Key topic

PIDs

Scientific domain

Life Science

Leading organisation

UNIMAN
University of Manchester

FAIR Use Cases 07 June 2023

FAIR-IMPACT
Expanding FAIR solutions across EOSC

eosc

Encouraging and supporting researchers in producing FAIR computational workflows

This use case is based around the University of Manchester's work with Persistent Identifiers in data production workflows via its involvement in the **WorkflowHub** - a registry of computational FAIR workflows. WorkflowHub is sponsored by the European RI Cluster EOSC-Life, the European Research Infrastructure ELIXIR and multiple EOSC projects (BY-COVID, BioDT and EuroScienceGateway). Its initial users were from within the life sciences working with COVID-19 workflows, but is now used by over 140 research groups and projects across disciplines.

The overall goal of this use case is to encourage and support **FAIR Computational Workflows**, where workflow systems help researchers in producing FAIR data and recording provenance of their analysis, but also where workflows themselves become FAIR scholarly objects in their own right, appear in the scholarly knowledge graph, gets cited in academic papers, and so on.
Workflows of any type (e.g. Galaxy, CWL, Nextflow, Jupyter Notebook) are registered in WorkflowHub from existing repositories like GitHub, or can be deposited as a direct upload. Metadata is extracted from the workflow and augmented by the user. This is archived in the form of an **RO-Crate** that also contains a snapshot of the executable workflow definition. The metadata uses JSON-LD and **schema.org** vocabulary for Dataset, together with a **Bioschemas** profile for computational workflows. WorkflowHub also uses the standard **Common Workflow Language** (CWL) as a way to describe the workflow structure and detailed annotations such as tools and containers required.

Workflows can be composed of various types of research objects which need to be formally and persistently identified to enable their reuse by other researchers. There are various challenges that need to be addressed resulting from the diverse types of identifiers of the various workflow components. In FAIR-IMPACT we are therefore following several strands to improve persistent identifiers for computational workflows:

- Improve and document explicit identification and linking (FAIR Signposting) from WorkflowHub to PIDs, metadata and RO-Crate downloads
- Enable an automatic request and recording function of Software Heritage identifiers (SWHID) when archiving a Git-based workflow
- Capture and expose PIDs for tools used by workflows (e.g. bio.tools, Bioconda) from Galaxy
- Generate location-independent identifiers (RFC6920) for data generated by workflow runs, potentially large/sensitive, to be included in workflow provenance
- Leverage RO-Crate to capture and propagate workflow provenance outputs and related PIDs
- Create RO-Crate profiles for capturing the provenance of an execution of a computational workflow with increasing granularity

Contributors

Stian Soiland-Reyes (UNIMAN)

Nick Juty (UNIMAN)

Josefine Nordling (CSC)

Reference Materials

WorkflowHub page - Research Object Crate (RO-Crate)

Key topic

PIDs

PIDs

Scientific domain

Life Science

Leading organisation

UNIMAN
University of Manchester

FAIR Use Cases 07 June 2023

FAIR-IMPACT
Expanding FAIR solutions across EOSC
eosc

Encouraging and supporting researchers in producing FAIR computational workflows

Challenges that need to be addressed

There are several challenges related to multiple identifier types being used for various workflow components, such as software and data and their various inter-relations, for instance input and output files, and the details of the runs and types, such as workflow or process. Several different identifiers are already being used, e.g. **EDAM** terms and **biotools** identifiers, but inconsistently and are not propagated through workflow systems to the final results.
Additional challenges result from workflow identification across various repositories that may store the same workflow, or versions thereof, as well as from how to avoid identifier proliferation, for example where these offer various routes for exposure of workflows in the EOSC catalogue. It would also be useful to be able to identify components of workflows, and their related outputs. However, these may vary over time making it difficult to persistently refer back to.
Further challenges are manifested when considering other workflow systems, for instance Galaxy, which while highly used, makes identification of individual workflows difficult.

Expected impact of the Use Case

Through this use case work, we hope to achieve the necessary guidance for researchers to be able to get all relevant components of their computational workflows formally and persistently identified and thus be findable, reusable and citable by other researchers. This also makes it possible to capture provenance information. Furthermore, entire workflows can also become FAIR scholarly objects, equipped with their own identifiers and citing opportunities. All of these advancements significantly improve the traceability, transparency, reliability, and reproducibility of research.

Expected outputs

Improved documentation on identification of workflows and ways of linking workflows or parts of workflows to other research objects.

Contributors

Stian Soiland-Reyes
(UNIMAN)

Nick Juty
(UNIMAN)

Josefine Nordling
(CSC)

Reference Materials
WorkflowHub page - Research Object Crate (RO-Crate)