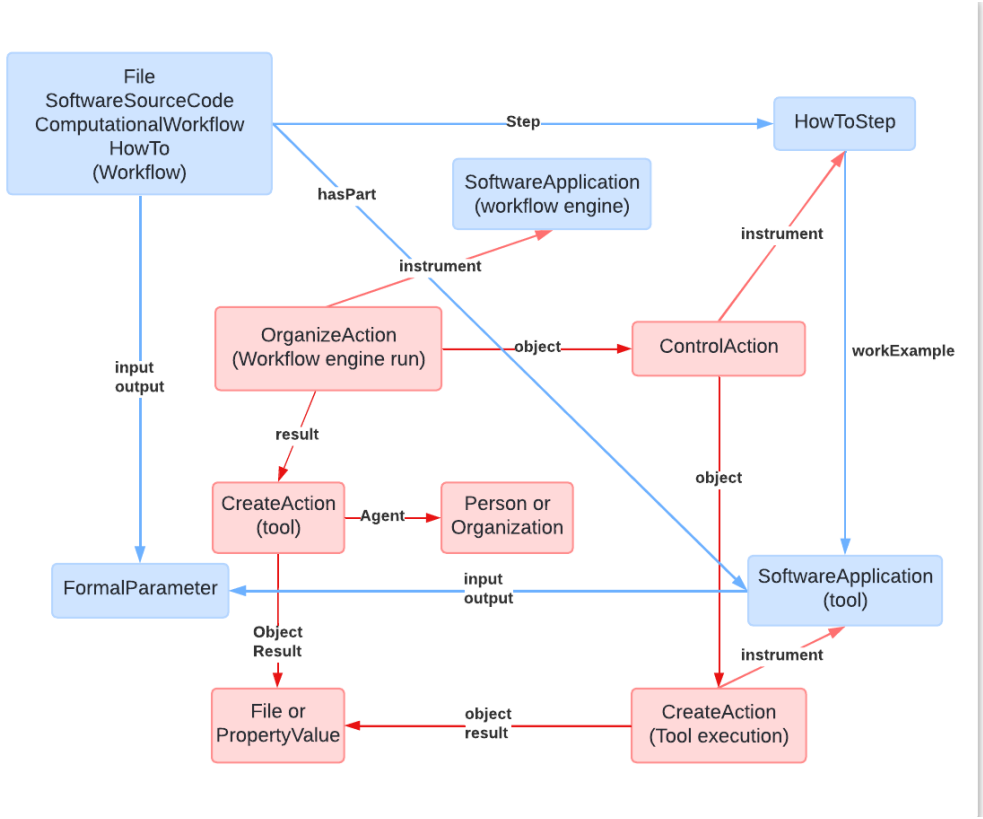
Capturing and recording provenance information from distributed workflows

Dr. Nelis Drost
nelis.drost@auckland.ac.nz

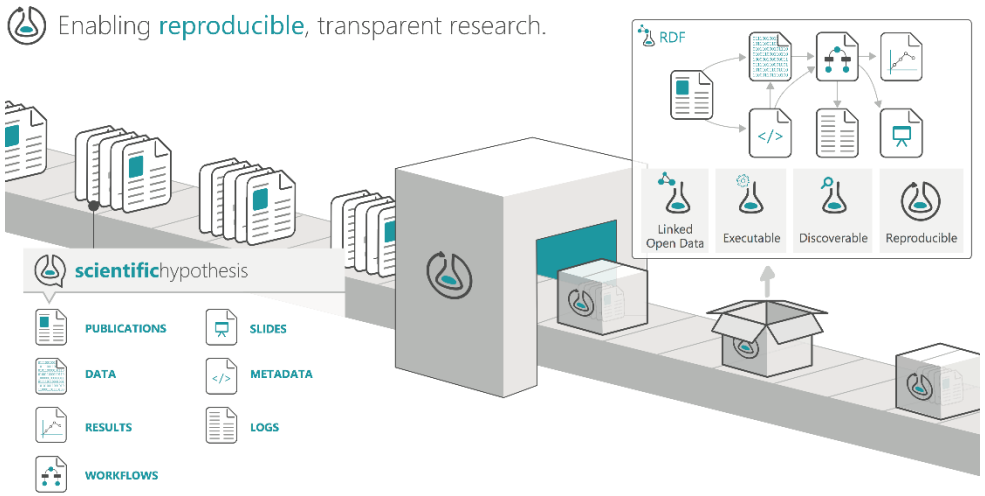
Augustus Ellerm
Prof. Mark Gahagen
Prof. Benjamin Adams
Dr. Ryan Chard
Prof. Ian Foster



What is data provenance?

