WorkflowHub knowledge graph task force

Feel free to help take minutes during the calls!
Minutes are in reverse chronological order (newest first)

2025 Schedule: Every 2 weeks on Wednesday 14:00 BST / 15:00 CEST

2025-08-27 #10

- **When**: 2025-08-27 14:00 BST / 15:00 CEST
- **Where**: https://zoom.us/j/91204435521?pwd=te2dSHwiMeYvXL7bf3D0bTXaHSQ0TM.1
- **Who:** Eli

Issues: https://github.com/workflowhub-eu/workflowhub-graph/issues

Notes

Code

PRs which need finishing ASAP:

- RO-Crate creation https://github.com/workflowhub-eu/workflowhub-graph/pull/70 (might need tweaks after other PRs)
- Replace arcp and add visualisation: https://github.com/workflowhub-eu/workflowhub-graph/pull/72 - almost done
- ORCID enrichment https://github.com/workflowhub-eu/workflowhub-graph/pull/71 - Oliver thinks it looks good and will test it on top of #72

Report

M 2.4 Integrated EuroScienceGateway knowledge graph

Oliver - screenshots of visualization, example queries?

2025-08-13 #9

- **When**: 2025-08-13 14:00 BST / 15:00 CEST
- **Where**: https://zoom.us/j/91204435521?pwd=te2dSHwiMeYvXL7bf3D0bTXaHSQ0TM.1
- **Who:** Eli, Volodymyr

Issues: https://github.com/workflowhub-eu/workflowhub-graph/issues

Notes

PR for RO-Crate creation: https://github.com/workflowhub-eu/workflowhub-graph/pull/70

Targeting data cleaning subissues: https://github.com/workflowhub-eu/workflowhub-graph/issues/48

- Not so much the ones about querying the graph - more the ones about cleaning up/de-duplicating the data.

Volodymyr to review enrichment scaffolding and try to do one or two cleanup tasks next week

Eli to get Oliver onto Matrix for discussion.

2025-07-30

Cancelled

2025-06-16 #8

- **When**: 2025-06-02 14:00 BST / 15:00 CEST
- **Where**: https://zoom.us/j/91204435521?pwd=te2dSHwiMeYvXL7bf3D0bTXaHSQ0TM.1
- **Who:** Alex, Eli, Oliver, Stian

Issues: https://github.com/workflowhub-eu/workflowhub-graph/issues

Notes

Enrichments:

- Example enrichment to connect all workflow languages to their wikidata entry - so all entities representing the same language connect to a single node
- Incorporate more WorkflowHub metadata - see https://github.com/workflowhub-eu/workflowhub-graph/issues/50 , maybe https://github.com/workflowhub-eu/workflowhub-graph/issues/49

New GH Actions workflow to build & publish the knowledge graph

To publish the milestone:

- Add back in the RO-Crate generation step
- Create the RO-Crate as an artifact in the GH Actions
- Upload the artifact to Zenodo (possibly using https://github.com/ResearchObject/ro-crate-inveniordm)
    - Could be automated or manual
    - The metadata should be manually checked before publication because the package is missing a couple of bits of metadata

