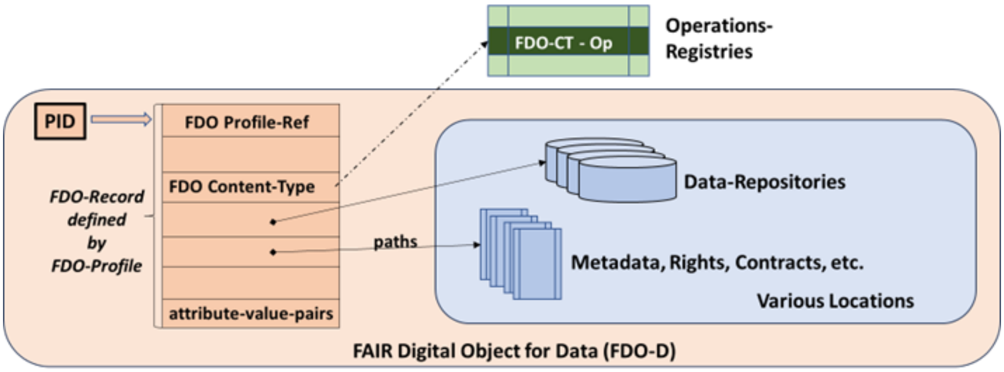


Figure 1: FAIR Digital Object for Data (FDO-D), where a persistent identifier (PID) resolves to an FDO Record, which structure is determined by the identified FDO Profile. The record references retrievable data from repositories, and separate metadata resources. Additional attribute/value pairs include the content type, which combined with operations registries enable additional operations on the data and the FDO. Reproduced from [Strawn 2024].

Evaluating FDO and Linked Data

There is a potentially large overlap across the FDO concept and established Linked Data practices, but FDO is technology-neutral in terms of implementations and protocols, with multiple realisations that can all be said to be following FDO principles at least loosely [Wittenburg 2022].

As part of EuroScienceGateway and with other EOSC-related projects, UNIMAN performed an evaluation of FAIR Digital Object and Linked Data, considering them as distributed object systems against multiple frameworks [Soiland-Reyes 2024a]. This extensive evaluation concluded that Linked Data technologies are not yet approachable for developers and further agreement on predictable implementations are needed, and as well as that FDO can learn from the earlier Semantic Web approaches to strike a balance between flexibility and rigidity.

The aforementioned evaluation article has been positively received by the FDO Forum and spurred several discussions, and forced a move to formalise the many "flavours" as *FDO Variants*. A new report is now being drafted by the FDO Technical Specification & Implementation Group (TSIG) that will list and compare 10 established use cases and FDO practices [Broeder 2024]. A part of this work is to formalise how the FDO-D requirements are implemented for each.

Signposting and RO-Crate

Packaging FAIR data with RO-Crate

Research Object Crate (**RO-Crate**) is a method for packaging of research data with structured metadata, building on established Web standards and supporting the FAIR principles for data sharing (Findable, Accessible, Interoperable, Reusable) [Soiland-Reyes 2022b]. The idea of the RO-Crate is to be a self-contained description of the Research Object with sufficient context for a human to be able to understand and reuse the data.

As RO-Crate is built on Web standards like JSON-LD it is easy to integrate the crate metadata with FAIR supporting systems, for instance building a knowledge graph across multiple crates combined with other FAIR resources enable complex queries using the SPARQL language and transformations to other metadata standards. The Linked Data background also gives clear mechanisms for extension vocabularies, although RO-Crate's default vocabulary schema.org does most of the heavy lifting and is compatible with search engine indexes like Google Dataset Search.

In EuroScienceGateway we have primarily used RO-Crate in these aspects, which are detailed in the rest of this deliverable:

- As archival and submission format for the https://workflowhub.eu/ workflow registry
- As provenance export of a workflow run, including from Galaxy
- As import and export format of a Galaxy history and its data, e.g. for depositing to the Zenodo repository.

Using Signposting for FAIR Digital Objects

Signposting is a way to give machines "just enough" navigation elements and metadata using existing HTTP mechanisms on the Web [Van de Sompel 2015]. Notably a fixed set of *link relations* are used to provide typed references from a HTML *landing page* to persistent identifier, downloadable resources and metadata (Figure 2).

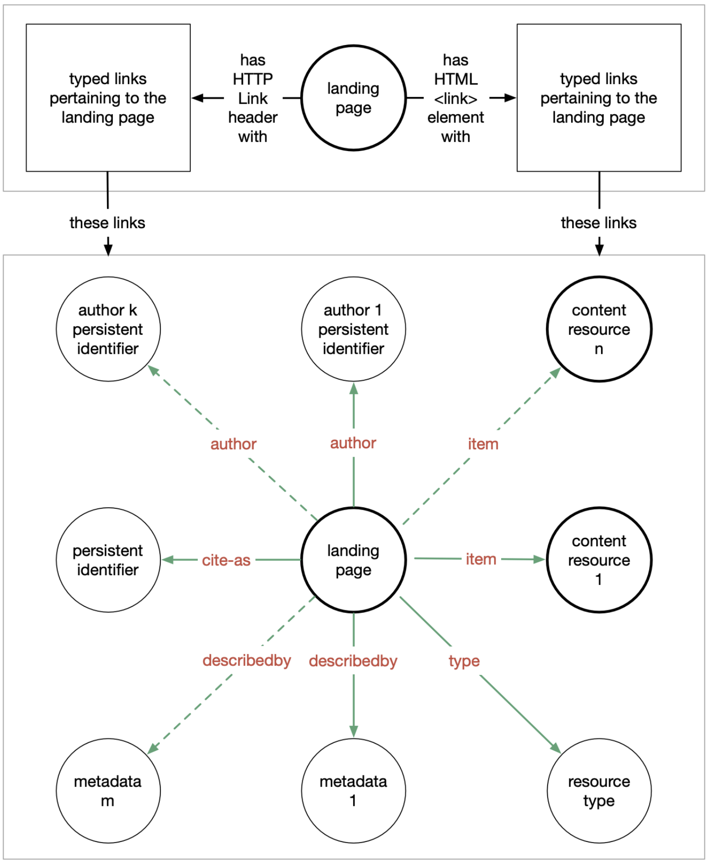


Figure 2: FAIR Signposting level 1 link relations. Reproduced from https://signposting.org/FAIR/

Taken together, the FAIR Signposting Profile [Van de Sompel 2023] has been specified as a community effort to identify the minimum of link relations needed to support the FAIR principles.

Simplified, we can say that FAIR consumption of a digital object involves:

1. Resolve persistent identifier, following any redirects
2. Find and retrieve data download
3. Find type and metadata for resource, and its expected format & profile
4. Parse metadata (e.g. into knowledge graph), query according to profile

Recent effort in the EOSC (including the FAIR-IMPACT project) has supported a growing uptake of Signposting by repositories, in particular to simplify FAIR consumption and to improve FAIR metrics [Wilkinson 2024a], as it was previously not very consistent how a client should do the FAIR resolution, causing differences in heuristics (particularly. in step 2 and 3 above) and thus measuring different metrics [Wilkinson 2022a]. Signposting helps by making the identification of constituent resources of a digital object explicit and consistent.

In EuroScienceGateway, we have implemented Signposting for exposing WorkflowHub entries as FAIR Digital Objects [Soiland-Reyes 2022a] and extended Signposting support for:

1. Retrieving JSON-LD using schema.org as a DCAT-like DataCatalog from https://workflowhub.eu/ following the Bioschemas profile for catalogues – this links to a Dataset for each of:
   Collections (e.g. Workflows in EuroScienceGateway [Soiland-Reyes 2024e]), Uploaded data files, Documents, Events, Institutions, Organisms (including SARS-CoV-2), People, Presentations, Spaces[2],Teams, Workflows
   For each grouping, a Dataset description (Figure 3) link to a complete dump of the corresponding entries as JSON-LD/Bioschemas. For instance, https://workflowhub.eu/workflows.jsonld?dump=true describes all the workflows.
2. JSON-LD metadata for each individual entry, following BioSchemas profiles
3. For each workflow, XML of metadata in Datacite Metadata Schema 4.4 [Datacite 2021]
4. Retrieving RO-Crate from WorkflowHub entries, e.g. from https://workflowhub.eu/workflows/29?version=3 to the crate download https://workflowhub.eu/workflows/29/ro_crate?version=3
5. Retrieving the persistent identifier for a workflow entry which has an assigned DOI

```
{
  "@context": "https://schema.org",
  "@id": "https://workflowhub.eu/workflows",
  "type": "Dataset",
  "dct:conformsTo": "https://bioschemas.org/profiles/Dataset/0.3-RELEASE-2019_06_14/",
  "creator": {
        "id": "https://about.workflowhub.eu/",
        "type": "Organization",
        "name": "WorkflowHub",
        "url": "https://about.workflowhub.eu/"
  },
  "description": "Workflows in WorkflowHub.",
  "distribution": {
        "type": "DataDownload",
        "contentSize": "3.8 MB",
        "contentUrl": "https://workflowhub.eu/workflows.jsonld?dump=true",
        "dateModified": "2024-06-11T00:11:09+01:00",
        "description": "A collection of public Workflows in WorkflowHub, serialized as an
array of JSON-LD objects conforming to Bioschemas profiles.",
        "encodingFormat": "application/ld+json",
        "name": "workflows-bioschemas-dump.jsonld"
  },
  "includedInDataCatalog": {
        "id": "https://workflowhub.eu"
  },
  "keywords": [],
  "license": "https://spdx.org/licenses/CC-BY-4.0",
  "name": "Workflows",
  "url": "https://workflowhub.eu/workflows"
}
```

Figure 3: Example Dataset description in JSON-LD, reformatted for readability from https://workflowhub.eu/workflows

---

[2] See https://about.workflowhub.eu/docs/guide-to-using-workflowhub/ for a guide to WorkflowHub's grouping of *Space*, *Team* and *Collection*. The linked JSON-LD dump is generated periodically..

For workflows, the biggest difference from Bioschemas markup and RO-Crate's metadata is that each RO-Crate also contain the workflow definition files (e.g. snapshotted from GitHub). The crate's description of the workflow itself will be equivalent to the BioSchema in the case of the crate being auto-created by WorkflowHub at definition file upload, but may contain extra annotations if registered as an RO-Crate directly (see section Encouraging research software best practices for workflows).

The Signposting support has been verified with the Python Signposting tool [Soiland-Reyes 2024d], see Figure 4.

```
$ curl -sI https://workflowhub.eu/workflows/415 | grep -i ^link
link: <https://workflowhub.eu/workflows/415?version=1> ; rel="describedby" ;
type="application/vnd.datacite.datacite+xml", <https://workflowhub.eu/workflows/415?version=1>
; rel="describedby" ; type="application/ld+json",
<https://doi.org/10.48546/workflowhub.workflow.415.1> ; rel="cite-as",
<https://workflowhub.eu/workflows/415/ro_crate?version=1> ; rel="item" ;
type="application/zip" ; profile="https://w3id.org/ro/crate"

$ signposting https://workflowhub.eu/workflows/415
Signposting for https://workflowhub.eu/workflows/415
CiteAs: <https://doi.org/10.48546/workflowhub.workflow.415.1>
DescribedBy: <https://workflowhub.eu/workflows/415?version=1>
                       application/vnd.datacite.datacite+xml
             <https://workflowhub.eu/workflows/415?version=1>
                       application/ld+json
Item: <https://workflowhub.eu/workflows/415/ro_crate?version=1> application/zip
```

Figure 4: Signposting for https://workflowhub.eu/workflows/415 explored from the HTTP Link header using curl, and as parsed by the Python signposting tool. The persistent identifier (PID) is indicated as `rel=cite-as`. The metadata linked to from `rel=describedby` makes explicit the ability to use HTTP Content Negotiation to retrieve metadata in either Datacite or JSON-LD formats. The ZIP download (`rel=item`) is likewise typed with a `profile` to indicate it is an RO-Crate.

Further work that has been identified as within scope for the remaining period of EuroScienceGateway include:

1. Automate FAIR metrics checking of WorkflowHub resources with Signposting
2. More specific Signposting and FAIR metadata to find individual WorkflowHub entries, e.g. `rel=item` from a WorkflowHub collection to the contained workflows
3. Additional Signposting on WorkflowHub extracted from metadata, e.g. `rel=type`, `rel=author`, `rel=license`
4. Signposting from Galaxy public workflow landing pages (e.g. https://usegalaxy.eu/published/workflow?id=466bdd8ba7b67264) to download (https://usegalaxy.eu/api/workflows/466bdd8ba7b67264/download?format=json)
5. Signposting from Galaxy public history landing pages to their RO-Crate export

Wider collaboration at the ELIXIR Biohackathon [Soiland-Reyes 2024c] helped demonstrate and further develop EuroScienceGateway's FDO approach.

FDO profile using Signposting and RO-Crate

The FDO Data requirements (**FDO-D**) [Strawn 2024] can be implemented using Signposting, RO-Crate, or as explored by EuroScienceGateway, their combination. Table 1 shows the profile we have developed to formalise these implementations.

FDO-D requirement	**Signposting implementation**	**RO-Crate implentation**
PID	HTTP redirect, `rel=cite-as`	HTTP redirect and/or signposting, `identifier`
FDO-Record	Signposting in HTTP header	RO-Crate metadata document, resolved using signposting or content negotiation from PID.
FDO-Profile	`rel=profile` and `profile="http://example.com/pid/1"` on `rel=describedby` and `rel=item`	`conformsTo` on data entity, `conformsTo` on crate. Defined as a Profile Crate as its own FDO.
Mandatory-FDO-D attributes	`rel=describedby, rel=item, rel=type`	Required properties: `name`, `license`, `description`, `datePublished`
Optional attributes	`rel=license`, `rel=author` Extensions by URI (see 111-fdo-gr4-attribute-uris/)	Multiple contextual attributes & types, extensible by profiles.
FDO Content-Type	`rel="type"` (semantic type), IANA media type as `type="text/html"` on `rel=item` and `rel=describedby` (syntactic type)	`encodingFormat` on data entity, with detailed file format info
Bitsequence reference	`rel=item` to download	Data entity, including web-based and directory archives
Metadata reference	`rel=describedby` with `type=` and `profile=`	Additional metadata resources linked using `subjectOf` and file format profile.

