



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



MS6 Integrated EuroScienceGateway knowledge graph

Work Package 2

Technical References

<i>Project Acronym</i>	<i>ESG</i>
<i>Project Title</i>	<i>EuroScienceGateway</i>
<i>Project Coordinator</i>	<i>Albert-Ludwigs-University Freiburg</i>
<i>Project Duration</i>	<i>September 2022 - August 2025</i>

<i>Document</i>	<i>MS6 Integrated EuroScienceGateway knowledge graph</i>
<i>Work Package</i>	<i>Work Package 2</i>
<i>Task</i>	<i>T 2.3 and T 2.4</i>
<i>Dissemination Level*</i>	<i>PU</i>
<i>Lead Beneficiary</i>	<i>UNIMAN</i>
<i>Contributing Beneficiaries</i>	<i>VIB, EPFL, BSC, UP, EGI, UiO, UNIMAN</i>
<i>Due Date of Milestone</i>	<i>31st August 2025</i>



Actual Submission Date	2025-08-29
------------------------	------------

* PU = Public

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

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

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

Version	Date	Beneficiaries	Author
1.0	2025-08-29	UNIMAN, VIB, EPFL, BSC, UP, EGI, UiO	Eli Chadwick, Oliver Woolland, Volodymyr Savchenko, Finn Bacall, Alexander Hambley, José María Fernández, Armin Dadras, Stian Soiland-Reyes



Funded by
the European Union

Acknowledgements

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

The authors acknowledge Qiwen Wu and Carole Goble, both from The University of Manchester, for helpful discussions and user feedback on the knowledge graph.

Cite as

Eli Chadwick, Oliver Woolland, Volodymyr Savchenko, Finn Bacall, Alexander Hambley, José María Fernández, Armin Dadras, Stian Soiland-Reyes (2025):

EuroScienceGateway MS6: Integrated EuroScienceGateway knowledge graph

Zenodo

<https://doi.org/10.5281/zenodo.16992674>

Disclaimer

This work may rely on data from sources external to the members of the ESG project Consortium. Members of the Consortium do not accept liability for loss or damage suffered by any third party as a result of errors or inaccuracies in such data. The information in this document is provided “as is” and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and neither the European Community nor any member of the ESG Consortium is liable for any use that may be made of the information.

© Members of the ESG Consortium



**Funded by
the European Union**

Executive Summary

The Workflowhub Knowledge Graph has been improved and its generation made more robust.

When this work was last reported, a complete knowledge graph had been generated but several criticisms were made. The previous graph was:

- Verbose and hard for a human to read or navigate
- Had unresolvable URIs as root data entities
- Contained many duplicate entries
- Contained sparse metadata from only a single source

Work has successfully been undertaken to address all of these points. The graph now uses partially resolvable, more human readable, URIs for root data entities. Steps have been added to the generation software to add metadata from additional sources (enrichment) and to remove duplicate entries (consolidation).

Several areas of the codebase have been refactored and improved, to help ensure repeatability and longevity.

The new knowledge graph still has areas that could be improved. Partially resolvable URIs should be migrated to fully resolvable alternatives. Further enrichment processes should be added which affords greater de-duplication.



Table of contents

Acknowledgements	3
Cite as	3
Disclaimer	3
Executive Summary	4
Introduction	6
Data Sources	6
Handling relative paths in WorkflowHub's RO-Crate	6
Workflow for building workflow graph	7
1. Fetch RO-Crates	8
2. Combine RO-Crates	8
3. Enrich with additional metadata	9
Workflow Languages	9
People & ORCID identifiers	9
4. Consolidate and de-duplicate metadata	9
5. Capture provenance and publish	10
Visualisation	10
Future Work	14
References	14



Introduction

This report describes the workflow that is used to create the Integrated knowledge graph (MS6) containing metadata from all workflows registered on WorkflowHub. This augments the previous Initial Knowledge Graph [[Hambley 2024](#)] (MS5).

The knowledge graph and the workflow to generate it are archived on Zenodo [[Chadwick 2025](#)]. The source code, including a configuration for visualizing the knowledge graph, is publicly available on GitHub: <https://github.com/workflowhub-eu/workflowhub-graph> [[swh:1:rev:e25e18a764da61437e5ad687876457928b792f9d](#)].

