Demonstrator #7
Accessing human sensitive data from analytical workflows available to everyone in EOSC-Life

EOSC-Life

Populating EOSC-Life: Success stories from the Demonstrator Projects

BSC • CRG-EGA • CNAG
Demonstrator #7: Accessing human sensitive data from analytical workflows available to everyone in EOSC-Life
Demonstrator #7: Accessing human sensitive data from analytical workflows available to everyone in EOSC-Life

https://github.com/inab/Wetlab2Variations/tree/eosc-life
Demonstrator #7: Putting the focus on the workflow

https://github.com/inab/Wetlab2Variations/tree/eosc-life
Demonstrator #7: Putting the focus on the workflow

A relevant challenge for the analysis deposited in EGA - the European Genome-phenome Archive - is the need of analysis reproducibility and data provenance > many times users repeat exactly the same analysis because they are not aware of data have been processed.

EGA seeks for a stable repository where workflows can be deposited and/or link to > EOSC-Life WorkflowHub.

RO-Crate is a mechanism to bundle workflows and guarantee provenance. It offers the flexibility of being a single digital object that encapsulate relevant metadata and data (or link to them), e.g. software containers, input and output datasets, etc.

As demonstrator we have contributed to the core development of WorkflowHub as well as to RO-Crate, both the specifications and the python library to use them (researchobject.org/ro-crate/1.1/; github.com/ResearchObject/ro-crate-py)
Demonstrator #7: Putting the focus on the workflow

- Distributed analysis require datasets being processed by the same workflow to contribute to reduce technical noise/biases allowing fair comparisons and performing meta-analysis.
  - Significative gene enrichment in deleterious variants.
  - Common CNV calling in order to have an homogeneous comparison background.
  - Distributed machine learning through federated datasets.

- Having the same workflow available at different locations/installations speed-up the analysis capabilities and allows complying with existing regulation without compromising carrying on research.
  - 10,000 human WGS/WES processed in a single facility vs. 1,000 samples in 10 facilities.
  - GDPR limitations on genomics data leaving specific jurisdictions e.g. Finland.
Demonstrator #7: Why RO-Crate?

Enabling reproducible, transparent research.

Slide from Stian Soiland-Reyes
Demonstrator #7: Executing workflows

... still we need a back-end engine that is able to interpret RO-Crate and execute any workflow contained within.

![Diagram showing the proliferation of standards](https://xkcd.com/927/)

... we are not creating (or proposing to create) the 300th workflow manager.
Demonstrator #7: Executing workflows

We are developing a back-end engine compatible with RO-Crate and GA4GH cloud workstream specifications (TRS/WES/TES):

- Materialize Nextflow and CWL workflows from WorkflowHub.
- Materialize remote repositories that contains the materialized workflows.
- Materialize inputs and reference data sets to launch the workflows.
- Setup Nextflow and CWL engines.
- Validate workflows and materialize their containers.
- Launch the workflows in an execution environment.
- Create execution provenance.
- Generate RO-Crate from execution provenance.

https://github.com/inab/WfExS-backend
Demonstrator #7: Specifying vs. executing workflows

Galaxy

nextflow

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 824087
Demonstrator #7: Putting the focus on the workflow

1. Download (using TRS API)
   - WorkflowHub

2. Materialize
   - RO-Crate

3. Create
   - Launch

4. Create
   - Execution provenance

5. Generate
   - RO-Crate

Workflow repository
CWL/Nextflow engine
Inputs + Reference files
Containers
Demonstrator #7: Access to sensitive data programmatically through EGA

Slide from Fred Aziza and Jordi Rambla
Demonstrator #7: Access to sensitive data programmatically through EGA
Demonstrator #7: Making accessible the workflow

[Image of a diagram]

https://dev.workflowhub.eu/workflows/34
https://dev.workflowhub.eu/projects/4#workflow
Demonstrator #7: Workflow execution

Demo case on the execution of the selected workflow

- Back-end engine installation
  [Link](https://asciinema.org/a/385285)

- Workflow execution files
  [Link](https://asciinema.org/a/385268)
Demonstrator #7: Workflow execution

Demo case on the execution of the selected workflow

- Workflow execution of [https://asciinema.org/a/385420](https://asciinema.org/a/385420)

Executed workflow is described at [https://dev.workflowhub.eu/workflows/34](https://dev.workflowhub.eu/workflows/34)
Demonstrator #7: Going big ...

Update on recomputing variants for WGS/WES for RD-Connect data based on similar workflow at MN4@BSC
Demonstrator #7: Beyond the demonstrator

We plan to leverage the learnings from this demonstrator into different projects and efforts.

TransBioNet is the Spanish Translational Bioinformatics Network of units and groups at healthcare facilities. There is a dedicated working group on Standards and Benchmarking.

OpenEBench is the ELIXIR platform for scientific benchmarking led by communities with common interests and challenges and to facilitate the technical monitoring of Life Sciences software applying the FAIR principles.
Demonstrator #7: Beyond the demonstrator

1. Definition of Benchmarking efforts including input and golden datasets as well as performance metrics
2. Use of OpenEBench for hosting results.
3. Make benchmarked workflows available to the community

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 824087
Demonstrator #7: Beyond the demonstrator

The first TransBioNet Benchmarking effort is ongoing with 15 different contributions from 8 different labs across Spain.

[Link to TransBioNet](gitlab.bsc.es/transbionet/wg1_standards_benchmarking/structuralvariants)
[Another Link](https://dev-openebench.bsc.es/scientific/OEBC005)
Demonstrator #7: Lessons learnt while writing WFs

It is important to recognize early in the process the workflow complexity, e.g. existence of scatter/gather operations, loops, decision points, etc. as some languages/managers do not support (yet) them.
Demonstrator #7: Lessons learnt while writing WF

Ten steps to prepare a portable and fully reproducible bioinformatics workflows. (On-going manuscript)

01. Get to know the domain of your workflow.
02. Get to know the inputs and outputs of your workflow.
03. Identify software containers.
04. Draw a conceptual workflow of your analysis.
05. Modular sub-workflows for common tasks.
06. Workflow Specification vs. Execution. Use of workflow languages.
07. Workflow Managers. What is best for my use-case? Who is going to use it?
08. Guarantee analysis provenance and workflow executability.
09. Make FAIR your workflow.
10. Contribute to community-led efforts to maintain commonly used workflows.
Demonstrator #7: Your experience with EOSC-Life

BCC and CoFest 2020

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