Table 1: Fulfilling FDO-FD requirements using Signposting and RO-Crate. Adapted from https://s11.no/2024/webby-fdos/#tab:relations

Further work is undergoing within the FDO TSIG working group to document all "FDO flavours" similarly [Broeder 2024], where EuroScienceGateway is responsible for documenting the Signposting and RO-Crate approaches.

Training and outreach

As FAIR Digital Objects implemented with RO-Crate is an emerging solution receiving broad interest, we have also developed training material, initially for the Galaxy Smörgåsbord and ELIXIR communities, but since expanded into full tutorials at Open Science and FAIR venues:

- Galaxy Smörgåsbord 2023, virtual, 2023-05-23/–26
    - Module: FAIR data and provenance with RO-Crate and Galaxy
- ELIXIR All Hands 2023, Dublin+virtual, 2023-06-05/–08
    - Workshop: Building lightweight FAIR data packages with Bioschemas and RO-Crate 2023-06-06
- Open Science Festival 2023, 2023-07-04/–05, Cologne, Germany.
    - Workshop: Data Exchange with RO-Crate and Knowledge Graphs 2023-07-05
- 15th International SWAT4HCLS Conference (Semantic Web Applications and Tools for Health Care and Life Sciences), 2024-02-26/–29 Leiden, Netherlands
    - Tutorial: Improving FAIRability of your research outcomes with RO-Crates, SignPosting and Bioschemas
- International FAIR Digital Objects Implementation Summit (FDOF2024), 2024-03-20/--21, Berlin, Germany
    - Training: Practical web-based FDOs with RO-Crate and FAIR Signposting

These practical tutorials include a template GitHub repository that is then modified to be published with Signposting and FAIR metadata using GitHub Pages:

- Signposting tutorial: https://github.com/stain/signposting-tutorial

Workflows as scholarly objects

In EuroScienceGateway WP2 have considered primarily one type of FAIR Digital Objects, where computational workflows become scholarly objects.

Using RO-Crate for workflows

Building on early work on WorkflowHub, in EuroScienceGateway we have expanded its support for generating and consuming **Workflow RO-Crate** as a package of the workflow and its supporting resources. Workflow RO-Crate is a profile of RO-Crate for describing the workflow and its metadata based on the Bioschemas ComputationalWorkflow profile [Bacall 2022] with additional definitions such as constants for known workflow systems and licences.

In WorkflowHub, workflows uploaded as deposits are wrapped into Workflow RO-Crates, storing the metadata filled in by the registering user. This means the metadata can travel with the workflow definitions as they are downloaded or retrieved.

Encouraging research software best practices for workflows

Development of mature workflows is increasingly treated like development of any research software, by following best practices, e.g. using source control repositories like GitHub or GitLab, and accompanying continuous integration (CI) testing such as Jenkins CI or GitHub Actions.

For instance, Nextflow's mature nf-core pipelines are maintained by its community through individual repositories under the https://github.com/nf-core/ organisation, while Galaxy's *Intergalactic Workflow Commission* (IWC) has mature workflows listed in the single repository https://github.com/galaxyproject/iwc and maintained in individual git repositories under https://github.com/iwc-workflows.

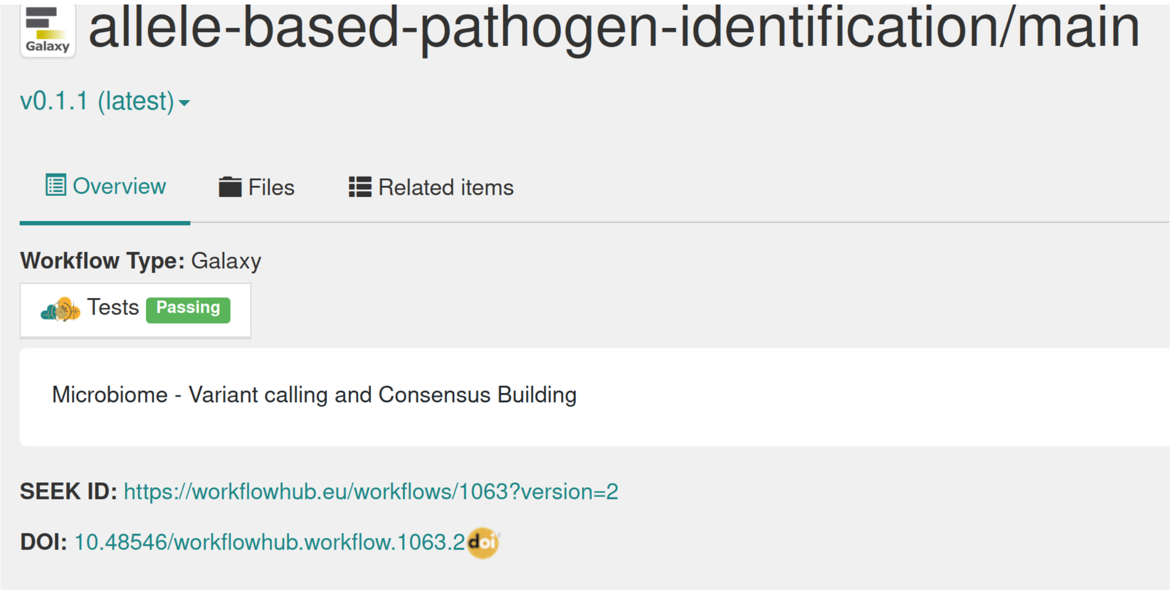
For both communities, this way of maintaining workflows enables mature software development techniques such as pull requests, automatic testing and deployments (e.g. to https://usegalaxy.eu/workflows/list_published).

As part of EuroScienceGateway, to support and encourage this way of creating workflow scholarly objects, we have expanded WorkflowHub's method of importing workflows from git repositories. By default this works similar to upload in that the user has to manually select the workflow file and workflow diagram as well as provide textual descriptions. However, if the repository includes an `ro-crate-metadata.json` file, it means it is an RO-Crate, which will then be parsed by WorkflowHub to extract this metadata. This functionality is now also available via WorkflowHub APIs.

This means metadata can be maintained upstream by workflow authors and the community, and updated along with workflow changes. This method has now been adopted throughout IWC and is used by the WorkflowHub Bot which propagates tagged GitHub releases to update the corresponding Workflowhub entry [Soiland-Reyes 2024e]. For instance, GitHub repository iwc-workflows/allele-based-pathogen-identification has a ro-crate-metadata.json that provides the metadata for Workflowhub entry [Nasr 2024] (with some caveats to be ironed out in its generation, such as formatting of ORCID identifiers). In EuroScienceGateway we have now engaged with the nf-core community to expand our support importing their workflows, there the nf-core command line is generating the RO-Crate by converting from nf-core metadata files.

Continuous Integration and Testing for workflows

With the EOSC-Life and BY-COVID project we have also integrated further with the LifeMonitor service, which expects the Workflow Testing RO-Crate profile, a specialisation of Workflow RO-Crate that defines test scaffolding in the Git repository and CI services. We have expanded WorkflowHub to look up corresponding LifeMonitor test status if the crate is following this profile, which means the two services are integrated to indicate workflow stability, as shown in Figure 5 and Figure 6.



allele-based-pathogen-identification/main

v0.1.1 (latest)

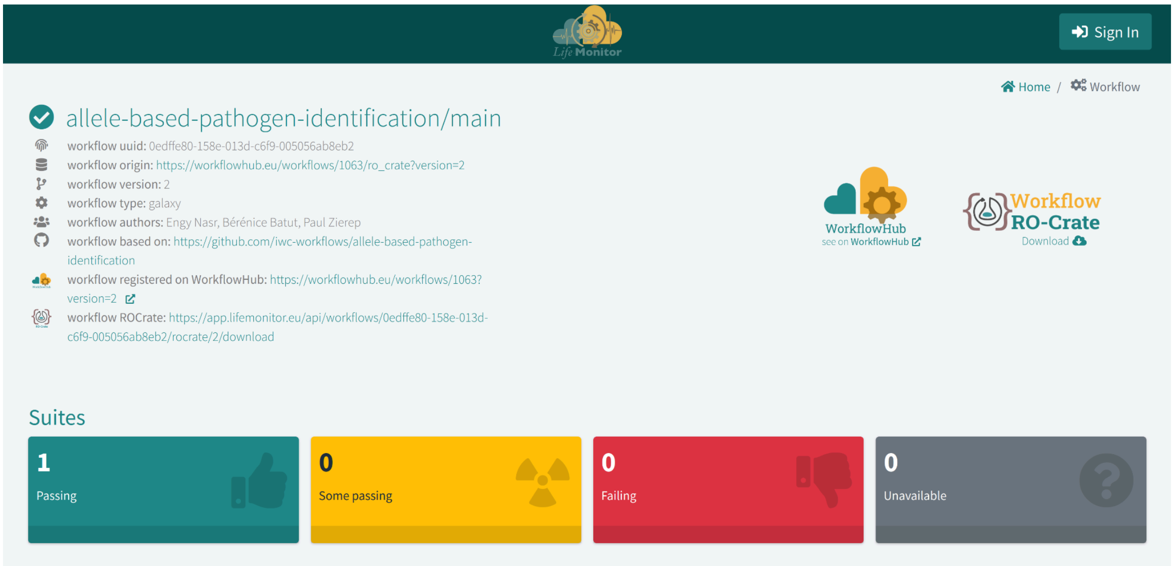
Workflow Type: Galaxy

Microbiome - Variant calling and Consensus Building

SEEK ID: https://workflowhub.eu/workflows/1063?version=2

DOI: 10.48546/workflowhub.workflow.1063.2

Figure 5: LifeMonitor test indicates the current IWC testing status of workflows [[Nasr 2024](#)] as inspected by LifeMonitor, shown in Figure 6.



allele-based-pathogen-identification/main

workflow uuid: 0edffe80-158e-013d-c6f9-005056ab8eb2
workflow origin: https://workflowhub.eu/workflows/1063/ro_crate?version=2
workflow version: 2
workflow type: galaxy
workflow authors: Engy Nasr, Bérénice Batut, Paul Zierep
workflow based on: https://github.com/iwc-workflows/allele-based-pathogen-identification
workflow registered on WorkflowHub: https://workflowhub.eu/workflows/1063?version=2
workflow ROCrate: https://app.lifemonitor.eu/api/workflows/0edffe80-158e-013d-c6f9-005056ab8eb2/rocrate/2/download

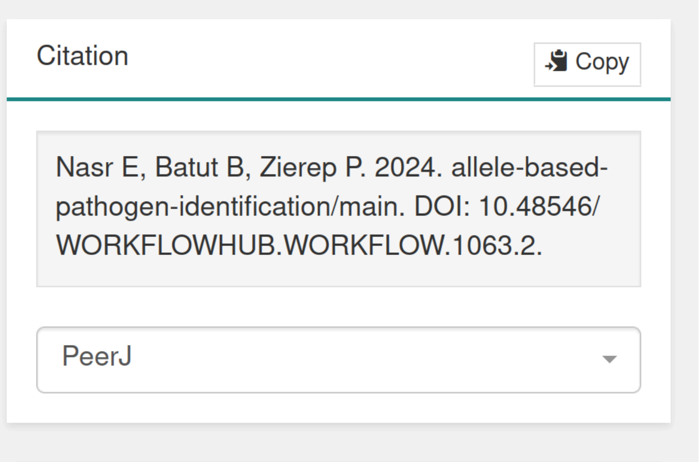
Suites

1	0	0	0
Passing	Some passing	Failing	Unavailable

Figure 6: LifeMonitor tests for [[Nasr 2024](#)], [inspected](#) from the [GitHub Action](#) executions, as indicated by the Workflow Testing RO-Crate metadata.

Encouraging workflow in publishing practices

WorkflowHub encourages [minting of DOIs](#) to help make public workflows citable, in which case we include a *Citation* box for the workflow, as shown in Figure 7. Users are encouraged to double-check the metadata before freezing to get a DOI, as this is a persistent identifier (PID) for that particular workflow version, which can't be changed after being minted. As highlighted earlier, such PIDs are also provided as Signposting by WorkflowHub for programmatic agents.



Nasr E, Batut B, Zierep P. 2024. allele-based-pathogen-identification/main. DOI: 10.48546/WORKFLOWHUB.WORKFLOW.1063.2.

Figure 7: Citation box for workflow [Nasr 2024] allowing selection of citation style. The DOI is registered with DataCite where bibliographic information has been propagated from WorkflowHub

The publisher GigaScience Press has recently raised awareness of FAIR Computational Workflows [Goble 2022] and encouraged use of WorkflowHub as part of their policies [Edmunds 2024]. GigaScience highlights the publication [Niehues 2024] by the Netherlands X-omics Initiative, which utilised RO-Crate and WorkflowHub to fully describe their Nextflow workflow [de Visser 2024] as a FAIR Digital Object. Here the RO-Crate metadata file is included in the upstream GitHub repository, where it has been generated by a Jupyter Notebook in order to add detailed annotations including Docker containers, ISA (Investigation, Study, Assay) structure and EDAM ontology annotations. The structure of the workflow scripts/steps is also listed.

As part of EuroScienceGateway we have also initiated a **Workflow Publisher Forum**, which in its inaugural meeting had representatives from several major publishers in the life sciences including representatives from Elsevier, GigaScience, PLoS, and Taylor & Francis [Goble 2024]. Several publishers in the forum were supportive of the idea of recommending registries like WorkflowHub as part of their author guidelines, but without making this mandatory. For our suggestion of improving workflow citation practices, this was received well, but the publishers flagged that it remains a challenge to get authors to add data citations and software citations in general.

A worry from several publishers was that a *proliferation of PIDs* may actually make citation tracking harder, and can decrease consistency in software citations. For instance, a computational workflow used in an analysis may have associated:

1. WorkflowHub entry, which the authors may have cited by versioned DOI or as direct workflowhub.eu URL.
   Unfortunately observed practice is commonly the latter using footnotes (or in Availability statements), rather than a formal citation under References. Some publishers still have outdated author guidelines that only recognise peer-reviewed scholarly articles.