2025-06-02 #7

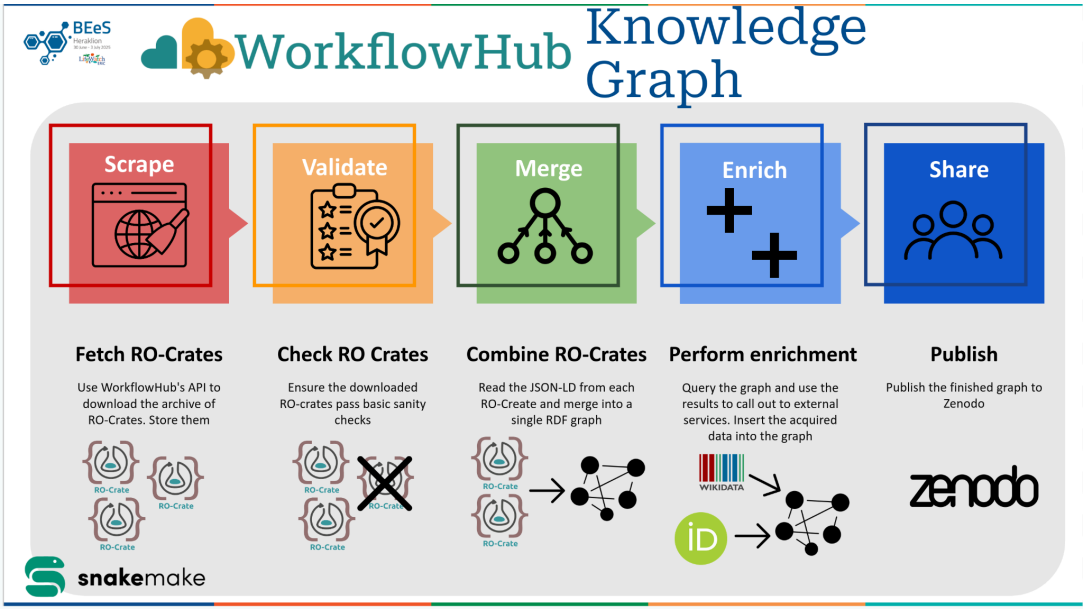
- **When**: 2025-06-02 14:00 BST / 15:00 CEST
- **Where**: ~~https://zoom.us/j/91204435521?pwd=te2dSHwiMeYvXL7bf3D0bTXaHSQ0TM.1~~
  - **https://zoom.us/j/91364934432 (changed)**
- **Who:** Stian, Oliver
- Apologies: Eli (leave)

Issues: https://github.com/workflowhub-eu/workflowhub-graph/issues

Attending

Notes

Status of enrichment support in knowledge graph workflow [Oliver]



Added to https://github.com/workflowhub-eu/workflowhub-graph/tree/develop/workflowhub_graph/workflowhub_graph/enrichment_strategies called by https://github.com/workflowhub-eu/workflowhub-graph/blob/develop/Snakefile#L57

Possible enrichments:

- https://graph.openaire.eu/ e.g. publications, datasets, EU projects, authors

- ORCID Linked Data https://info.orcid.org/ufaqs/does-orcid-support-schema-org-linked-open-data-and-json-ld/
    - Can add list of publications from JSON-LD output. But you don't know if any of these are related to the workflow or not.
- https://orkg.org/
- http://bio.tools/ e.g. using API https://biotools.readthedocs.io/en/latest/api_reference.html or better via https://research-software-ecosystem.github.io/ content in https://github.com/research-software-ecosystem/content
    - For instance https://github.com/research-software-ecosystem/content/blob/master/data/2d-page/2d-page.bioschemas.jsonld
    - Our KG should replace https://github.com/research-software-ecosystem/content/blob/master/datasets/workflow_hub.ttl
- Galaxy Codex https://github.com/galaxyproject/galaxy_codex/tree/main
- EOSC graph / https://datacite.org/blog/introducing-the-pid-graph/
- https://europepmc.org/search?query=workflowhub.eu shows mentions of "workflowhub.eu" https://europepmc.org/article/PPR/PPR976394 https://workflowhub.eu/workflows/1103?version=3
- EDAM ontology
- Classifications

Can do one named graph per enrichment, as each of them come with a level of

→ Oliver will

2025-05-21 #6

- **When**: 2025-05-21 14:00 BST / 15:00 CEST
- **Where**: https://zoom.us/j/91204435521?pwd=te2dSHwiMeYvXL7bf3D0bTXaHSQ0TM.1

Issues: https://github.com/workflowhub-eu/workflowhub-graph/issues

Attending

- Eli Chadwick (UNIMAN)
-

Notes

No progress by Eli (as expected)

From Oliver on Teams:

- I tried out what we already have

- I got build errors which I fixed

- I get runtime errors which I *kinda* fixed but could be fragile

- started working on the workflow, still getting used to Snakemake

2025-05-07 #5

- **When**: 2025-05-07 14:00 BST / 15:00 CEST
- **Where**: https://zoom.us/j/91204435521?pwd=te2dSHwiMeYvXL7bf3D0bTXaHSQ0TM.1

Issues: https://github.com/workflowhub-eu/workflowhub-graph/issues

Invited/Attending

- Oliver Woolland (UNIMAN)
- Eli Chadwick (UNIMAN)
- Volodymyr Savchenko (EPFL)
- Stian Soiland-Reyes (UNIMAN)
- Armin Dadras (ALU-FR)