Data Sources

The primary data source is Workflow RO-Crates downloaded from WorkflowHub. Workflow RO-Crate [[Bacall 2022](#)] is a profile of RO-Crate for describing workflows, including workflow-specific metadata such as input and output parameters. All workflows registered with WorkflowHub can be exported as Workflow RO-Crates. The exported metadata may be derived from an imported RO-Crate, an imported Git repository, or information provided manually by the user that registered the workflow.

Each workflow's RO-Crate takes the form of a ZIP file containing the workflow files and an [RO-Crate Metadata Document](#), which is an RDF document in [JSON-LD](#) format. These metadata files are combined together to form the knowledge graph, following the method detailed below.

The knowledge graph enrichments pull in data from additional sources. At the time of writing, those sources are Wikidata and ORCID, though more could be added in future. These enrichments can provide canonical sources of information that individual workflows can be linked to, such as a Wikidata entry for a workflow language.

Handling relative paths in WorkflowHub's RO-Crate

RO-Crate metadata may use relative paths to identify files within the crate. However, when many RO-Crates are merged into a single knowledge graph, there may be duplicate relative paths coming from different crates, and those paths are no longer resolvable as the metadata has been abstracted away from the data itself. To alleviate this, a unique identifier is determined for each RO-Crate, and that identifier is prepended to all relative paths within the crate. We call this the “base URI” as it corresponds to `@base` in JSON-LD.



In the first version of the knowledge graph, each RO-Crate's base URI was an ARCP URI derived from the URL of the workflow on WorkflowHub. ARCP is designed for this purpose of creating URIs for files within a ZIP archive [[Soiland-Reyes & Cáceres 2018](#)], so it seemed a natural choice. However, in the current version of the knowledge graph, we switched to using the download link for the RO-Crate as the base URI, followed by a trailing slash (e.g. https://workflowhub.eu/workflows/1698/ro_crate/). This was done for multiple reasons:

- University of Manchester researchers using the knowledge graph found the ARCP URIs difficult to understand, as they obfuscate the identifiers of the workflow and are not resolvable
- The same users' feedback also noted that metadata fields such as `distribution` and `contentURL` are often missing from the RO-Crates downloaded from WorkflowHub, meaning individual files cannot be retrieved using the metadata in the knowledge graph. This exacerbates the difficulty of fetching data based on ARCP base URIs.
- The base URI is now directly resolvable. We also believed, based on some initial testing, that relative file paths within crates would also be semi-resolvable with this base URI (resolving to the crate download, but not the file itself). However, this turned out not to be true in practice.

If improvements are made to the metadata provided by WorkflowHub, then it may be more practical to return to using ARCP URIs in future. A [GitHub issue](#) has been raised to request these improvements in FAIRDOM-SEEK (the platform on which WorkflowHub is based).

Workflow for building workflow graph

The development of the first version of the graph, which was published in August 2024, is described in the earlier EuroScienceGateway deliverable D2.1 [[Soiland-Reyes 2024](#)].

Since then we made some changes to how the graph is constructed.