2. Web page in a workflow-specific repository, e.g. https://nf-co.re/rnaseq/3.14.0/ with textual description.
3. Direct GitHub/GitHub source code repository URL
4. DOI of Zenodo entry, auto-generated from GitHub
5. SoftWare Heritage persistent IDentifiers (SWHIDs) with versioned commit (not yet commonly used)

From a business perspective, journal publishers are of course interested in increasing citations to their publications, and are also encouraging publications about Research Software artefacts (e.g. Application Notes), but citing data and software through registries like WorkflowHub can disincentive traditional article-to-article citations, which reduces the perceived "impact" in established publisher/journal/article metrics calculations, when in reality the impact of a journal's author guidelines can be seen as being increased if authors followed its recommended workflow & software citation practices.

It was raised as a bigger concern by publishers that authors using workflows may not have the right guidelines and practical knowledge for robust workflow design [Möller 2017] and how to follow best practices to ensure reproducibility, interoperability and long term sustainability of the workflow. Indeed the concern of *workflow decay* was raised more than a decade ago [Hettne 2012].

In EuroScienceGateway we see here a bridge to ongoing work with the Workflows Community Initiative (see next section) to fully define FAIR principles for workflows. Gathering of existing best practices for different workflow systems (e.g. for nf-core, IWC, CWL, Snakemake) and distilling these to general workflow best practices will be an important next step. Future meetings with the Workflow Publisher Forum are planned to be organised by EuroScienceGateway, in order to define common goals across publishers and to agree on such recommendations.

FAIR computational workflows and Workflows Community Initiative

The idea of considering computational workflows as FAIR objects in their own right was established as *FAIR Computational Workflows* [Goble 2022]. There are two aspects of this: Firstly, a workflow definition is a specialisation of FAIR Research Software [Lamprecht 2020] and so the workflow should be treated as a citable scholarly output (see previous section); secondly, a workflow can be an important consumer and producer of FAIR data, and with the help of the workflow engine, should assist users in capturing and propagating the associated metadata.

In EuroScienceGateway we have engaged with the Workflows Community Initiative, which has spun out of previous Workflows Community Summits [Ferreira da Silva 2023]. In the task group we are formalising the FAIR Computational Workflow principles based on best practices in several Workflow Management Systems (WfMS) (including on HPC) and the FAIR Research Software guidelines. The current draft of the principles [Wilkinson 2024b] is listed in Table 2.

PRINCIPLE
F1. A workflow is assigned a globally unique and persistent identifier.
F1.1. Components of the workflow representing levels of granularity are assigned distinct identifiers.
F1.2. Different versions of the workflow are assigned distinct identifiers.
F2. A workflow and its components are described with rich metadata.
F3. Metadata clearly and explicitly include the identifier of the workflow, and workflow versions, that they describe.
F4. Metadata and workflow are registered or indexed in a searchable FAIR resource.
A1. Workflow and its components are retrievable by their identifiers using a standardised communications protocol.
A1.1. The protocol is open, free, and universally implementable.
A1.2. The protocol allows for an authentication and authorization procedure, when necessary.
A2. Metadata are accessible, even when the workflow is no longer available.
I1. Workflow abstraction and its metadata use a formal, accessible, shared, transparent, and broadly applicable language for knowledge representation.
I2. Metadata and workflow use vocabularies that follow FAIR principles.
I3. Workflow is specified in a way that allows its components to read, write, and exchange data (including intermediate), in a way that meets domain-relevant standards.
I4. Metadata (about a workflow) and workflow include qualified references to other objects and the workflow's components.
R1. Workflow is described with a plurality of accurate and relevant attributes.
R1.1. Workflow is released with a clear and accessible licence.
R1.2. Components of the workflow representing levels of granularity are given clear and accessible licences.
R1.3. Workflow is associated with detailed provenance.
R2. Workflow includes qualified references to other workflows.
R3. Workflow meets domain-relevant community standards.

Table 2: Draft of FAIR Computational Workflow principles, adapted from [Wilkinson 2024b]

In EuroScienceGateway we see these principles as important to formalise FAIR Digital Objects for workflows and find requirements to expand the existing Workflow RO-Crate profile.

Raised by this work is a very important distinction from research software in general: The concept of a *workflow component* that itself should be treated as a FAIR scholarly object. What makes workflows different from software is that they can more easily be broken down into *steps,* which help to explain the scientific computational method, but also are often using software written by someone else than the workflow authors. This necessarily complicates software citation practices [Smith 2016], as a single computational workflow may easily use 20 of such tools, and workflows themselves can become nested.

We see an example of this such annotations done manually in the previously mentioned Nextflow example [Niehues 2024] with multiple containers. Earlier work also highlighted the need for complex software citations when the workflows use building blocks that wrap underlying software [Soiland-Reyes 2022c], as is common in both Galaxy [Galaxy 2024] and Nextflow. Clearly it needs to be the role of a FAIR supporting WfMS to propagate this information, and that is part of EuroScienceGateway's effort in this work package.

In Galaxy for instance, tool citation is often available as part of its wrapper, in terms of a preferred citation (e.g. a journal paper), although not in terms of software release (e.g. Zenodo deposit from a GitHub release). The underlying GitHub repository of the software may have CodeMeta annotations [Jones 2023] that provides the full list of tool authors etc. but the source code repository is not easily located from a deployment, Conda package or Docker image.

However, currently this information is not easily available, nor propagated to the workflow definition or the corresponding RO-Crate in WorkflowHub, as it is only available on the server where the tool definitions are installed. In EuroScienceGateway we are therefore looking at ways to augment this information so it becomes part of the workflow scholarly object (see later section on knowledge graphs).

Workflow provenance helps explain workflow use

As mentioned in previous section, WfMS can be instrumental in making FAIR data. One aspect of this is to record *provenance* of workflow outputs, connecting them to the workflow execution, and ideally the particular step executions that produced them within the workflow. This then builds a chain of provenance that goes back to the origin data and parameters, which, with sufficient data citations and additional provenance, can be traced further.

Capturing workflow provenance is also an important element of ensuring reproducibility, the previously mentioned *workflow decay* can be partially addressed by having a detailed trace - the workflow may no longer be executable, but with sufficient provenance and metadata can still be explained and recreated using different tools and settings.

This has been a motivating principle for workflows since the early days [Atkinson 2017], and earlier work like CWLProv [Khan 2019] has demonstrated methods to capture workflow execution provenance from WfMS as Research Objects and using the W3C PROV standard. However these methods are not as connected to the workflow as a scholarly object, and tend to expose many execution details of the workflow engine itself, rather than explain the workflow independently. There are also many pragmatic cases where a workflow is conceptually implied, but not formally defined in a WfMS.

In EuroScienceGateway we have therefore helped lead the effort to develop Workflow Run Crate (WRROC), a set of RO-Crate profiles that capture the execution of one or more processes, which may be organised by a WfMS [Leo 2024].

The three profiles defined, as shown in Figure 8:

1. Process Run Crate [WRROC 2024a] – a computational process was executed, which consumed and produced some files. The tool may be identified by URL to its homepage or source code. Multiple processes, where one tool's output is consumed as input by another tool, indicates an *implied workflow*.
2. Workflow Run Crate [WRROC 2024b] – a computational process was executed, and it was defined by a computational workflow. The workflow definition is included in the crate and described by the Workflow RO-Crate profile (see section Using RO-Crate for workflows).
3. Provenance Run Crate [WRROC 2024c] – the execution of the workflow is detailed for each tool (as in Process Run Crate) and related to a *prospective provenance* step definition within the workflow. Further details on the workflow engine is also included.

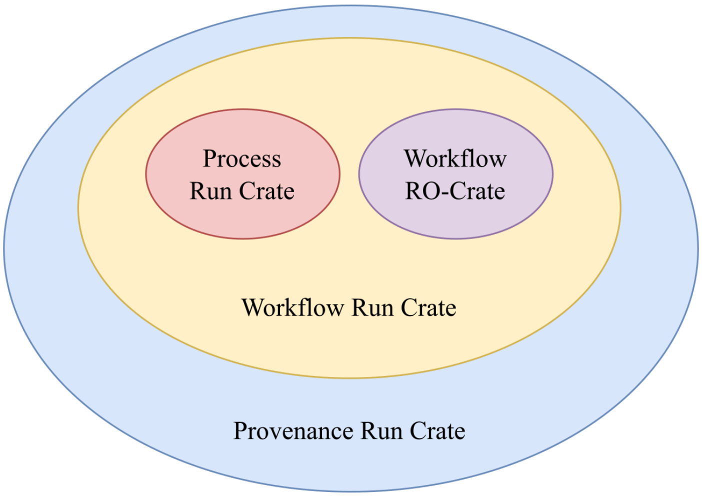
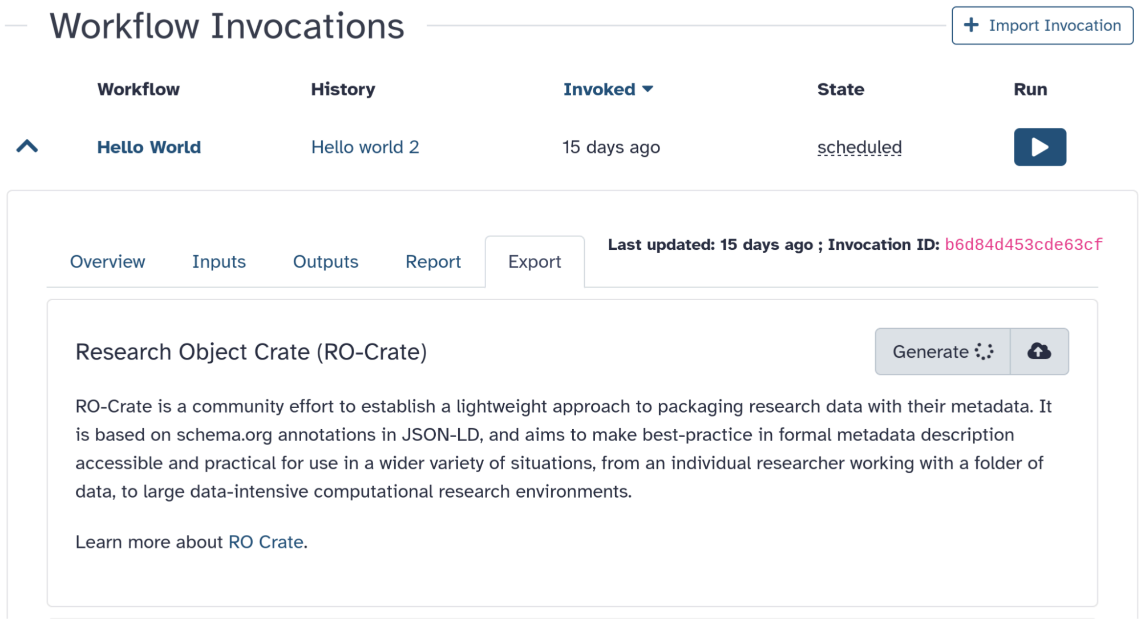


Figure 8: Venn diagram of the specifications for the various RO-Crate profiles. Workflow Run Crate inherits the specifications of both Process Run Crate and Workflow RO-Crate. Provenance Run Crate, in turn, inherits the specifications of Workflow Run Crate. Reproduced from [Leo 2024].

By having multiple profiles, different provenance detail levels are possible depending on the WfMS capabilities, as suggested by [Khan 2019]. The current implementations of WRROC, shown in Table 3, are generating such RO-Crate according to different profiles.



	Workflow	History	Invoked	State	Run
^	Hello World	Hello world 2	15 days ago	scheduled	▶

Last updated: 15 days ago ; Invocation ID: `b6d84d453cde63cf`

Research Object Crate (RO-Crate)

RO-Crate is a community effort to establish a lightweight approach to packaging research data with their metadata. It is based on schema.org annotations in JSON-LD, and aims to make best-practice in formal metadata description accessible and practical for use in a wider variety of situations, from an individual researcher working with a folder of data, to large data-intensive computational research environments.

Learn more about RO Crate.

Figure 9: Exporting a Galaxy Workflow Invocation as RO-Crate.

In addition to defining the WRROC profiles and supporting the developers of these WfMS, as part of EuroScienceGateway (building on BY-COVID work [De Geest 2022]) we have continued development of WRROC support in Galaxy, shown in Figure 9. The export can either be downloaded locally, or transferred by Galaxy to a user-defined file store such as an S3 bucket or an institutional Nextcloud/ownCloud endpoint.

WfMS	Profile	Description	Primary domain
runcrate	Provenance	Command line tool and Python library for Workflow Run Crate profiles. Can re-execute CWL workflow runs,	Generic
Galaxy	Workflow	Web-based workflow system, can export and import Workflow Run Crates.	Life sciences
COMPSs	Workflow	HPC-centric workflow system with big data	Simulation, modelling
StreamFlow	Provenance	HPC-centric container-based workflow system.	Bioinformatics
WfExS	Workflow	Workflow Execution Service, wraps existing engines, captures their provenance and rerun.	Life sciences
Sapporo	Workflow	WES execution service, wraps existing engines.	Genomics
Autosubmit	Workflow	HPC-centric workflow system focused on climate research	Climate research
Nextflow	Provenance	Script-like cloud-native workflow system, popular in genomics.	Genomics
Snakemake	(in development)	File-based workflow system with pluggable executions	Generic

Table 3: Implementations of Workflow Run Crate profiles across different workflow management systems (WfMs) and their primary science domains. Adapted from [Leo 2024]

Current implementation of WRROC in Galaxy is done as the *Workflow Run Crate* profile, meaning the RO-Crate includes:

1. The workflow definition (classical .ga format and newer.gxwf.yml format)
2. Workflow abstractions (as Abstract CWL, consumed by WorkflowHub; and a HTML diagram)
3. Workflow inputs and output values, copied from the history
4. Execution details for the overall workflow, linking these files
5. Additional internal state representations from Galaxy, e.g. collections_attrs.txt list the Galaxy data types of the data files, and invocation_attrs.txt has details of the invocation.

The additional representations allows Galaxy to recreate the workflow execution state on import, however they are not interoperable with other Workflow Run Crate implementations. Further work being investigated by EuroScienceGateway WP2 is to translate from these to lift the additional details on step execution, making a more granular *Provenance Profile Crate* as demonstrated with *runcrate*, *Streamflow* and *Nextflow*.

Additional reproducibility work on WRROC, to resolve the workflow component citation issue mentioned earlier, is to fully define capturing of software containers at the time of execution, along with provenance of how these containers were built. EuroScienceGateway work on the workflow execution service (WfExS) (with WP3) has already implemented this in terms of capturing containers, and have recently developed deep inspection of Snakemake workflows [Iborra 2024]. This shows that it is not necessary to be deeply integrated in the workflow engine, however further revision of the WRROC profiles may be needed to better support this kind of mixture of the Workflow Run and Provenance Run profiles.