Notes

- Enrich the knowledge graph with e.g. ORCID data
- Productionise the graph - e.g. update every week
- Consolidate / de-duplicate entities such as authors, workflow systems
- Classification of WFH teams into domains - there's a spreadsheet somewhere…
- Look through suggestions in D2.1
  D2.1 Reproducible FAIR Digital Objects for workflows.docx
    - Generating issues on this thread – https://github.com/workflowhub-eu/workflowhub-graph/issues/48
- Visualisation options tried at UNIMAN
    - https://github.com/SemanticComputing/sampo-ui?tab=readme-ov-file
    - https://mappingmanuscriptmigrations.org/ (instance of Sampo-UI)
    - https://github.com/zazuko/blueprint?tab=readme-ov-file#introduction
- GTN workflow search: https://training.galaxyproject.org/training-material/news/2023/11/20/workflow-search.html
- M2.4 Integrated EuroScienceGateway knowledge graph - WP2 M36 Zenodo data deposit; Report of integrations and queries
- Oliver to look at how to fit enrichment into existing pipeline as scaffolding. With pytest.
- Oliver: GitHub Actions workflow for running the knowledge graph pipeline.
- Volodymyr: Expand the use cases. Each use case is an idealised query. The encrichments then will add things to the workflows to help one or more use cases. Make test cases.

- Eli: Limited time until July, can pick up some individual consolidation tasks during June
- Use case: Use micropublications from paper and look up corresponding/relevant workflows (may need LLM or wikidata enrichment)
-

2024-06-28 #4

- **When**: 2024-06-28 16:00 BST / 17:00 CEST
- **Where**: https://zoom.us/j/96948642698

Issues: https://github.com/workflowhub-eu/workflowhub-graph/issues

Invited/Attending

- Stian Soiland-Reyes
- Volodymyr Savchenko (EPFL)
- Denys Savchenko (UP):
- Stuart Owen (UNIMAN)
- Paul De Geest (VIB):
- **Oliver Woolland (UNIMAN)**
- **Alex Hambley (UNIMAN)**
- **José Mª Fernández (BSC)**
- **Finn Bacall (UNIMAN)**
- Nick Juty (UNIMAN)
- **Doug Lowe (UNIMAN)**
- Eli Chadwick (UNIMAN)
- Francois Antoine Morier-Genoud

Repository: https://github.com/workflowhub-eu/workflowhub-graph
See branch https://github.com/workflowhub-eu/workflowhub-graph/tree/feature-source-ro-crates

Tasks: https://github.com/workflowhub-eu/workflowhub-graph/issues

Notes

- Sample graph (limited to 10 workflows): https://github.com/workflowhub-eu/workflowhub-graph/blob/develop/merged.ttl
- Still working on workflow
- Graphs from RDF: https://sketch.zazuko.com/
- https://monarch-initiative.github.io/ontogpt/

2024-05-24 #3

- **When**: 2024-05-24 16:00 BST / 17:00 CEST
- **Where**: https://zoom.us/j/96948642698

Issues: https://github.com/workflowhub-eu/workflowhub-graph/issues

Invited/Attending

- **Stian Soiland-Reyes**
- **Volodymyr Savchenko (EPFL)**
- Denys Savchenko (UP):
- Stuart Owen (UNIMAN)
- Paul De Geest (VIB):
- **Oliver Woolland (UNIMAN)**
- **Alex Hambley (UNIMAN)**
- José Mª Fernández (BSC)
- Finn Bacall
- Nick Juty (UNIMAN) (unavailable)
- Francois Antoine Morier-Genoud

Repository: https://github.com/workflowhub-eu/workflowhub-graph
See branch https://github.com/workflowhub-eu/workflowhub-graph/tree/feature-source-ro-crates

Tasks: https://github.com/workflowhub-eu/workflowhub-graph/issues

Notes

Alex: Versioning of RO-Crates in WorkflowHub. Make it work without too many changes. Avoid reiterating old crates. On Tuesday, refactor the code to make more of a module. Getting version through the identifier.

Volodymyr: Needed a more consistent way to find the version.

A: Need to find the version. There is a "version" text field in the RO-Crate, but need to check extensively.

V: The version in RO-Crate could be inconsistent. We need the one in the URL.