WorkflowHub Knowledge Graph

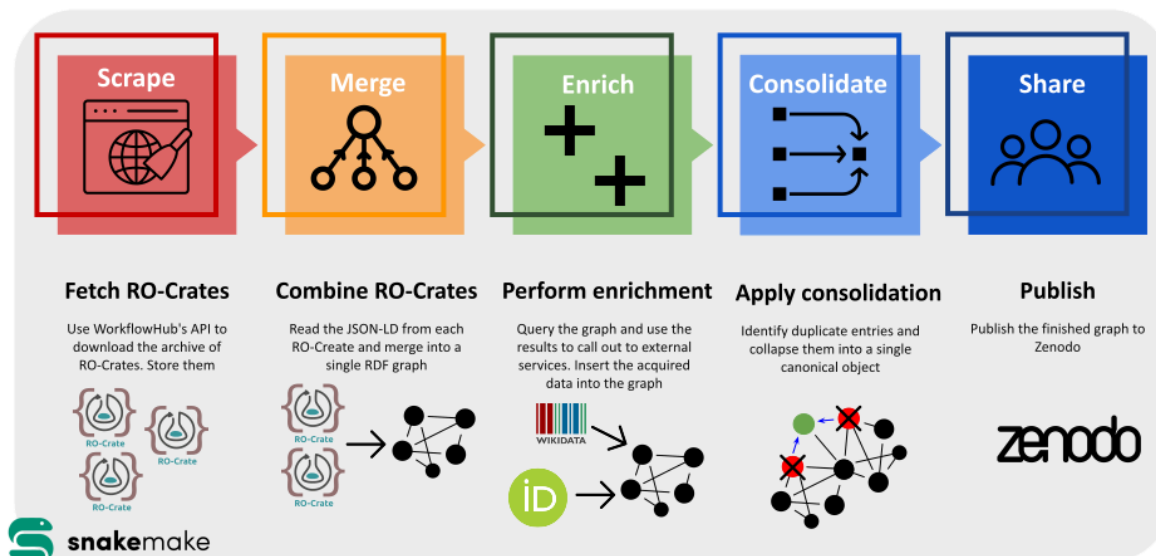


Figure 1: A visual summary of the Snakemake workflow steps.

The knowledge graph is built using a Snakemake workflow with the following steps:

1. Fetch RO-Crates
2. Combine RO-Crates
3. Enrich with additional metadata
4. Consolidate and de-duplicate metadata
5. Capture provenance and publish

1. Fetch RO-Crates

Retrieve the list of known workflows in WorkflowHub, then use WorkflowHub's API to download the RO-Crate metadata file for each workflow and store them. There is also an option to download the full ZIP archive for each workflow and extract the RO-Crate metadata from there, but this is avoided by default to minimize storage use.

2. Combine RO-Crates

Merge JSON-LD files from each RO-Crate into a single RDF graph, and save in RDF turtle format.



Funded by
the European Union

3. *Enrich with additional metadata*

Enrich the metadata to increase its usefulness, incorporating metadata from sources like Wikidata and ORCID.

Enrichments are included in a modular fashion. There is an abstract base class to support the development of new enrichments. The enrichments included in the current version of the graph are:

Workflow Languages

By convention, WorkflowHub RO-Crates describe their workflow language using entities with URIs provided by the Workflow RO-Crate profile. For example, <https://w3id.org/workflowhub/workflow-ro-crate#cw/> will be used to represent the Common Workflow Language. These URIs resolve to relevant sections in the Workflow RO-Crate profile description. However, some RO-Crates use local entities (entities which are not identified by a URI) to describe the workflow language instead.

Canonical entities are added for workflow languages based on Wikidata entries. Entities which have local identifiers within an RO-Crate but appear to match one of these canonical entities have a `sameAs` relationship added to connect the two.

People & ORCID identifiers

Where a Person entity in the graph includes an `identifier` property with an ORCID, additional information is retrieved from the ORCID API about that person, including family names, given names, and institutions. This metadata is added to the knowledge graph.

Some Person entities in the graph are not identified with any PID, only their name. In some cases there are many duplicate entities referring to the same person, and sometimes at least one duplicate has an ORCID associated. A future enrichment will aim to connect all duplicate entities to the same ORCID. This may pose challenges at scale, as of course there is no guarantee that two entities which have the same name definitely refer to the same person. However, the benefits of de-duplicating currently outweigh this risk for the data on WorkflowHub.

4. *Consolidate and de-duplicate metadata*

After all enrichments are completed, there is a final stage of consolidation and de-duplication. Entities which are marked as `sameAs` each other are combined into a single entity, and cross-references are updated to use that combined entity. This



effectively de-duplicates entities that appear in multiple crates, even if they did not use the same identifier originally (most common where local identifiers have been used instead of PIDs).

In future a quality assurance stage could be added here, including the generation of some statistics about the graph.

5. Capture provenance and publish

Generate a Workflow Run RO-Crate to capture the configuration used, the date of graph generation, and so on. Both the knowledge graph and the workflow source code are included in the RO-Crate, as well as intermediate files representing the base graph, the metadata added with each enrichment, and the merged graph prior to consolidation.

