



# Making workflow provenance FAIR across workflow systems with Workflow Run RO-Crate



Simone Leo<sup>1</sup>, Laura Rodríguez-Navas<sup>2</sup>, José M. Fernández<sup>2</sup>, Paul De Geest<sup>3</sup>, Luca Pireddu<sup>1</sup>, Michael R. Crusoe<sup>6</sup>, Daniel Garijo<sup>7</sup>, Iacopo Colonnelli<sup>8</sup>, Raül Sirvent<sup>2</sup>, Stian Soiland-Reyes<sup>4,5</sup>

<sup>1</sup> CRS4, Pula (CA), Italy

<sup>2</sup> Barcelona Super Computing Center (BSC-CNS), Spain

<sup>3</sup> VIB-Ugent Center for Plant Systems Biology, Ghent, Belgium

<sup>4</sup> The University of Manchester, United Kingdom

<sup>5</sup> University of Amsterdam, The Netherlands

<sup>6</sup> Forschungszentrum Jülich, Germany

<sup>7</sup> Universidad Politécnica de Madrid, Spain

<sup>8</sup> Università degli Studi di Torino, Italy

## Packaging a workflow run as RO-Crate

Workflow Run RO-Crate is a set of profiles of RO-Crate that capture *workflow provenance* in a lightweight FAIR data package, in order to support traceability, reproducibility and interoperable description of diverse computational analysis.

To capture a workflow run, this crate includes or references:

1. The tools that have been executed
2. The computational workflow or script that coordinated their execution
3. The language of the workflow/script needed to execute it
4. The original input files, parameters and configuration
5. The output files and values from the workflow
6. Any files from the workflow management systems

By using RO-Crate as foundation, the provenance structure and metadata is described in a programmatically accessible way along with data files in the same package.

The underlying Linked Data nature of the metadata file help preserve identifiers and enable interoperability with the FAIR ecosystem and existing RO-Crate tooling.

## Interoperable implementations

The workflow run crate profiles are implemented by multiple workflow systems and platforms including **Galaxy**, **COMPSS**, and two CWL implementations (**StreamFlow**, **Autosubmit**).

The workflow execution services **WfExS** and **Sapporo**, both of which support GA4GH APIs, can execute CWL and Nextflow workflows and report outputs as Workflow Run Crates.

The command line tool **runcrate** can convert from the precursor **CWLProv**, and display or validate crates according to the profiles. This tool includes a prototype of repeating an earlier execution.