Through the profile inheritance shown in Figure 8, crates following Workflow Run Crate or Provenance Run Crate will also be implementing Workflow RO-Crate and so technically be possible to deposit in WorkflowHub, of which a handful of examples already exist.

It however is not in EuroScienceGateway's vision that WorkflowHub will become a global host of WRROC workflow runs, as these will include workflow output files and potentially container images, they can become large or complex, and require different treatment as data-like rather than as software-like scholarly objects. In addition, a large majority of workflow runs will have failed in some respect or not be interesting for broader publication. Naturally, one workflow definition may have many workflow runs, and some of these may be good exemplars to help explain the workflow. This is one aspect of Workflow scholar objects we will explore further within EuroScienceGateway.

Depositing Workflow Run Crates

In general, EuroScienceGateway would encourage users to publish the workflow runs to general repositories like Zenodo, ideally providing links back to the WorkflowHub entry.

Further work on this being explored is to traverse such repositories to identify matching runs of known workflows (see knowledge graph section), or to provide a *pingback mechanism* for RO-Crate upload mechanisms like in Galaxy to notify WorkflowHub about the publishing of the related workflow run.

Zenodo uploader

In order to support general uploading of RO-Crate as FAIR Digital Objects, in EuroScienceGateway we have developed the rocrate-zenodo command line tool and Python library (https://github.com/ResearchObject/ro-crate-zenodo) [Chadwick 2024]. This tool has two functionalities:

1. Extract RO-Crate metadata and transform to Zenodo's metadata format
2. Upload the RO-Crate to the configured Zenodo instance, zipping if necessary

The tool can be configured to work against https://sandbox.zenodo.org/ for testing, and needs a Zenodo developer token for authentication. It is also configurable if the uploaded record should be immediately published, or left in draft stage for further editing in the Zenodo Web UI.

This uploader uses the "classic" official Zenodo REST API, which still remains the official API of Zenodo. The mapping includes some heuristics for selecting the open source license, as many different identifiers are used in RO-Crate. For consistent results, SPDX identifiers should be used for the license in the RO-Crate.

However, as of autumn 2023, Zenodo.org has been updated to be based on the open source InvenioRDM, which has its own API and metadata based on the Datacite Metadata schema [Datacite 2021]. InvenioRDM is also used by several institutional repositories, including by EuroScienceGateway partner Freiburg (https://freidata.uni-freiburg.de/).

For this reason, we have also contributed and released ro-crate-inveniordm [Beer 2024], a fork of the open source beerphilipp/ro-crates-deposit [Beer 2023]. This tool was enhanced from Beer & Szente's version to add automated tests, new command line options, and support for environment variables for credentials. Some minor bugs in the original tool have also been fixed.

Moving forward, we suggest using and developing further *ro-crate-inveniordm* rather than *ro-crate-zenodo*, although for now we will maintain both options pending Zenodo's decision on their official API. It should be noted that *ro-crate-inveniordm* also has a more complete and configurable mapping of authors and contributors than our initial *ro-crate-zenodo*.

Using and enriching workflow FDOs

Knowledge graph considerations

In general sense, the term *knowledge graph* refers to a collection of facts expressed through named nodes and qualified edges, which can be examined and queried in multiple ways, without having one particular top node. Using knowledge graphs have become established as a powerful method for data analysis and insight, and compared to relational databases have strengths such as flexibility, extensibility, mergeability and transformability.

In practical applications, different ways to implement knowledge graphs build on existing data structures and formats, and are typically prepared from underlying databases and other data sources for use in particular knowledge graph software. JSON-based knowledge graphs such as ElasticSearch and Neo4J can index such data and expose it with APIs such as GraphQL, but have a disadvantage that such graphs must be merged and prepared in advance for the intended set of queries and integrations, by *closing* the types of nodes and edges, and transforming local identifiers.

RDF is a method for expressing Linked Data on the Web (for a detailed history, see [Soiland-Reyes 2024a]), but has also become a format for building and querying knowledge graphs, where the edge and node identifiers are named using URIs. This allows future extensibility as different RDF graphs of various shapes can be merged by the data scientists, with nodes overlapping based on these global identifiers.

For instance, two repositories like Zenodo and WorkflowHub may both be expressing https://orcid.org/0000-0002-1825-0097 as the author of a dataset and a workflow correspondingly. By merging RDF graphs of this metadata from both repositories into a single knowledge graph, querying for this identifier will find both entities, or even querying for “datasets made by the same author as a workflow” will find the relation. If it is possible to retrieve Linked Data from such identifiers (as is possible from ORCID) then the graph can also be augmented dynamically with additional information.

While RDF allows each data source to use their own types and edges in such knowledge graphs (flexibility), to simplify such queries it is recommended to reuse *vocabularies* where possible. One such vocabulary that has grown in popularity for marking up Web content is schema.org – for instance both sources would declare the author as a http://schema.org/Person although they may vary in which particular attributes of that type are expressed.

Building the WorkflowHub knowledge graph

As detailed in section Using RO-Crate for workflows, in the WorkflowHub repository, each workflow is archived as an RO-Crate [Soiland-Reyes 2022a], which comply with the Workflow RO-Crate profile [Bacall 2022] that specify workflow-specific properties such as input/output parameters.

This builds on the schema.org vocabulary, as well as the Bioschemas ComputationalWorkflow profile. WIthin each workflow's RO-Crate ZIP, the RO-Crate Metadata document is expressed in the RDF-format JSON-LD using these vocabularies.

As part of EuroScienceGateway we have developed a method to build a joint knowledge graph of all the WorkflowHub Workflow RO-Crates (Milestone 5), detailed below.

Handling relative paths in WorkflowHub's RO-Crate

While JSON-LD as a format is compatible with knowledge graphs, as WorkflowHub crates are expressed within a ZIP file rather than directly exposed on the Web, RO-Crate's considerations for handling relative identifiers must also be observed. This is important as a knowledge graph that merges all the WorkflowHub entries may encounter several workflows with the same relative filename.

In short, a unique identifier can be assigned for a ZIP file based on its download URL, e.g. if downloading `https://workflowhub.eu/workflows/415/ro_crate?version=1` then a UUIDv5 can be calculated from hashing this URL: 4979a39d-d733-570f-b838-ad5fef0994eb – and from this a *base URI* of `arcp://uuid,4979a39d-d733-570f-b838-ad5fef0994eb/` (signifying the root of that ZIP file) which can be used when parsing the JSON-LD, so that say a relative filename `conesearch.cwl` becomes `arcp://uuid,4979a39d-d733-570f-b838-ad5fef0994eb/conesearch.cwl`.

It is worth noting that this combines two identifier methods [RFC 4112, Soiland-Reyes 2018] but the resulting URI is not resolvable directly, it is only meaningful together with the RO-Crate download URI, which therefore must also be preserved in the knowledge graph.

We are exploring alternative ways to generate and reference such "inner" identifiers within RO-Crate, as WorkflowHub API also can expose individual files when the origin is a git repository, e.g. https://workflowhub.eu/workflows/502/git/4/raw/vaccine_effectiveness_analytical_pipeline/Dockerfile is a file `vaccine_effectiveness_analytical_pipeline/Dockerfile` from https://workflowhub.eu/workflows/502?version=4 – this challenge becomes relevant when referencing parts of one RO-Crate from another crate.

Workflow for building workflow graph

In order to build a single knowledge graph of all WorkflowHub entries we have developed a Snakemake workflow workflowhub-eu/workflowhub-graph that performs this process:

1. Retrieve list of known workflows in WorkflowHub
2. Retrieve the RO-Crate ZIP for each of the workflows
3. Merge JSON-LD files from each RO-Crate, mapping to global identifiers
4. Save knowledge graph in RDF Turtle format
5. (Quality assurance and statistics) *(planned)*
6. Generate RO-Crate Metadata for knowledge graph

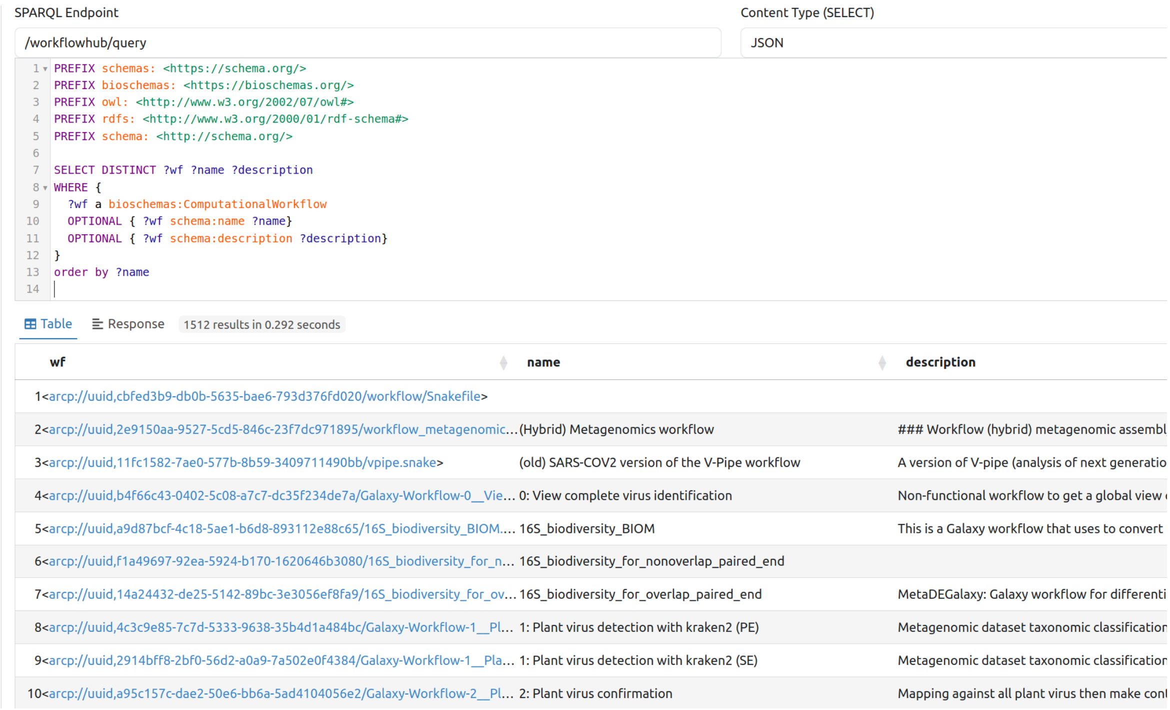
7. Upload to Zenodo using ro-crate-inveniordm *(manually)*

For testing purposes the workflow can be configured to only retrieve a limited set of workflows or to use the Sandbox instance https://dev.workflowhub.eu/ instead of the production instance https://workflowhub.eu/.

The generated WorkflowHub graph is output in the RDF Turtle format, can be loaded in a triple store like Apache Jena Fuseki, then examined using the SPARQL query language, as shown in Figure 10.

The upload to Zenodo [Hambley 2024] is currently done manually until we have integrated quality control measures in the workflow. This will do queries such as ensuring every downloaded crate has a corresponding `ComputationalWorkflow` entity in the graph. Some data cleaning needs have also been identified that will be added at this stage. This Q&A stage will also calculate further statistics that can be added to the outer RO-Crate for the knowledge graph itself.

The use of Snakemake allows repeated runs of the workflow without redownloading existing versioned RO-Crates, and we are planning to set up automatic deployment as part of the workflowhub.eu server in UNIMAN, which will regularly update Zenodo records with the latest knowledge graph dump, e.g. every week.



SPARQL Endpoint: /workflowhub/query

Content Type (SELECT): JSON

```
PREFIX schemas: <https://schema.org/>
PREFIX bioschemas: <https://bioschemas.org/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX schema: <http://schema.org/>

SELECT DISTINCT ?wf ?name ?description
WHERE {
  ?wf a bioschemas:ComputationalWorkflow
  OPTIONAL { ?wf schema:name ?name}
  OPTIONAL { ?wf schema:description ?description}
}
order by ?name
```

1512 results in 0.292 seconds

	wf	name	description
1	<arcp://uuid,cbfed3b9-db0b-5635-bae6-793d376fd020/workflow/Snakefile>		
2	<arcp://uuid,2e9150aa-9527-5cd5-846c-23f7dc971895/workflow_metagenomic…	(Hybrid) Metagenomics workflow	### Workflow (hybrid) metagenomic assembl
3	<arcp://uuid,11fc1582-7ae0-577b-8b59-3409711490bb/vpipe.snake>	(old) SARS-COV2 version of the V-Pipe workflow	A version of V-pipe (analysis of next generatio
4	<arcp://uuid,b4f66c43-0402-5c08-a7c7-dc35f234de7a/Galaxy-Workflow-0__Vie…	0: View complete virus identification	Non-functional workflow to get a global view
5	<arcp://uuid,a9d87bcf-4c18-5ae1-b6d8-893112e88c65/16S_biodiversity_BIOM.…	16S_biodiversity_BIOM	This is a Galaxy workflow that uses to convert
6	<arcp://uuid,f1a49697-92ea-5924-b170-1620646b3080/16S_biodiversity_for_n…	16S_biodiversity_for_nonoverlap_paired_end	
7	<arcp://uuid,14a24432-de25-5142-89bc-3e3056ef8fa9/16S_biodiversity_for_ov…	16S_biodiversity_for_overlap_paired_end	MetaDEGalaxy: Galaxy workflow for differenti
8	<arcp://uuid,4c3c9e85-7c7d-5333-9638-35b4d1a484bc/Galaxy-Workflow-1__Pl…	1: Plant virus detection with kraken2 (PE)	Metagenomic dataset taxonomic classification
9	<arcp://uuid,2914bff8-2bf0-56d2-a0a9-7a502e0f4384/Galaxy-Workflow-1__Pla…	1: Plant virus detection with kraken2 (SE)	Metagenomic dataset taxonomic classification
10	<arcp://uuid,a95c157c-dae2-50e6-bb6a-5ad4104056e2/Galaxy-Workflow-2__Pl…	2: Plant virus confirmation	Mapping against all plant virus then make con

Figure 10: SPARQL Query in Fuseki to select name and description for every workflow.