When the workflow is run through GitHub Actions, the generated RO-Crate is preserved as an artifact. This artifact can then be uploaded to Zenodo manually, using the RO-Crate InvenioRDM package to automatically populate much of the Zenodo metadata through the API. In future this upload stage can be automated so that the graph can be regenerated on a regular basis.

Visualisation

A Docker configuration for visualizing the knowledge graph is included in the source code. The stack includes an [Apache Jena Fuseki](#) SPARQL server, a [Zazuko Trifid](#) server which provides endpoint exploration tooling, and a [Zazuko Blueprint](#) frontend.

The Blueprint frontend allows exploration of the knowledge graph in a visual interface. The user can search for specific entities, explore entities by type (e.g. workflow languages), and view relationships between entities in a list or graph layout. Some example screenshots are included below.



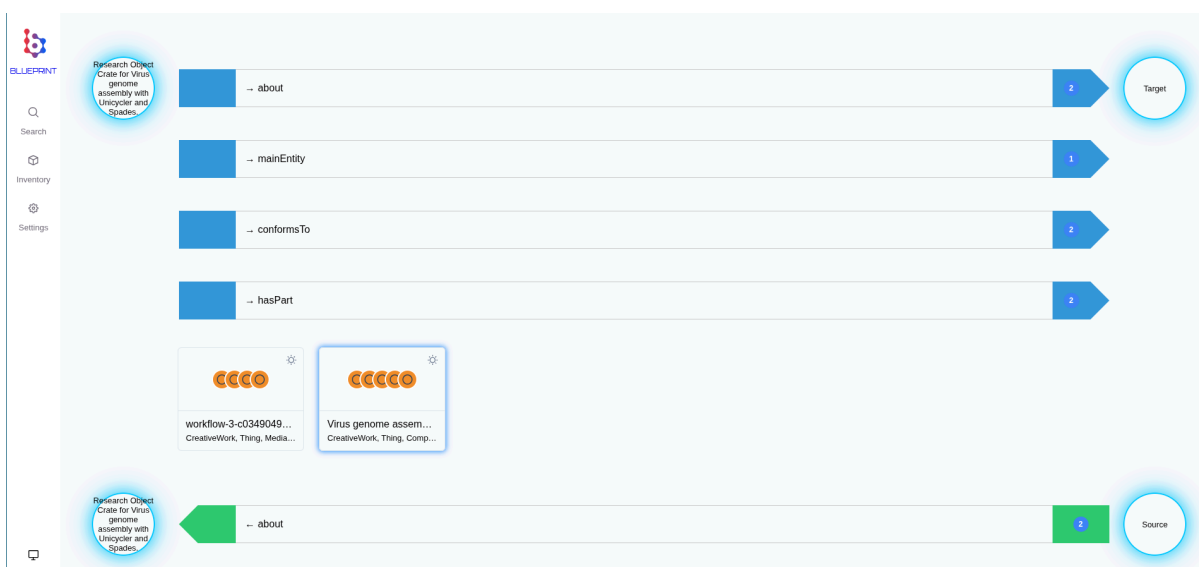


Figure 2: The relationships for an RO-Crate root dataset. Both forward and reverse relationships are included, and each relationship can be expanded to show the entities involved.