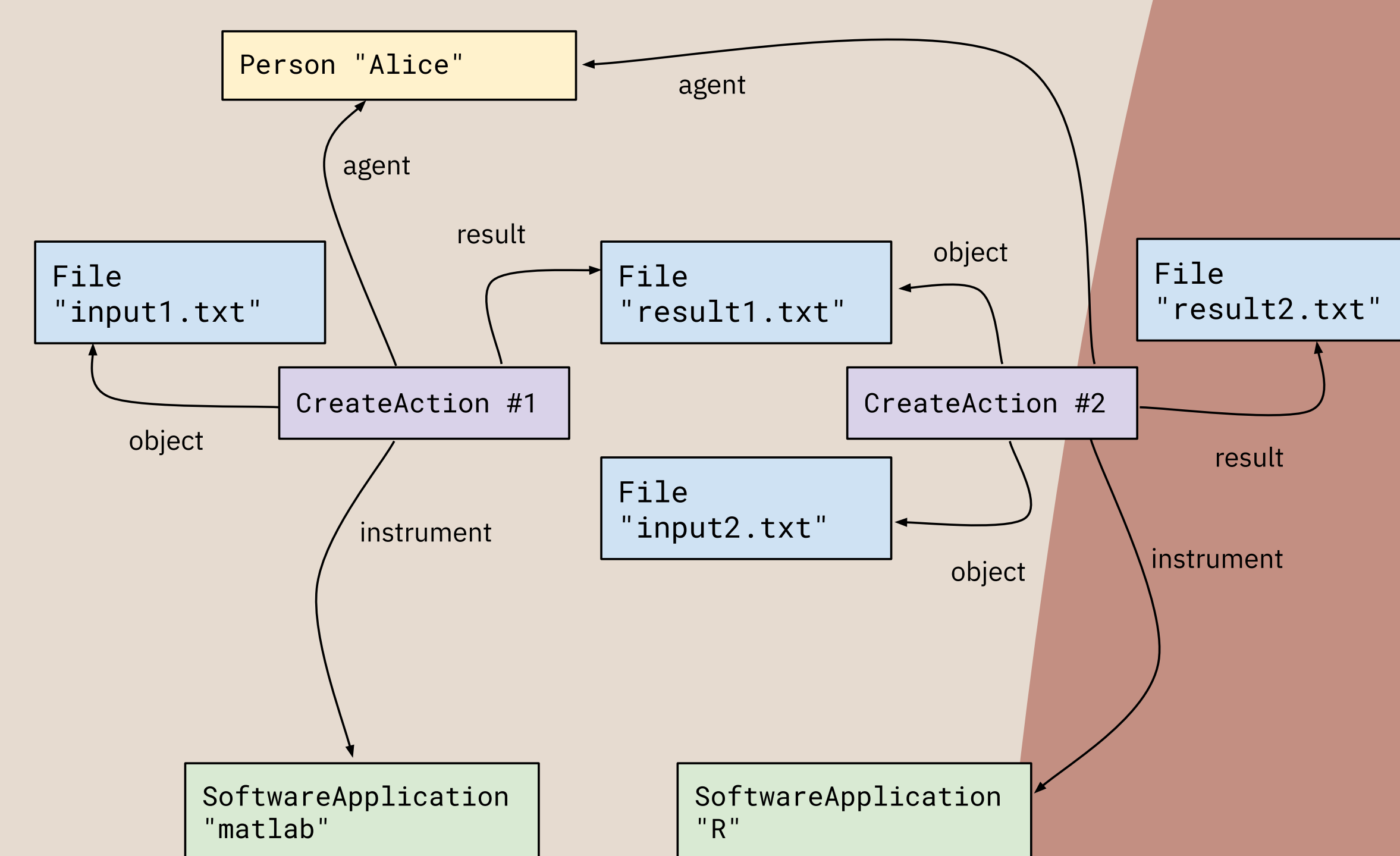
## Profile levels

The profiles are organised by increasing levels of details, allowing gradual adaptation.

- **Process Run Crate** can be used to describe the execution of one or more *command line tools*, even when there is no overall workflow system. The tools are documented, but not directly re-executable.
- **Workflow Run Crate** captures a coordinated execution of tools driven by a *computational workflow*. This crate also follows the WorkflowHub-compatible *Workflow RO-Crate* profile). Sufficient information is recorded to re-execute the workflow in a compatible WfMS.
- **Provenance Run Crate** adds the internal details of each step of the workflow, including their *input/output* values. This profile provides the most granular provenance and is thus the best option for full transparency; it can also be useful for debugging.

Additional detail levels planned for future profiles include: containers, software requirements, hardware resource usage and internal workflow management logs.

<https://www.researchobject.org/workflow-run-crate/profiles/>



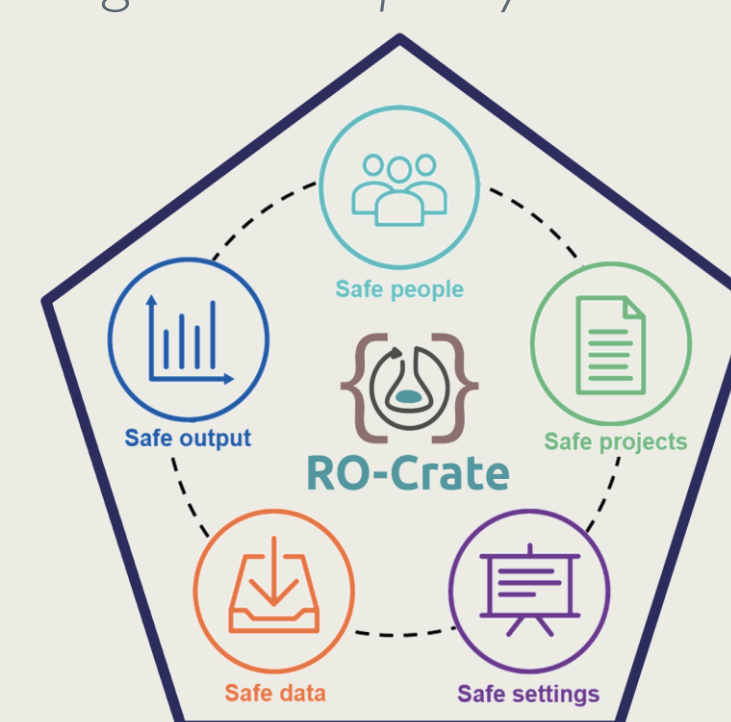
**Example Process Run Crate:** The tool *matlab* was executed by *Alice* to read *input1.txt* and create *result1.txt*, which subsequently was consumed by *R* to create *result2.txt*.

The relations are based on [schema.org/Action](https://schema.org/Action) as recommended by RO-Crate.

## Trusted Workflow Run Crate

The **TRE-FX** project has developed an extension of the Workflow Run Crate profile for use in *Trusted Research Environments* (TRE) to handle sensitive health data in federated workflow execution across TREs in the UK following the *Five Safes Framework*.

In this scenario, a crate with a *workflow run request* references a **pre-approved workflow** and project details for manual and automated assessment according to the TRE's *agreement policy* for the sensitive dataset.



The crate goes through multiple *phases* internal to the TRE, including **sign-off**, **workflow execution** and **disclosure control**. The final crate is then safe to be made public.

**TRE-FX**

<https://trefx.uk/>

<https://w3id.org/trusted-wfrun-crate/0.3>

This extension supports the **human review process** – important for transparency on TRE data usage.

## Community

The **Workflow Run working group** collaborates across >8 ELIXIR nodes and EU-wide projects (**BY-COVID**, **EOSC-Life**, **EJP-RD**, **EuroHPC**, **eFlows4HP**, **EuroScienceGateway**) as well as national projects.

The growing community (>30 members) meets bi-weekly and is open for anyone to participate.



This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

Profiles: <https://w3id.org/ro/wfrun/>

Poster: <https://doi.org/10.5281/zenodo.7996434>