Initial statistics and example queries

The SPARQL query for Figure 10 selects name and description for each workflow:

```
PREFIX bioschemas: <https://bioschemas.org/>
PREFIX schema: <http://schema.org/>

SELECT DISTINCT ?wf ?name ?description
WHERE {
  ?wf a bioschemas:ComputationalWorkflow
  OPTIONAL { ?wf schema:name ?name}
  OPTIONAL { ?wf schema:description ?description}
}
ORDER BY ?name
```

This query for counts number of workflows per licence, shown in Table 4:

```
PREFIX bioschemas: <https://bioschemas.org/>
PREFIX schema: <http://schema.org/>

SELECT ?license (COUNT(?wf) AS ?workflows)
WHERE {
  ?wf a bioschemas:ComputationalWorkflow .
  ?wf schema:license ?license .

}
GROUP BY ?license
ORDER BY DESC(?workflows)
```

license	**workflows**
https://spdx.org/licenses/MIT	594
https://spdx.org/licenses/Apache-2.0	305
https://spdx.org/licenses/CC-BY-4.0	110
https://spdx.org/licenses/GPL-3.0	79
https://spdx.org/licenses/CC-BY-NC-4.0	12
(unknown)	12
https://spdx.org/licenses/LGPL-3.0	10
https://spdx.org/licenses/CC0-1.0	10
https://spdx.org/licenses/BSD-2-Clause	8
https://choosealicense.com/no-permission/	8
https://spdx.org/licenses/BSD-3-Clause	4

https://spdx.org/licenses/GPL-3.0+	1
https://spdx.org/licenses/GPL-2.0	1
https://spdx.org/licenses/CECILL-2.1	1
https://spdx.org/licenses/CC-BY-SA-4.0	1
https://spdx.org/licenses/CC-BY-NC-SA-4.0	1
https://spdx.org/licenses/AGPL-3.0-or-later	1
https://spdx.org/licenses/AGPL-3.0	1
https://spdx.org/licenses/AFL-3.0	1

Table 4: Specific workflow licences in WorkflowHub. Note that all WorkflowHub entries have also got a licence for the overall RO-Crate, which may differ from the above.

This query selects how many properties have been used to describe each type of entity (Table 5). The inner subquery here selects which properties `?prop` are used for each `?class`, while the outer query counts them per aggregated class:

```
PREFIX bioschemas: <https://bioschemas.org/>
PREFIX schema: <http://schema.org/>

SELECT ?class (COUNT(?prop) AS ?properties)
WHERE {
 SELECT DISTINCT ?class ?prop WHERE {
  ?any a ?class ;
        ?prop ?obj .
    }
}
GROUP BY ?class
ORDER BY DESC(?properties)
LIMIT 14
```

class	properties
http://schema.org/Dataset	41
http://schema.org/MediaObject	37
http://schema.org/SoftwareSourceCode	32
https://bioschemas.org/ComputationalWorkflow	30
http://schema.org/SoftwareApplication	13
http://schema.org/CreateAction	11
http://schema.org/CreativeWork	11

http://schema.org/ComputerLanguage	10
http://schema.org/PropertyValue	8
http://schema.org/DataDownload	7
http://schema.org/ImageObject	7
http://schema.org/Organization	7
http://schema.org/ScholarlyArticle	7
arcp://uuid,01c8b5d3-81a5-52db-876b-545a09674f28/WorkflowSketch	6

Table 5: Types and their number of unique properties used within all RO-Crates in WorkflowHub. Note that RO-Crate `File` is an alias for http://schema.org/MediaObject

Select how many RO-Crates have declared which RO-Crate profiles (Table 6):

```
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX bioschemas: <https://bioschemas.org/>
PREFIX schema: <http://schema.org/>

SELECT ?profile (COUNT(?ro) AS ?crates)
WHERE {
       ?ro schema:about ?dataset .
       { ?ro dct:conformsTo ?profile } UNION { ?dataset dct:conformsTo ?profile }
}
GROUP BY ?profile
ORDER BY DESC(?crates)
```

profile	crates
https://w3id.org/ro/crate/1.1	3787
https://w3id.org/workflowhub/workflow-ro-crate/1.0	2461
https://w3id.org/ro/wfrun/process/0.1	73
https://w3id.org/ro/wfrun/workflow/0.1	64
https://w3id.org/ro/wfrun/process/0.4	20
https://w3id.org/ro/wfrun/workflow/0.4	20
https://w3id.org/ro/wfrun/process/0.5	4
https://w3id.org/ro/wfrun/workflow/0.5	4

Table 6: RO-Crate profiles and how many deposits declares conformance

Notable from Table 6 is that there are about 80 crates with Workflow Run Crate profiles (section Workflow provenance helps explain workflow use). Although WorkflowHub has not (currently) got any specific support for WRROC, as the provenance profiles expands Workflow RO-Crate (section Using RO-Crate for workflows), these crates are nevertheless compatible with WorkflowHub.

Further knowledge graph developments

We will be further developing the knowledge graph to improve its usability and interoperability.

In particular we have identified some data cleaning needs:

- Deposits with the older version of the Workflow RO-Crate profile use relative identifiers like "#galaxy" for programming language, meaning that with absolute URIs (section Handling relative paths in WorkflowHub's RO-Crate) in the graph the same workflow systems differ across crates. Data cleaning can merge known workflow systems to to their PIDs e.g. https://w3id.org/workflowhub/workflow-ro-crate#galaxy
- Some GTN deposits provide ORCID using local identifiers like `#0000-0001-9842-9718` instead of https://orcid.org/0000-0001-9842-9718 - data cleaning can recognize the particular ID pattern of ORCIDs and transform.
- Licences on the RO-Crates are expressed in many different ways, typically as strings like `"GPL-2.0"`. They should be unified to SPDX identifiers as shown in Table 4.
- The type `WorkflowSketch` is inadvertently mapped to `arcp://uuid,01c8b5d3-81a5-52db-876b-545a09674f28/WorkflowSketch` etc as it is not in the JSON-LD context. (Table 5)

We will also add corresponding updates and fixes to Workflow RO-Crate profile and tooling based on these identified issues. We will however only be updating the RO-Crates that have been generated by WorkflowHub, and not retrospectively modify any RO-Crates that have been submitted as-is (section Encouraging research software best practices for workflows) - here we will rather report the issues upstream.

Additional data can be added to the knowledge graph from the organisational structure of WorkflowHub, which is not yet fully shown in the RO-Crate, but has Bioschemas metadata (see section Using Signposting for FAIR Digital Objects).

- Submitting User and their Organisation
- Collections containing workflow
- Teams and Spaces that "own" the workflow
- Assigned DOI
- List of versions (currently only latest workflow version is included in graph)

Additional data can be added from external sources, which can be mapped or consumed as FAIR resources and included in the graph:

- For each author, find other schema.org data from ORCID, e.g. country, affiliation, publications
- ROR organisation identifiers (e.g. https://ror.org/027m9bs27 for University of Manchester)
- SPDX licence information from https://github.com/spdx/license-list-data/tree/main/rdfturtle
- Details of tools used by Galaxy workflows, including EDAM annotations of their purpose
- RDF transcription of CWL workflows, e.g. using CWL Viewer

Alternative formats and subsets of the knowledge graph can also be generated:

- Named graphs in RDF TriG format, e.g. to distinguish properties such as an author's full name depending on which crate stated it.
- Every RO-Crate JSON-LD as-is (alternative ways to parse these can however easily modify the Snakemake workflow)
- JSON-LD using *Framing* to create a nested JSON tree of selected objects (e.g. a `ComputationalWorkflow`) – this can be consumed by GraphQL and other JSON-based knowledge graphs.

Annotating and sharing workflows

Galaxy have developed a workflow annotation mechanism for graphically grouping and describing various parts of the pipeline. We have now made it possible for any web page to embed interactive Galaxy workflow diagrams that show these descriptions [Los 2024]. This is a powerful explanation mechanism by documenting the workflow visually, as demonstrated in Figure 11.

🤯 Is that a Workflow on the Community Hub?

It sure is!

And it's also the first feature we'd like to show you: Workflow Embeds

Full Analyse Argo data

Galaxy Earth system 🌍
https://earth-system.usegalaxy.eu/
Come try our earth system **subdomain** and test our tutorials with the **earth-system** tag !

Pangeo in Galaxy
There is no single software package called "pangeo"; rather, the Pangeo project serves as a coordination point between scientists, software, and computing infrastructure. The Pangeo software ecosystem involves open source tools such as xarray as you can see in te tolls used below.

Argo
It's an Argo is an international program that collects information from inside the ocean using a fleet of robotic instruments that drift with the ocean currents and move up and down between the surface and a mid-water level.

The studied area here is the Carraibean sea from April to May 2021. Feel free to change to the ocean part that you're interested in !

Can be temperature, pressure, salinity, ...

Figure 11: Workflow https://usegalaxy.eu/published/workflow?id=a80f9b926ba43892 embedded with interactive navigation in the Galaxy Community Hub blog post [[Los 2024]]. The Community Hub is rendered as static HTML pages [from Markdown sources], where the workflow preview is included using `<iframe>`, similar to embedded YouTube videos.

This is a new Galaxy feature which we're exploring how users make best use of. Earlier workflow systems that explored such "free hand" annotations include KNIME [[Fillbrunn 2017]] and Taverna Data Playground [[Gibson 2009]]. This way of explaining a pipeline presents both a challenge and great opportunity for Workflow FAIR Digital Objects and WorkflowHub:

1. Tools grouped together typically perform some scientific function; common *workflow motifs* include Data retrieval, Data cleaning etc. [[Garijo 2013]]. This is clearly important for explainability, and can be connected to established FAIR resources like the [EDAM ontology]. Such semantic grouping can be useful for instance for enhancing workflow discovery by their methods, and also for finding the purpose of individual tools.
   However the Galaxy annotation is not currently semantically linked to the tools in Galaxy's saved JSON `.ga` representation, rather the tools are geographically "near" the annotation in x,y coordinates. This means the implied motif grouping is not directly machine-readable. The grouping is however available within Galaxy's code

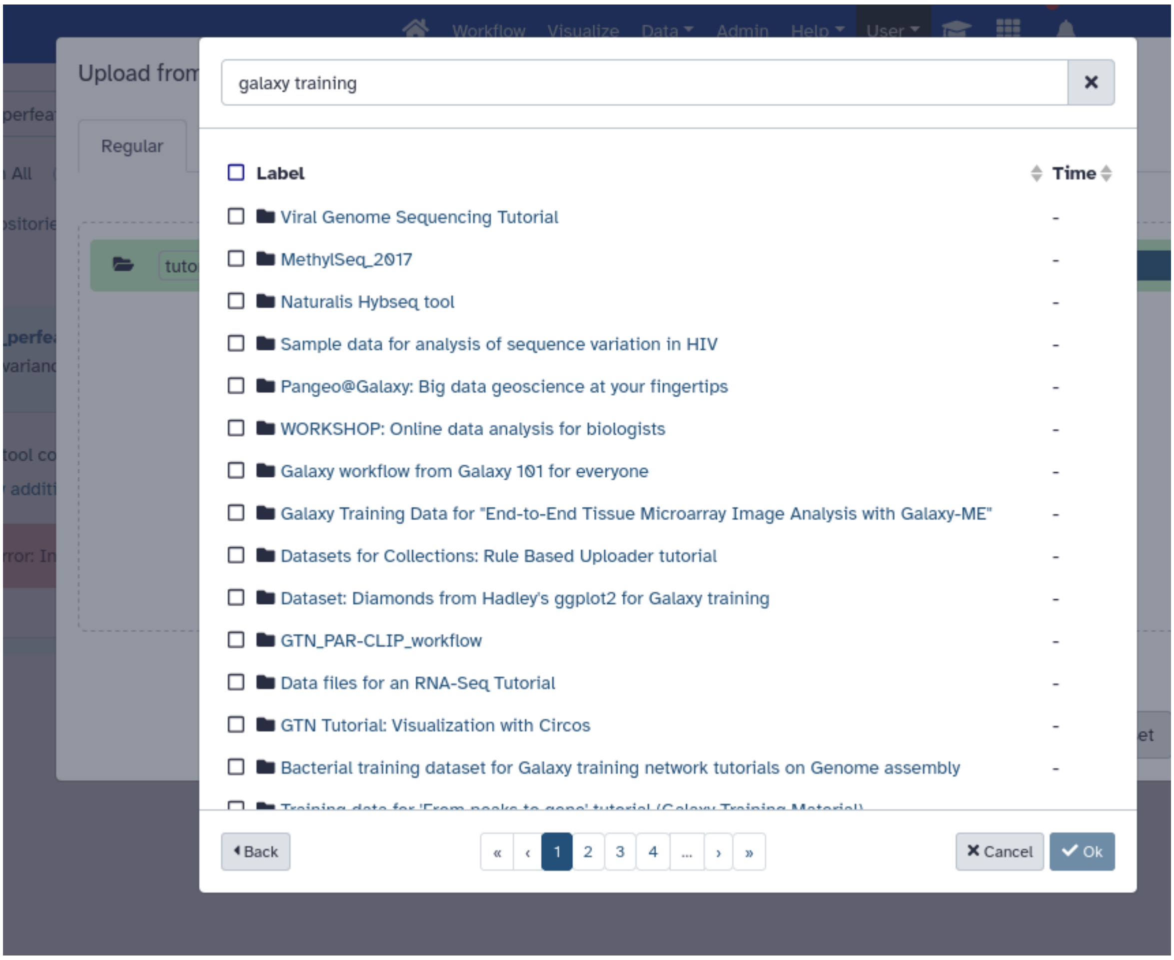
while editing, so this could be saved as a secondary annotation to make it accessible and FAIR.

2. Established FAIR annotation models include *Web Annotation Data Model* [Sanderson 2017], which include powerful selector mechanisms and rich annotation properties. However existing selectors assume an image of fixed dimensions, while the Galaxy workflow embedding is effectively a *read only* view of the workflow editor, not a static image. Repositioning a tool graphically also means moving its annotation.
3. WorkflowHub could be expanded to embed the interactive diagram, however the iframe embeds from a "live" entry at a Galaxy server like usegalaxy.eu, while currently WorkflowHub registration is done by upload or reference to a Git repository. Further APIs may be needed to query which Galaxy servers have a particular workflow version installed, or if the workflow is in IWC then this is already guaranteed for the latest version.
4. Adding Signposting (section Signposting and RO-Crate) for navigating between WorkflowHub DOIs, WorkflowHub entries, Galaxy workflows, and embedded workflows can help "wake up" a workflow (FDO machine actionability) and in the other direction make embedded views display citation information. However, this requires additional WorkflowHub discovery or notification of a workflow view having been made public by the owner at the Galaxy Server.