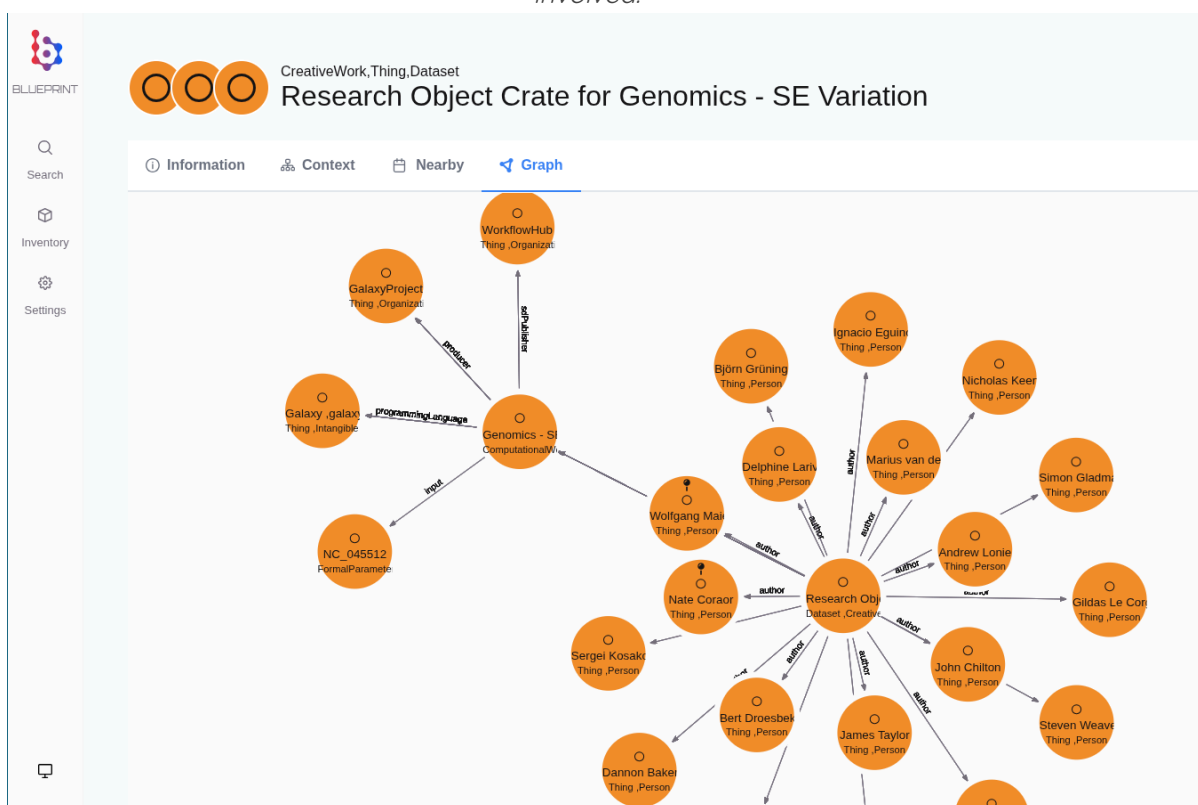


Figure 3: A graph view of the relationships within a RO-Crate, an alternative to Fig. 2. In this case, the root dataset is connected to many authors, plus the workflow file. The workflow file is in turn connected to its programming language, its inputs, and so on.