S: May need a second call to WFHub to ask what is the current version.

V: Will try to add a new API call to

A: Could not get the test working on the absolute URI branch.

V: Only works while in the action! But should also work locally. Check versions. Check the GitHub Action .github/workflows

A bit many branches! No code on main yet.
https://github.com/workflowhub-eu/workflowhub-graph/branches/active

→ A: Will merge into main.

Can also request the ro-crate-metadata with ?version=2 etc. e.g.
https://dev.workflowhub.eu/workflows/1048/ro_crate_metadata?version=1

Could be lots to do in validation / verification

“name” come in different shapes as well.

Check @id is valid

→ V: Start a “validation.md” file to note things to check and weird things that have been found in the wild.

Patching rdflib to work offline. Urllib retrieves the JSON-LD context. This can be modified for testing purposes.

Adding older versions asan optional graph output.

2024-05-17 #2

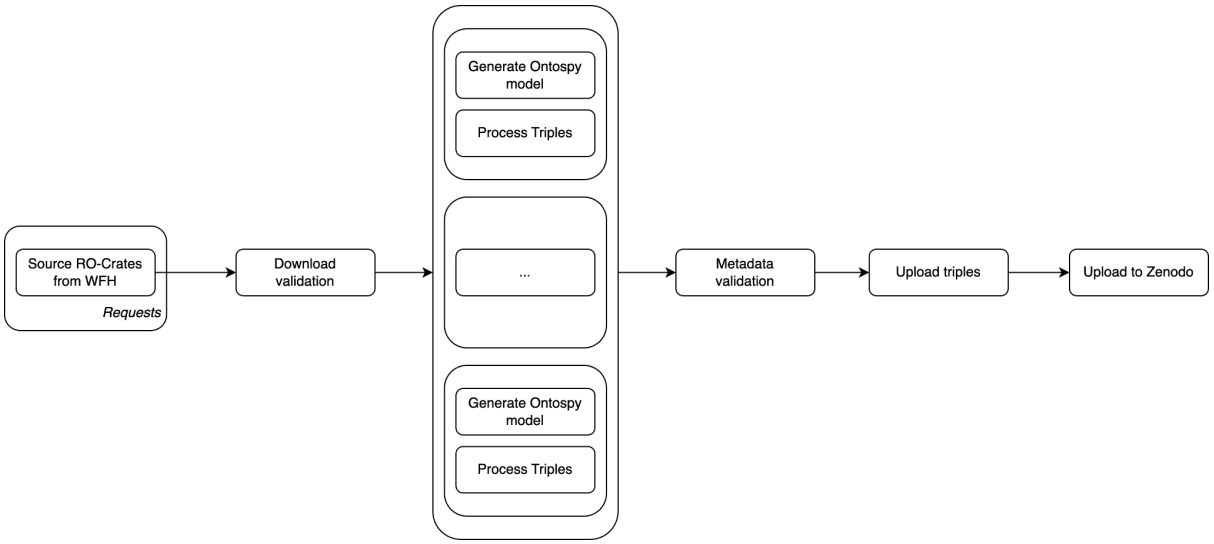
- **When**: Fri 2025-05-17 16:00 BST / 17:00 CEST
- **Where**: https://zoom.us/j/96948642698

Invited/Attending

- Stian Soiland-Reyes
- Volodymyr Savchenko (EPFL)
- Denys Savchenko (UP):
- Stuart Owen (UNIMAN)
- Paul De Geest (VIB):
- Oliver Woolland (UNIMAN)
- Alex Hambley (UNIMAN)
- Oliver Woolland (UNIMAN)
- José Mª Fernández (BSC)
- Nick Juty (UNIMAN) (unavailable)
-

Repository: https://github.com/workflowhub-eu/workflowhub-graph
See branch https://github.com/workflowhub-eu/workflowhub-graph/tree/feature-source-ro-crates

Notes



Alexander sketched out above and looked more on RO-Crate sourcing.

Finn started an endpoint that gives RO-Crate as single JSON-LD instance e.g.

https://dev.workflowhub.eu/workflows/1048/ro_crate_metadata – this has been pre-parsed so is valid JSON-LD.
https://dev.workflowhub.eu/workflows.json

Also looked at how to get only the most recent workflows using a filter.

https://dev.workflowhub.eu/workflows.json?filter[updated_at]=P1M

Workflow could for instance run every month on the recent workflows. This is on the dev instance but will be made live on public instance “soon”.