Using FAIR digital objects from Galaxy

In Galaxy we have extended the support for file storage backends to support institutional storage systems such as ownCloud and repositories like InvenioRDM. These are powerful ways to include large data in workflows, as Galaxy can refer to such data by reference, which can be taken advantage of by the Bring Your Own Data (BYOD) mechanism in Pulsar network (WP3) by computing the workflow near such data.

In EuroScienceGateway we have further improved this support to do a paginated filtered search and also added explicit connectors to the EU-wide Zenodo repository [López 2024], shown in Figure 12.



galaxy training

☐ Label	Time
☐ Viral Genome Sequencing Tutorial	-
☐ MethylSeq_2017	-
☐ Naturalis Hybseq tool	-
☐ Sample data for analysis of sequence variation in HIV	-
☐ Pangeo@Galaxy: Big data geoscience at your fingertips	-
☐ WORKSHOP: Online data analysis for biologists	-
☐ Galaxy workflow from Galaxy 101 for everyone	-
☐ Galaxy Training Data for "End-to-End Tissue Microarray Image Analysis with Galaxy-ME"	-
☐ Datasets for Collections: Rule Based Uploader tutorial	-
☐ Dataset: Diamonds from Hadley's ggplot2 for Galaxy training	-
☐ GTN_PAR-CLIP_workflow	-
☐ Data files for an RNA-Seq Tutorial	-
☐ GTN Tutorial: Visualization with Circos	-
☐ Bacterial training dataset for Galaxy training network tutorials on Genome assembly	-
☐ Training data for 'From peaks to genes' tutorial (Galaxy Training Material)	

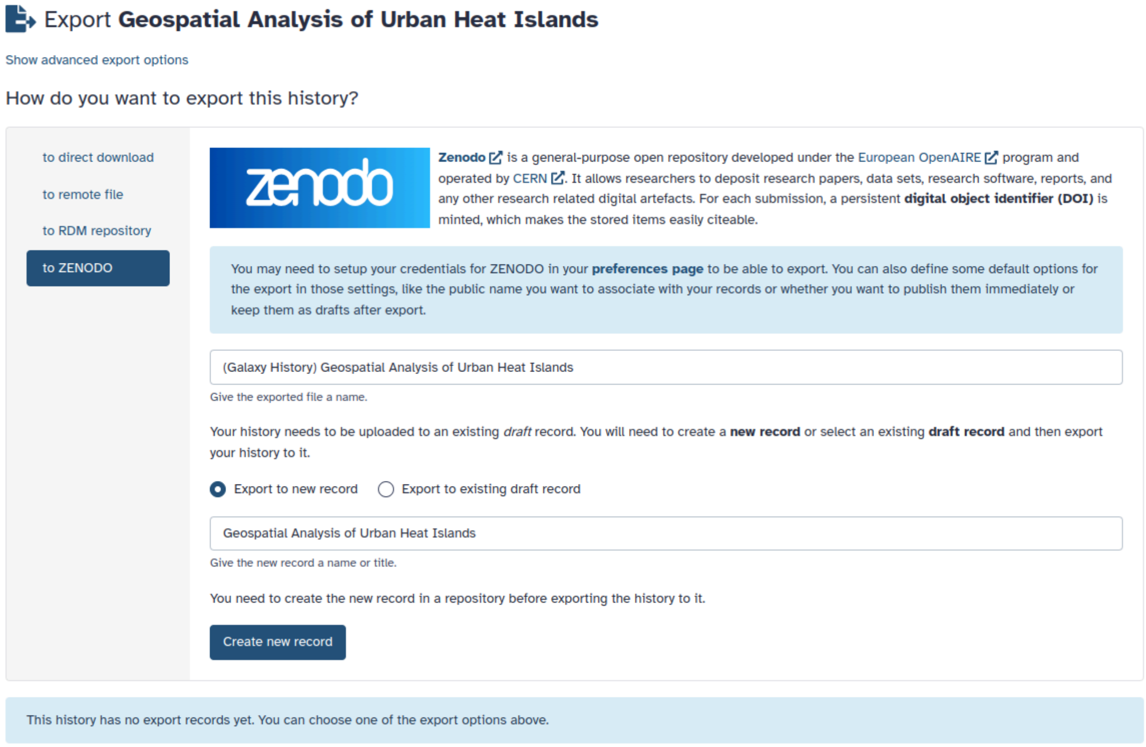
Figure 12: Galaxy data import showing a Dataset search for *galaxy training* from Zenodo, equivalent to https://zenodo.org/search?q=galaxy%20training

The current import (see screencast) uses the general file import mechanism in Galaxy, and does not have particular requirements on the underlying data sources. Envisioned further work to expand on this support from an FDO perspective include:

1. Import from any persistent identifier (e.g. Zenodo DOI), using Signposting to resolve to data
2. Propagation of metadata from upstream repository, for further embedding in RO-Crate (e.g. PID, title and author in order to comply with licences like CC-BY-SA 4.0)
3. Guided import of data sources that are published as RO-Crate, e.g. selection of particular resources based on their types. Matching to Galaxy data types.

In addition to importing, we have also improved Galaxy **export of histories**. The Galaxy *history* includes the data files that have been progressively used and generated by a Galaxy user, along with the Tool settings for each analysis. Note that in Galaxy the history does not necessarily imply a Galaxy *workflow*, however a workflow can be extracted from the history.

In EuroScienceGateway we have connected the export mechanism to the new file storage systems, including Zenodo, shown in Figure 13 and a screencast. The generated history includes all the history data and an RO-Crate description of each data item. The user may choose to store as a *draft* record to complete additional metadata in the Zenodo UI, or publish it directly. When the Zenodo record has been published, its generated DOI is recorded by Galaxy and shown as part of archived histories.



Export **Geospatial Analysis of Urban Heat Islands**

How do you want to export this history?

Zenodo is a general-purpose open repository developed under the European OpenAIRE program and operated by CERN. It allows researchers to deposit research papers, data sets, research software, reports, and any other research related digital artefacts. For each submission, a persistent **digital object identifier (DOI)** is minted, which makes the stored items easily citeable.

You may need to setup your credentials for ZENODO in your **preferences page** to be able to export. You can also define some default options for the export in those settings, like the public name you want to associate with your records or whether you want to publish them immediately or keep them as drafts after export.

(Galaxy History) Geospatial Analysis of Urban Heat Islands

Give the exported file a name.

Your history needs to be uploaded to an existing *draft* record. You will need to create a **new record** or select an existing **draft record** and then export your history to it.

● Export to new record ○ Export to existing draft record

Geospatial Analysis of Urban Heat Islands

Give the new record a name or title.

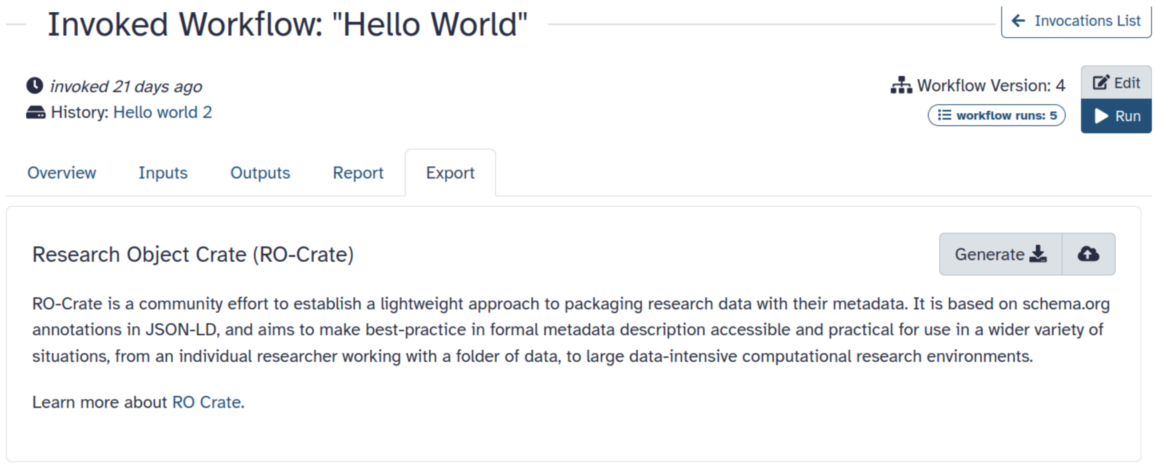
You need to create the new record in a repository before exporting the history to it.

This history has no export records yet. You can choose one of the export options above.

Figure 13: Galaxy export of execution history to create a new record in the Zenodo repository.

Histories published as such RO-Crates can later be reloaded by another Galaxy instance, showing each tool execution as if it had happened there.

In comparison, *Workflow Invocations* are tracked separately in Galaxy and connected to a workflow definition. These can also be exported to a selection of file storage systems. In this case the files are exported as a Workflow Run Crate that embeds the Galaxy workflow definition (section Workflow provenance helps explain workflow use), shown in Figure 14.



RO-Crate is a community effort to establish a lightweight approach to packaging research data with their metadata. It is based on schema.org annotations in JSON-LD, and aims to make best-practice in formal metadata description accessible and practical for use in a wider variety of situations, from an individual researcher working with a folder of data, to large data-intensive computational research environments.

Learn more about RO Crate.

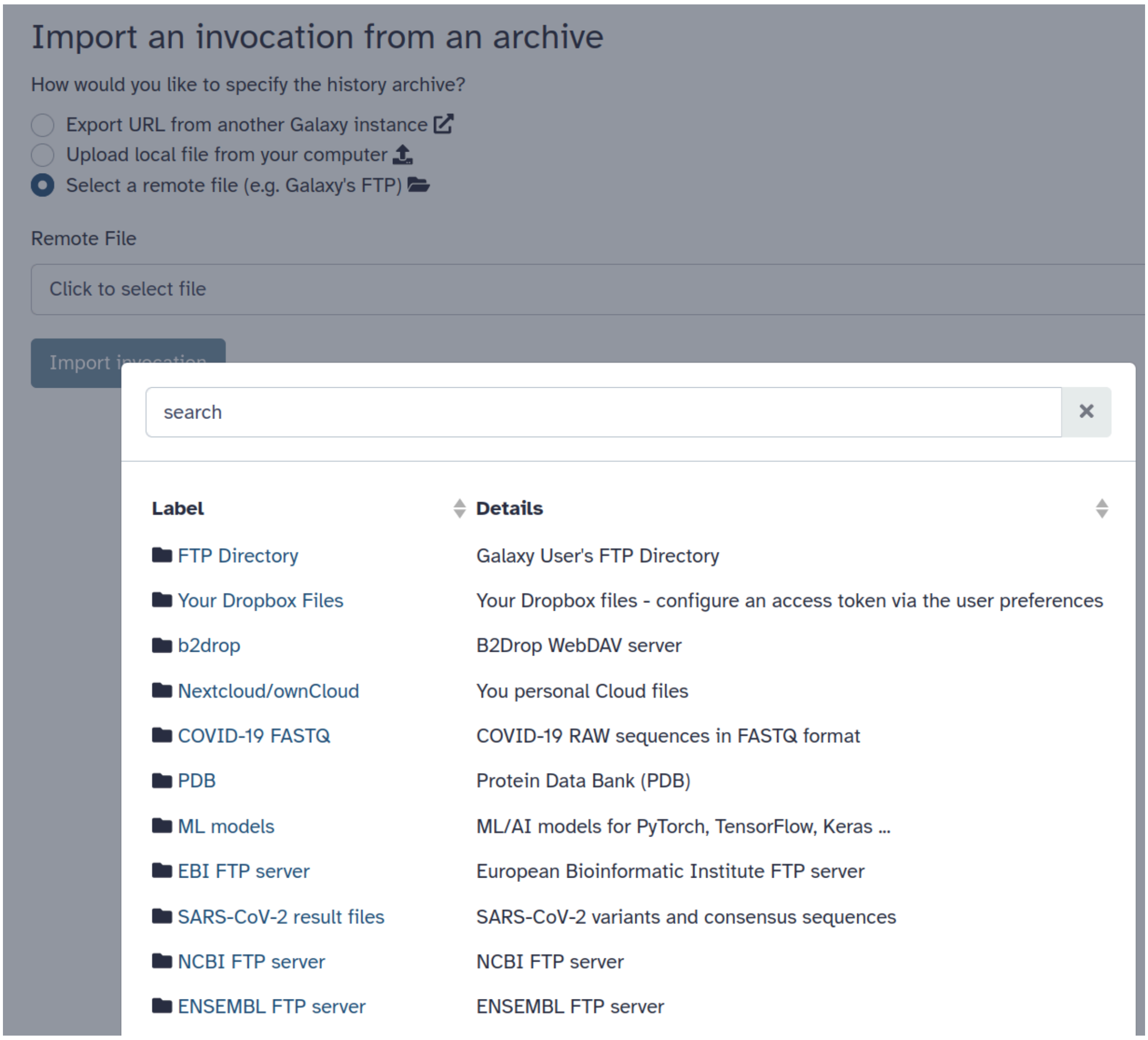
Figure 14: RO-Crate export of a Galaxy workflow invocation.

Such invocation RO-Crates can likewise be *Imported* from the supported file storage mechanisms, as well as from a URL (Figure 15).

Annotated Research Contexts (ARC) is a way to structure plant experiments with workflows in an RO-Crate [Beier 2023]. The NFDI DataPlant initiative has used ARC in Galaxy [Schaaf 2023], and recently also in the Molecular Adaptation to Land (MADland) programme [Varshney 2024]. With help from EuroScienceGateway, Galaxy has added a DataPlant git as a dedicated data source on UseGalaxy.eu that can be used in the UI, e.g. to import an ARC.