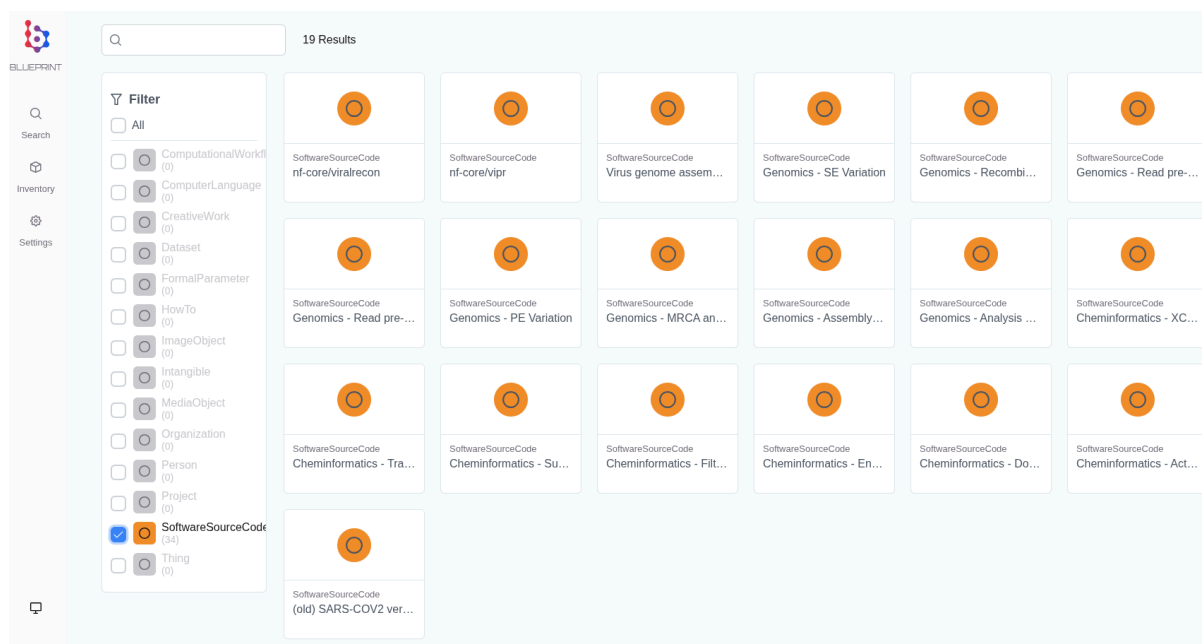







Figure 4. A list of all “SoftwareSourceCode” entities in the graph, which represent workflow files.



 Search
 Inventory
 Settings



Person, Thing
Ivan Topolsky

Information
Context
Relations
Graph

label

Ivan Topolsky

affiliation

University of Geneva Institute of Genetics and Genomics of Geneva
D-BSSE - ETHZ/Basel
Swiss Institute of Bioinformatics
Hôpitaux Universitaires de Genève
Geneva Bioinformatics SA
University of Geneva

familyName

Topolsky

givenName

Ivan

identifier

<https://orcid.org/0000-0002-7561-0810>

memberOf

University of Geneva Institute of Genetics and Genomics of Geneva
D-BSSE - ETHZ/Basel
Swiss Institute of Bioinformatics
Hôpitaux Universitaires de Genève
Geneva Bioinformatics SA
University of Geneva

Figure 5. The data page for an individual entity in the graph. In this case, additional information about a person is included from the ORCID enrichment.



Future Work

Improvements to the quality of metadata produced by WorkflowHub will improve the quality of the knowledge graph. The development work and user feedback helped to uncover some quality issues and desired features, such as the need to include the `contentUrl` property to help users to find the download link for each file in the graph.

Some workflows also do not fully conform to the Bioschemas ComputationalWorkflow profile, as their input and output parameters are not always defined. This limits the usefulness of the metadata, and it is a result of users not providing that information at the time of workflow registration. It may be possible to improve this with features in WorkflowHub or improvements in training to encourage users to add this metadata.

The WorkflowHub knowledge graph work could be combined with other knowledge graph work currently being developed in the [BIOINDUSTRY 4.0 project](#) (Horizon Europe project 101094287). This uses the triple store already integrated and embedded in FAIRDOM-SEEK. WorkflowHub could generate fresh RO-Crates and send them to the triple store when metadata is updated.

Other outstanding issues for improving the knowledge graph can be found in the GitHub repository: <https://github.com/workflowhub-eu/workflowhub-graph/issues>

References

[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>

[Chadwick 2025] Eli Chadwick, Oliver Woolland, Alexander Hambley, Volodymyr Savchenko, and Stian Soiland-Reyes (2025):

Workflowhub Knowledge Graph

Zenodo

<https://doi.org/10.5281/zenodo.16995374>.

[Hambley 2024] Alexander Hambley, Eli Chadwick, Oliver Woolland, Stian Soiland-Reyes, Volodymyr Savchenko (2024):

WorkflowHub Knowledge Graph. (Dataset)

Zenodo

<https://doi.org/10.5281/zenodo.13362051>

[Soiland-Reyes & Cáceres 2018] Stian Soiland-Reyes, Marcos Cáceres (2018):



**Funded by
the European Union**

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> arXiv:1809.06935

[Soiland-Reyes 2024] Stian Soiland-Reyes, Eli Chadwick, Finn Bacall, José M. Fernández, Björn Grüning, and Hakan Bayındır (2024):

Eurosciencegateway D2.1: Reproducible FAIR Digital Objects for Workflows

Zenodo

<https://doi.org/10.5281/zenodo.13225792>



**Funded by
the European Union**