Alexander adding flag for dev vs production

Oliver: Looking at Ontospy. Some notes on pyld ??

More wider: On Question & Answer LLM connection to SPARQL. Stian to find BioCompute Object example.

JM SPARQL queries from WfExs source code? They are split because of different parts in the RO-Crate.
→ JM could add them to https://github.com/workflowhub-eu/workflowhub-graph as sparql files?

JM: Could WorkflowHub recognize from CWL to RO-Crate to match with similar steps for instance. Currently the structure of workflow is only available after execution or within CWL or other workflow languages, not in RO-Crate kept in WFHub.

In Galaxy you can get tool ID and you need to look up in Galaxy instance further information. WFHub can do this now, but only if you ask the right instance that knows that particular tool. Only then look up from usegalaxy.eu at the moment.

Paul: Can you get it from the toolshed directly?

WorkflowHub & Galaxy

WorkflowHub currently queries: https://usegalaxy.eu/api/tools to get set of tools from which to look up bio.tools IDs from Galaxy tool IDs

https://github.com/seek4science/seek/tree/main/lib/galaxy

Can this be synchronized with Zenodo for instance. May be some citation information but textual only.

How to do absolute URIs
https://www.researchobject.org/ro-crate/1.2-DRAFT/appendix/relative-uris.html for instance with @context and @base but when does it change over time?

Volodymyr: Wanted to help with the gathering.

How to move forward the code? Alexander shows download code on
https://github.com/workflowhub-eu/workflowhub-graph/tree/feature-source-ro-crates

Oliver considering adding docker-compose for Jena. Could use https://github.com/mcuadros/ofelia to orchestrate one-off jobs.

→ Oliver to add to GitHub repo.

Adding Q&A step? Could do two steps before and after adding to graph.

In some issues with Docker could be SSL certificate errors, so perhaps some warm-up checks as well. Certifi package.

OntoSpy was used to convert from JSON-LD to triples. If we have a directory of JSON files we load them into ontospy (but with the correct @base). This may be where we control the @base and @graph name.

Remaining tasks are more after we've build a graph.

Volodymyr: Can we help on the retrieval? Or is this server-side?

Could the file be hosted by WorkflowHub or by Zenodo?

For next week or so:

- Alexander: Continue with sourcing before starting on big graph building
- Oliver: Add the Docker Compose to repo to spin up Fuseki and a "regular" job that we can plug into
- Finn: Awaiting bug reports on endpoint - deployment on production
- Paul: Could look more on how WorkflowHub collects from Toolshed and if there are alternatives to talk directly to Toolshed.
  Finn: Source code:
- JM: In branch add the SPARQL queries so we can look at with some useful filenames
- Volodymyr: Could look at base URI handling.

Next: 2024-05-24 16:00 BST / 17:00 CEST

2024-05-03 #1

- **When**: Fri 2025-05-03 16:00 BST / 17:00 CEST
- **Where**: https://zoom.us/j/96948642698

Invited/Attending

- **Stian Soiland-Reyes**
- **Volodymyr Savchenko** (EPFL) volodymyrss
- **Denys Savchenko** (UP): dsavchenko
- **Stuart Owen** (UNIMAN)
- **Paul De Geest** (VIB): pauldg
- **Oliver Woolland** (UNIMAN)
- **Alex Hambley** (UNIMAN)
- **Oliver Woolland** (UNIMAN)
- **Nick Juty** (UNIMAN) (unavailable)
- José Mª Fernández (BSC) (unavailable, sick)