From a FDO perspective, further possibilities in Galaxy's RO-Crate mechanisms include:

1. Process executions can be documented as a Process Run Crate [WRROC 2024a] with multiple tool executions and an implied workflow where output and input data match across steps. (It may not be reliable to always extract the workflow, as some steps may have been removed from the history by the user, or a tool was run multiple times)
2. Existing metadata (e.g. from data imports) should be mentioned with citations in the RO-Crate
3. Import of any RO-Crate into the history with graceful "upgrade" if it was a previous Galaxy history or Galaxy Workflow Invocation. Currently different export and import mechanisms are needed.
4. Import of RO-Crate from a PID, using Signposting to match to a supported file storage or URL download.



Import an invocation from an archive

How would you like to specify the history archive?

○ Export URL from another Galaxy instance
○ Upload local file from your computer
● Select a remote file (e.g. Galaxy's FTP)

Remote File

Click to select file

Label	Details
FTP Directory	Galaxy User's FTP Directory
Your Dropbox Files	Your Dropbox files - configure an access token via the user preferences
b2drop	B2Drop WebDAV server
Nextcloud/ownCloud	You personal Cloud files
COVID-19 FASTQ	COVID-19 RAW sequences in FASTQ format
PDB	Protein Data Bank (PDB)
ML models	ML/AI models for PyTorch, TensorFlow, Keras ...
EBI FTP server	European Bioinformatic Institute FTP server
SARS-CoV-2 result files	SARS-CoV-2 variants and consensus sequences
NCBI FTP server	NCBI FTP server
ENSEMBL FTP server	ENSEMBL FTP server

Figure 15: Galaxy invocation import from configured data sources

Reproducing workflow runs in Galaxy and WfExS

A motivation for doing Workflow Run Crate (WRROC) export from workflow management systems like Galaxy and WfExS is to support *reproducibility*. The simplest form of reproducibility of computational workflows is *rerunability*, that is to execute with the same inputs again on an equivalent platform, to verify that the computational tools produce the expected outputs. In this document, *Replicability* is reproducibility where one or more factors are modified, e.g. different inputs, different installation.

The distinction between rerun, replicate and reproduce is a sliding scale for computational analyses, as even in the simplest case, some factors are necessarily different. For instance, running the very same workflow with the very same inputs and very same tools on the same Galaxy instance may still experience small technical changes over time, giving differences in output, e.g. different compute nodes may be scheduled for the tasks, or the tool relies on external data sources or random seeds.

Reproducibility in Galaxy

After importing an existing Galaxy workflow invocation, it is possible to re-execute it. Existing inputs are shown as in the original run (rerun), but with the possibility to modify some of these (light reproducibility) . As the workflow is included in the RO-Crate and is editable in Galaxy, users can further modify it, to do a slightly different analysis (reuse).

Reexecution for imported Galaxy workflow histories are more complex, as each step must be executed in order, and this produces new outputs that must be reconnected in the corresponding next step rather than the old output (the overall workflow is implied).

For Galaxy histories, practical re-runnability would be to extract a workflow before re-executing it. It is however an advantage if this is rather done by the original author, which is better informed to clean up the workflow for unnecessary steps. However, one advantage of "step by step" reproducibility is that it is possible to bypass a tool which no longer produces a valid or desired result, by using the old value from the history for subsequent steps.

Challenges with reproducibility include:

1. Provenance of a rerun or reproduced RO-Crate should cite the original, which may have been executed by someone else.
2. Edited workflows from an imported WRROC should propagate any citation information of the original author (e.g. at WorkflowHub), which may not be the same as the users who executed the workflow and made the first WRROC.
3. Provenance of derived WRROC implies a provenance from one RO-Crate to another, not just for its individual files. For this, versioned identifiers must be ensured, e.g. WorkflowHub DOIs (Figure 7).
4. Tools used by the workflow should be available at the Galaxy instance where the WRROC is imported. In Galaxy, only system administrators are able to add new tools. The installed tool may be in a different version than used by the original workflow, but this can be highlighted by Galaxy.

Reproducibility in WfExS

The workflow orchestrator WfExS has support for generating WRROC for any of the supported workflow systems (currently Nextflow and CWL) [Fernández 2024]. WfExS can also export the used container images as part of the RO-Crate. We have recently also expanded WfExS to support *rerunnability* of WRROC crates at different compliance levels, with the potential to override particular inputs (*reproducibility and replicability*).

The potential of rerunning with container image snapshots is very powerful, as computational tools can be captured in the version and binary form used at a particular time. Workflow systems like Common Workflow Language support container image references, but these are frequently not versioned. There is also the potential of infrequently used images being deleted after some time, as is the policy of Docker Hub, meaning workflows with versioned containers may no longer run just 6 months later.

While researchers generally prefer running the latest version of tools in a workflow, sometimes these evolve beyond the retrocompatibility, requiring changes to the workflow. This mechanism would allow more precise reproducibility of workflows using older tools with new parameters.

This WfExS feature is also being explored by the EOSC-ENTRUST project and HDR UK Federated Analytics program, as a mechanism for moving a workflow's container dependencies inside the "airlock" of a Trusted Research Environment (TRE), where strict firewalls prevent direct software downloads e.g. from Docker Hub. In this case the workflow can be executed as a "dry run" outside the TRE (using synthetic/test inputs) to populate the containers, with the full WRROC moved inside the TRE, to be used as a base for the actual execution on sensitive data (*automatic replicability*).

Reproducibility and replicability efforts can be hindered by workflows which have steps depending on external services (e.g. they could not be reached within a TRE, they are discontinued or temporarily unavailable). In these scenarios, metadata and data gathered by WRROC snapshots are crucial to ease authors to modify the workflow to avoid such external services. For this, further FDO aspects such as moving Data RO-Crates along with the Workflow RO-Crate may be needed.

References

[Anders 2023] Ivonne Anders, Christophe Blanchi, Daan Broder, Maggie Hellström, Sharif Islam, Thomas Jejkal, Larry Lannom Larry, Karsten Peters-von Gehlen, Robert Quick, Alexander Schlemmer, Ulrich Schwardmann, Stian Soiland-Reyes, George Strawn, Dieter van Uytvanck, Claus Weiland, Peter Wittenburg, Carlo Zwölf (2023):
FAIR Digital Object technical overview. Version PEN 2.0.
FDO Specification Documents Full FDO Overview PEN-2.0-v2
FAIR Digital Objects Forum
https://doi.org/10.5281/zenodo.7824714

[Atkinson 2017] Malcolm Atkinson, Sandra Gesing, Johan Montagnat, Ian Taylor (2017):
Scientific workflows: Past, present and future.
Future Generation Computer Systems 75
https://doi.org/10.1016/j.future.2017.05.041

[Bacall 2022] Finn Bacall, Alan R. Williams, Stuart Owen, Stian Soiland-Reyes (2022):
Workflow RO-Crate Profile 1.0.
WorkflowHub community
https://w3id.org/workflowhub/workflow-ro-crate/1.0

[Beer 2023] Philipp Beer, Milan Szente (2023):
RO-Crates Data Deposit. v1.0.2
Zenodo
https://doi.org/10.5281/zenodo.8127644

[Beer 2024] Philipp Beer, Milan Szente, Eli Chadwick (2024):
rocrate-inveniordm. v2.0.3
Zenodo
https://doi.org/10.5281/zenodo.13366072

[Beier 2023] Sebastian Beier, Timo Mühlhaus, Cyril Pommier, Stuart Owen, Dominik Brilhaus, Heinrich Lukas Weil, Florian Wetzels, Gavin Chait, Daniel Arend, Manuel Feser, Gajendra Doniparthi, Jonathan Bauer, Sveinung Gundersen, Pável Vázquez (2024):
BioHackEU23 report: Enabling continuous RDM using Annotated Research Contexts with RO-Crate profiles for ISA.
BioHackrXiv
https://doi.org/10.37044/osf.io/7y2jh

[Broeder 2024] D. Broeder, A. Fouilloux, E. Schultes, P. Wittenburg (eds):
FDO Related Solutions.
FDO Forum (internal draft)

[Chadwick 2024] Eli Chadwick, Stian Soiland-Reyes (2024):
rocrate-zenodo. v0.1.1
Zenodo
https://doi.org/10.5281/zenodo.13365999

[Datacite 2021] DataCite Metadata Working Group (2021).
DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs. Version 4.4.
DataCite e.V.
https://doi.org/10.14454/3w3z-sa82

[De Geest 2022] Paul De Geest, Frederik Coppens, Stian Soiland-Reyes, Ignacio Eguinoa, Simone Leo (2022):
Enhancing RDM in Galaxy by integrating RO-Crate.
1st International Conference on FAIR Digital Objects (FDO 2022) (poster)
Research Ideas and Outcomes **8**:e95164
https://doi.org/10.3897/rio.8.e95164 [poster]

[Edmunds 2024]
Scott Edmunds (2024):
A Decade of FAIR – and what next? Q&A on FAIR workflows with the Netherlands X-omics Initiative.
GIGABlog, 2024-01-14
http://gigasciencejournal.com/blog/fair-workflows/

[Erxleben 2024] Anika Erxleben (2024):
Horizon Europe project EOSC EuroScienceGateway at EOSC Winter School 2024.
Galaxy Community Hub
https://galaxyproject.org/news/2024-02-27-eosc-winter-school-2024/

[FDO-Specs] **FAIR Digital Object Forum specifications**.
https://fairdo.org/specifications/

[Fernández 2024] José M. Fernández, Paula Iborra, Sébastien Moretti, Arun Isaac, Paul De Geest, Stian Soiland-Reyes (2024):
BioHackEU23: FAIR Workflow Execution with WfExS and Workflow Run Crate.
BioHackrXiv
https://doi.org/10.37044/osf.io/7f94w

[Ferreira da Silva 2023] Rafael Ferreira da Silva, Rosa M. Badia, Venkat Bala, Debbie Bard, Peer-Timo Bremer, Ian Buckley, Silvina Caino-Lores, Kyle Chard, Carole Goble, Shantenu Jha, Daniel S. Katz, Daniel Laney, Manish Parashar, Frederic Suter, Nick Tyler, Thomas Uram, Ilkay Altintas, Stefan Andersson, William Arndt, Juan Aznar, Jonathan Bader, Bartosz Balis, Chris Blanton, Kelly Rosa Braghetto, Aharon Brodutch, Paul Brunk, Henri Casanova, Alba Cervera Lierta, Justin Chigu, Taina Coleman, Nick Collier, Iacopo Colonnelli, Frederik Coppens, Michael Crusoe, Will Cunningham, Bruno de Paula Kinoshita, Paolo Di Tommaso, Charles Doutriaux, Matthew Downton, Wael Elwasif, Bjoern Enders, Chris Erdmann, Thomas Fahringer, Ludmilla Figueiredo, Rosa Filgueira, Martin Foltin, Anne Fouilloux, Luiz Gadelha, Andy Gallo, Artur Garcia Saez, Daniel Garijo, Roman Gerlach, Ryan Grant, Samuel Grayson, Patricia Grubel, Johan Gustafsson, Valerie Hayot-Sasson, Oscar Hernandez, Marcus Hilbrich, AnnMary Justine, Ian Laflotte, Fabian Lehmann, Andre Luckow, Jakob Luettgau, Ketan Maheshwari, Motohiko Matsuda, Doriana Medic, Pete Mendygral, Marek Michalewicz, Jorji Nonaka, Maciej Pawlik, Loic Pottier, Line Pouchard, Mathias Putz, Santosh Kumar Radha, Lavanya Ramakrishnan, Sashko Ristov, Paul Romano, Daniel Rosendo, Martin Ruefenacht, Katarzyna Rycerz, Nishant Saurabh, Volodymyr Savchenko, Martin Schulz, Christine Simpson, Raul Sirvent, Tyler Skluzacek, Stian Soiland-Reyes, Renan Souza, Sreenivas Rangan Sukumar, Ziheng Sun, Alan Sussman, Douglas Thain, Mikhail Titov, Benjamin Tovar, Aalap Tripathy, Matteo Turilli, Bartosz Tuznik, Hubertus van Dam, Aurelio Vivas , Logan Ward, Patrick Widener, Sean Wilkinson, Justyna Zawalska, Mahnoor Zulfiqar (2023):
Workflows Community Summit 2022: A Roadmap Revolution.
arXiv:2304.00019
https://doi.org/10.48550/arXiv.2304.00019

[Fillbrunn 2017] Alexander Fillbrunn, Christian Dietz, Julianus Pfeuffer, René Rahn, Gregory A. Landrum, Michael R. Berthold (2017):
KNIME for reproducible cross-domain analysis of life science data
Journal of Biotechnology **261**
https://doi.org/10.1016/j.jbiotec.2017.07.028

[Galaxy 2024] The Galaxy Community (2024):
The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update.
Nucleic Acids Research **52**(W1)
https://doi.org/10.1093/nar/gkae410

[Garijo 2013] Daniel Garijo, Pinar Alper, Khalid Belhajjame, Oscar Corcho, Yolanda Gil, Carole Goble (2013):
Common motifs in scientific workflows: An empirical analysis.
Future Generation Computer Systems **36**
https://doi.org/10.1016/j.future.2013.09.018

[Gibson 2009] Andrew Gibson, Matthew Gamble, Katy Wolstencroft, Tom Oinn, Carole Goble, Khalid Belhajjame, Paolo Missier (2009):
The data playground: An intuitive workflow specification environment
Future Generation Computer Systems **25**(4)
https://doi.org/10.1016/j.future.2008.09.009

[Goble 2022] 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

[Goble 2024] Carole Goble (2024):
WorkflowHub Publishers and Journal Forum.
Galaxy Community Hub
https://galaxyproject.org/news/2024-08-03-workflow-publisher-forum/

[Hambley 2024] Alexander Hambley, Eli Chadwick, Oliver Woolland, Stian Soiland-Reyes, Volodymyr Savchenko (2024):
WorkflowHub Knowledge Graph. 2024-08-22
Zenodo
https://doi.org/10.5281/zenodo.13362051

[Hettne 2012] Kristina Hettne, Katy Wolstencroft, Khalid Belhajjame, Carole Goble, Eleni Mina, Harish Dharuri, Lourdes Verdes-Montenegro, Julián Garrido, David de Roure, Marco Roos (2012):
Best practices for workflow design: how to prevent workflow decay.
Proceedings of the 5th international workshop on semantic web applications and tools for life sciences (SWAT4LS 2012)
CEUR Workshop Proceedings **952**:23
http://ceur-ws.org/Vol-952/paper_23.pdf

[Iborra 2024] Paula Iborra, José M. Fernández, Salvador Capella-Gutierrez (2024):
Onboarding Snakemake: Progress towards WfExS-backend integration.
F1000Research **13**(ELIXIR):551 (poster)
https://doi.org/10.7490/f1000research.1119725.1

[Jones 2023] Matthew B. Jones, Carl Boettiger, Abby Cabunoc Mayes, Arfon Smith, Morane Gruenpeter, Valentin Lorentz, Thomas Morrell, Daniel Garijo, Peter Slaughter, Kyle Niemeyer, Yolanda Gil, Martin Fenner, Krzysztof Nowak, Mark Hahnel, Luke Coy, Alice Allen, Mercè Crosas, Ashley Sands, Neil Chue Hong, Patricia Cruse, Daniel S. Katz, Carole Goble, Bryce Mecum, Alejandra Gonzalez-Beltran, Noam Ross (2023):
CodeMeta: an exchange schema for software metadata. Version 3.0.
CodeMeta project
https://w3id.org/codemeta/v3.0

[Khan 2019] [Khan 2019] Farah Zaib Khan, Stian Soiland-Reyes, Richard O. Sinnott, Andrew Lonie, Carole Goble, Michael R. Crusoe (2019):
Sharing interoperable workflow provenance: A review of best practices and their practical application in CWLProv.
GigaScience **8**(11)
https://doi.org/10.1093/gigascience/giz095

[Lamprecht 2020] Anna-Lena Lamprecht, Leyla Garcia, Mateusz Kuzak, Carlos Martinez, Ricardo Arcila, Eva Martin Del Pico, Victoria Dominguez Del Angel, Stephanie Van De Sandt, Jon Ison, Paula Andrea Martinez, Peter Mcquilton, Alfonso Valencia, Jennifer Harrow, Fotis Psomopoulos, Josep Ll. Gelpi, Neil Chue Hong, Carole Goble, Salvador Capella-Gutierrez (2020):
Towards FAIR principles for research software.
Data Science **3**(1) pp. 37–59.
https://doi.org/10.3233/DS-190026

[Leo 2024] Simone Leo, Michael R. Crusoe, Laura Rodríguez-Navas, Raül Sirvent, Alexander Kanitz, Paul De Geest, Rudolf Wittner, Luca Pireddu, Daniel Garijo, José M. Fernández, Iacopo Colonnelli, Matej Gallo, Tazro Ohta, Hirotaka Suetake, Salvador Capella-Gutierrez, Renske de Wit, Bruno de Paula Kinoshita, Stian Soiland-Reyes (2024):
Recording provenance of workflow runs with RO-Crate.
arXiv:2312.07852
PLOS One (accepted)
https://doi.org/10.48550/arXiv.2312.07852
https://doi.org/10.1371/journal.pone.0309210

[Los 2024] Laila Los (2024):
Workflows Workflows Workflows!
Galaxy Community Hub
https://galaxyproject.org/news/2024-04-26-workflows-workflows-workflows/

[López 2023] David López, Stian Soiland-Reyes (2023):
Exporting structured data like RO-Crate or BioComputeObjects.
Galaxy Community Hub
https://galaxyproject.org/news/2023-02-23-structured-data-exports-ro-bco/

[López 2024] David López (2024):
InvenioRDM integration in Galaxy.
Galaxy Community Hub
https://galaxyproject.org/news/2024-05-03-inveniordm-integration/

[Möller 2017] Steffen Möller, Stuart W. Prescott, Lars Wirzenius, Petter Reinholdtsen, Brad Chapman, Pjotr Prins, Stian Soiland-Reyes, Fabian Klötzl, Andrea Bagnacani, Matúš Kalaš, Andreas Tille, Michael R. Crusoe (2017):
Robust cross-platform workflows: How technical and scientific communities collaborate to develop, test and share best practices for data analysis.
Data Science and Engineering **2**
https://doi.org/10.1007/s41019-017-0050-4

[Nasr 2024] Engy Nasr, Bérénice Batut, Paul Zierep (2024):
allele-based-pathogen-identification/main. (Galaxy workflow)
WorkflowHub
https://doi.org/10.48546/workflowhub.workflow.1063.2

[Niehues 2024] Anna Niehues, Casper de Visser, Fiona A Hagenbeek, Purva Kulkarni, René Pool, Naama Karu, Alida S D Kindt, Gurnoor Singh, Robert R J M Vermeiren, Dorret I Boomsma, Jenny van Dongen, Peter A C 't Hoen, Alain J van Gool (2024):
A multi-omics data analysis workflow packaged as a FAIR Digital Object.
GigaScience **13**:giad115
https://doi.org/10.1093/gigascience/giad115

[Nyberg Åkerström 2024] Wolmar Nyberg Åkerström, Kurt Baumann, Oscar Corcho, Romain David, Yann Le Franc, Yann, Bénédicte Madon, Barbara Magagna, András Micsik, Marco Molinaro, Milan Ojsteršek, Silvio Peroni, Andrea Scharnhorst, Lars Vogt, Heinrich Widmann (2024):
Developing and implementing the semantic interoperability recommendations of the EOSC Interoperability Framework.
EOSC-A Semantic Interoperability Task Force
https://doi.org/10.5281/zenodo.10843882

[RFC 4112] Paul J. Leach, Rich Salz, Michael H. Mealling (2005):
A Universally Unique IDentifier (UUID) URN Namespace.
RFC Editor, RFC 4122
https://doi.org/10.17487/rfc4122

[Sanderson 2017] Robert Sanderson, Paolo Ciccarese, Benjamin Young (eds.) (2017):
Web Annotation Data Model.
W3C Recommendation 23 February 2017, Web Annotation Working Group
https://www.w3.org/TR/2017/REC-annotation-model-20170223/

[Schaaf 2023] Sebastian Schaaf, Anika Erxleben-Eggenhofer, Bjoern Gruening (2023):
Galaxy and RDM: Being More Than a Workflow Manager: Living the Data Life Cycle.
Proceedings of the Conference on Research Data Infrastructure **1**
https://doi.org/10.52825/cordi.v1i.421

[Smith 2016] Arfon M. Smith, Daniel S. Katz, Kyle E. Niemeyer, FORCE11 Software Citation Working Group (2016):
Software citation principles.
PeerJ Computer Science **2**:e86
https://doi.org/10.7717/peerj-cs.86

[Soiland-Reyes 2018] Stian Soiland-Reyes, Marcos Cáceres (2018):
The Archive and Package (arcp) URI scheme.
2018 IEEE 14th International Conference on e-Science (e-Science).
https://doi.org/10.1109/eScience.2018.00018

[Soiland-Reyes 2022a] Stian Soiland-Reyes, Peter Sefton, Leyla Jael Castro, Frederik Coppens, Daniel Garijo, Simone Leo, Marc Portier, Paul Groth (2022):
Creating lightweight FAIR Digital Objects with RO-Crate.
1st International Conference on FAIR Digital Objects (FDO 2022) (poster)
Research Ideas and Outcomes **8**:e93937
https://doi.org/10.3897/rio.8.e93937

[Soiland-Reyes 2022b] Stian Soiland-Reyes, Peter Sefton, Mercè Crosas, Leyla Jael Castro, Frederik Coppens, José M. Fernández, Daniel Garijo, Björn Grüning, Marco La Rosa, Simone Leo, Eoghan Ó Carragáin, Marc Portier, Ana Trisovic, RO-Crate Community, Paul Groth, Carole Goble (2022):
Packaging research artefacts with RO-Crate.
Data Science **5**(2)
https://doi.org/10.3233/DS-210053

[Soiland-Reyes 2022c] Stian Soiland-Reyes, Genís Bayarri, Pau Andrio, Robin Long, Douglas Lowe, Ania Niewielska, Adam Hospital, Paul Groth (2022):
Making Canonical Workflow Building Blocks interoperable across workflow languages.
Data Intelligence **4**(2)
https://doi.org/10.1162/dint_a_00135

[Soiland-Reyes 2024a] Stian Soiland-Reyes, Carole Goble, Paul Groth (2024):
Evaluating FAIR Digital Object and Linked Data as distributed object systems.
PeerJ Computer Science **10**:e1781
https://doi.org/10.7717/peerj-cs.1781

[Soiland-Reyes 2024b] Stian Soiland-Reyes, Peter Sefton, Simone Leo, Leyla Jael Castro, Claus Weiland, Herbert Van de Sompel (2024):
Practical webby FDOs with RO-Crate and FAIR Signposting: Experiences and lessons learned.
International FAIR Digital Objects Implementation Summit 2024, Berlin, Germany, 2024-03-20/–21.
Open Conference Proceedings **4** (submitted)
https://research.manchester.ac.uk/en/publications/48574fab-924b-4879-9c88-914552b1214f

[Soiland-Reyes 2024c] Stian Soiland-Reyes, Leyla Jael Castro, Rohitha Ravinder, Claus Weiland, Jonas Grieb, Alexander Rogers, Christophe Blanchi, Herbert Van de Sompel (2024):
BioHackEU23 report: Enabling FAIR Digital Objects with RO-Crate, Signposting and Bioschemas.
BioHackrXiv
https://doi.org/10.37044/osf.io/gmk2h

[Soiland-Reyes 2024d] Stian Soiland-Reyes, Bruno P. Kinoshita, Vincent Emonet (2024):
Signposting link parser library.
https://doi.org/10.5281/zenodo.10471965

[Soiland-Reyes 2024e] Stian Soiland-Reyes, Björn Grüning, Paul De Geest (2024):
EuroScienceGateway MS3: Initial EuroScienceGateway workflows registered. (Milestone)
Zenodo
https://doi.org/10.5281/zenodo.1072892

[Strawn 2024] G. Strawn, P. Wittenburg, D. Broeder, Christophe Blanchi (eds.) (2024):
Data FDO: Simplified Requirements.
FAIR Digital Objects Forum (internal draft)
https://docs.google.com/document/d/1O_Q3Qa9rliiI3mX8paYCSEUXIpcgJlYO_-TXddHEol8/edit

[Van de Sompel 2015] Herbert Van de Sompel, Michael L. Nelson (2015):
Reminiscing About 15 Years of Interoperability Efforts.
D-Lib Magazine **21**(11/12)
https://doi.org/10.1045/november2015-vandesompel

[Van de Sompel 2023] Herbert Van de Sompel, Martin Klein, Shawn Jones, Michael L. Nelson, Simeon Warner, Anusuriya Devaraju, Robert Huber, Wilko Steinhoff, Vyacheslav Tykhonov, Luc Boruta, Enno Meijers, Stian Soiland-Reyes, Mark Wilkinson (2023):
FAIR Signposting Profile. (version 20231002)
https://signposting.org/FAIR/

[Varshney 2024] Deepti Varshney, Saskia Hiltemann, Stefan A. Rensing, Romy Petroll, Björn A. Grüning (2024):
MAdLand Computational resources through Galaxy.
Galaxy community conference (GCC) 2024
F1000Research **13**:776 (slides)
https://doi.org/10.7490/f1000research.1119789.1

[de Visser 2024] Casper de Visser, Anna Niehues (2022):
X-omics ACTIONdemonstrator analysis workflow.
WorkflowHub
https://doi.org/10.48546/workflowhub.workflow.402.8

[Wilkinson 2022a] Mark D. Wilkinson, Susanna-Assunta Sansone, Grootveld Marjan, Josefine Nordling, Richard Dennis, David Hecker (2022):
FAIR Assessment Tools: Towards an "Apples to Apples" Comparisons.
EOSC FAIR Metrics and Data Quality Task Force
EOSC Association / Zenodo
https://doi.org/10.5281/zenodo.7463421

[Wilkinson 2024a] Mark D Wilkinson, Susanna-Assunta Sansone, Marjan Grootveld, Richard Dennis, David Hecker, Robert Huber, Stian Soiland-Reyes, Herbert Van de Sompel, Andreas Czerniak, Milo Thurston, Allyson L. Lister, Alban Gaignard (2024):
Report on "FAIR Signposting" and its uptake by the community.
EOSC FAIR Metrics and Data Quality Task Force
https://doi.org/10.5281/zenodo.10490289

[Wilkinson 2024b] Sean Wilkinson, Meznah Aloqalaa, Michael Crusoe, Luiz Gadelha, Daniel Garijo, Carole Goble, Johan Gustafsson, Nick Juty, Simone Leo, Sehrish Kanwal, Farah Zaib Khan, Bruno Kinoshita, Johannes Koster, Karsten Peters-von Gehlen, Line Pouchard, Randy Rannow, Stian Soiland-Reyes, Ziheng Sun, Baiba Vilne, Merridee Wouters, Denis Yuen, Sarah Wait Zaranek, Mahnoor Zulfiqar et al. (2024):
The FAIR Principles for Computational Workflows.
(in preparation)

[Wittenburg 2022] Peter Wittenburg, Ivonne Anders, Christophe Blanchi, Merret Buurman, Carole Goble, Jonas Grieb, Alex Hardisty, Sharif Islam, Thomas Jejkal, Tibor Kálmán, Christine Kirkpatrick, Laurence Lannom, Thomas Lauer, Giridhar Manepalli, Karsten Peters-von Gehlen, Andreas Pfeil, Robert Quick, Mark van de Sanden, Ulrich Schwardmann, Stian Soiland-Reyes, Rainer Stotzka, Zachary Trautt, Dieter Van Uytvanck, Claus Weiland, Philipp Wieder (2022):
FAIR Digital Object Demonstrators 2021.
Report,
FAIR Digital Objects Forum / Zenodo
https://doi.org/10.5281/zenodo.5872645

[WRROC 2024a] Workflow Run RO-Crate working group (2024):
Process Run Crate specification. Version 0.5.
Zenodo
https://w3id.org/ro/wfrun/process/0.5
https://doi.org/10.5281/zenodo.12158562

[WRROC 2024b] Workflow Run RO-Crate working group (2024):
Workflow Run Crate specification. Version 0.5.
Zenodo
https://w3id.org/ro/wfrun/workflow/0.5
https://doi.org/10.5281/zenodo.12159311

[WRROC 2024c] Workflow Run RO-Crate working group (2024):
Provenance Run Crate specification. Version 0.5.
Zenodo
https://w3id.org/ro/wfrun/provenance/0.5
https://doi.org/10.5281/zenodo.12160782