Semi-random notes

- Building "initial" Knowledge graph from WorkflowHub
    - Task #1 Build initial pipeline based on WorkflowHub APIs and RO-Crate
        - .. and/or Bioschemas? Some duplication.
        - Validation Workflow RO-Crate + Validate any other profiles
        - Discover any extensions ontologies etc.
    - Task #2: Prepare deposit for Zenodo as a Dataset (RO-Crate + DCAT?)
    - → Later
        - Augment with ORCID/EOSC/OpenAIRE/DataCite information
        - Should we host a live SPARQL endpoint?
    - For deliverable:
        - Document and develop example queries
    - Build a knowledge graph from gathering all the RO-Crates in WorkflowHub as JSON-LD. Iterate over the APIs.
    - Oliver has looked at Apache Jena as possible graph database. Uncovered issues.
    - BioIndustry running Virtuoso as triple store next to SEEK. SEEK can also generate RDF internally for now and populate it, but doesn't do so for workflows.
    - In WorkflowHub there is already RO-Crate.
    - Oliver: Had to convert from JSON-LD to other RDF format first.
    - Neo4J is another common graph database.
    - Build some initial queries.
    - Carole should be finding competency questions from student's earlier work
        - Background materials from student projects etc:
          Knowledge Graphs and Workflows
    - Oliver will demo the Jena prototype graph and how it was gathered.
    - Deciding on architecture. Make it reproducible!

Additional note from JMF: SPARQL query to capture the different RO-Crate profiles (used later to detect bare RO-Crate, Workflow RO-Crate and Workflow Run RO-Crate workflow)
https://github.com/inab/WfExS-backend/blob/full_circle/wfexs_backend/workflow.py#L1334-L1360

Who is doing what?

- Explore reports from previous students
    - → Find competency questions and queries
- Scrape ROCrates from workflow hub
- Upload dataset to zenodo
- Upload dataset to triplestore (Jena?)
- Describe the dataset
- Design some SPARQL queries for extracting triple data
- Decide / investigate a mechanism for created a knowledge graph
- Design / investigate separation of ROCrates by unique ID
- Quality Control
- Explore neo4j as alternative knowledge – GraphQL etc. - property graph
-

Notes from meeting

Oliver had a go making a triple dump into Apache Jena. Shell script, iterated over workflows in WorkflowHub, download RO-crate, unzipped, took ro-crate-metadata.json parsed to Python library "ontospy" (?) that gave triples. Had an Apache Jena running, and chucked the triples into that. Total of 80.000 triples. This took maybe 6 hours! Spent lots of time unzipping.

Do we have an API endpoint that gives RO-Crate? Has been a suggestion. If the file has been generated you can get it from specific file. But some workflows don't have the RO-Crate metadata and then it's not exposed.

SEEK which WorkflowHub is based on, has RDF support as well, which can be generated on changes. This can provide more structural information such as Institution.
https://github.com/seek4science/seek/blob/main/lib/seek/rdf/rdf_mappings.csv

Matched by object type. Makes an RDF for this, it has been connected to triple store, only tested with Virtuoso, but anything with a Ruby adapters will work (e.g. Sesame). This will then be updated on the fly. Stuart is extending this table for BioIndustry. Workflow not currently making RDF. If you enable it will at rest do the '*' ones, rest can be new methods.

Could just do the triple-convert per Workflow from the ro-crate-metadata.json

Do we expose this as a triple store? This makes a public and a private graph! Private includes also non-public entries.

Two strands:

-
- https://workflowhub.eu/workflows.jsonld?dump=true is Bioschemas dump. Only includes properties that WorkflowHub understands. This includes the identifiers. Extracted from provided RO-Crates.

Can scrape RDF e.g. https://workflowhub.eu/people/47.person

What is about the Workflows vs. what is the Workflow Record? Describing the workflow file “inside” vs “outside”.

Submitter vs creator. Multiple ways to upload to WorkflowHub affect how some metadata are extracted.

Could be a GitHub Action for instance.

Fill into https://github.com/workflowhub-eu/workflowhub-graph

Link to Workflow Run Crate and WorkflowHub.

Complex Citations and scholarly connections. Credit & Attribution.

bio.tools knowledge graph could be integrated with Galaxy.

Volodymyr: Could want to query the knowledge graph to find related information. Could do crosswalks etc.

More on provenance of usage of workflows.

Oliver: Industry also want to avoid rerunning similar analysis.

The structure of workflow may not be included.

API for listing the workflows?

Could be a WorkflowHub endpoint for retrieving the ro-crate-metadata.json

Would the scraper need to re-download? LastModified date?

→ Volodymyr to set up skeleton GitHub Action

→ Alexander to sketch out requirements

→ Oliver share shellscript https://github.com/workflowhub-eu/workflowhub-graph

→ Finn to do a one-off data-dump of every ro-crate-metadata.json if public – similar to the bioschemas one.

→