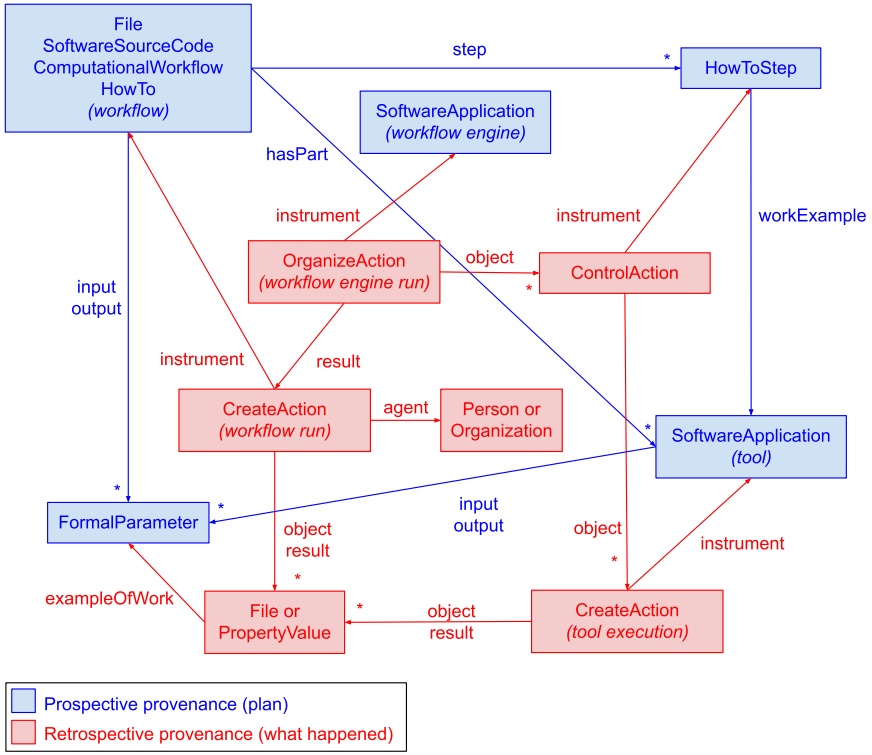
Metadata about how data was generated:

- Data sources: DOIs, UUIDs, etc.
- Authors, organisations, funders
- Instrument IDs
- Code repositories
- Configuration files
- Process



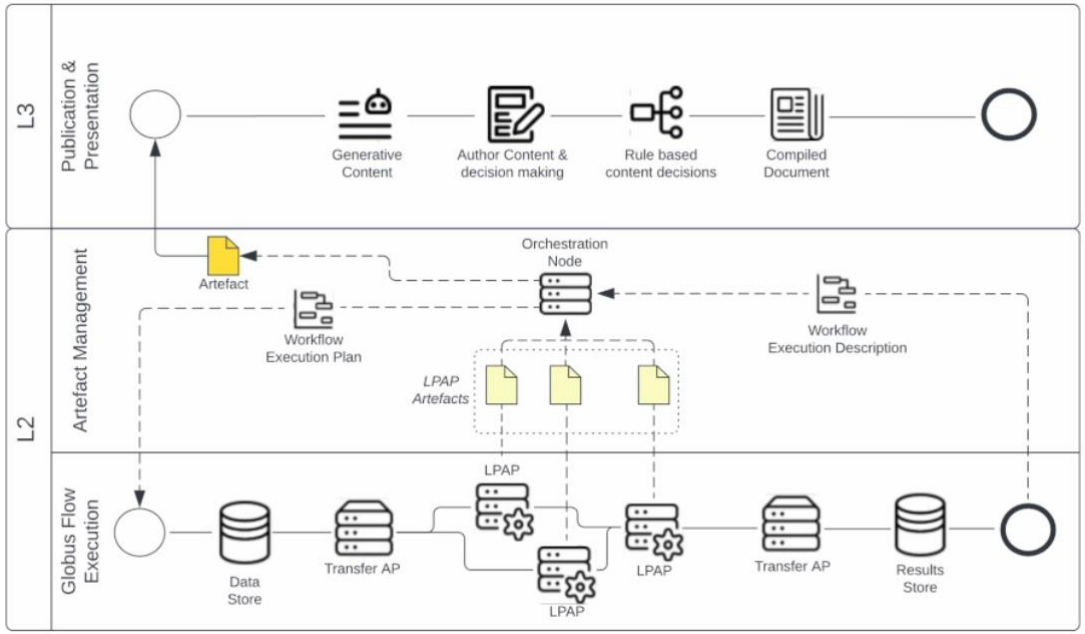
The RO-Crate Provenance Run Crate

- json-ld format
- Extends schema.org + bioschemas
- Details a workflow for generating data/results
- Describes both the planned (*prospective*) and actual (*retrospective*) steps taken



Live Publication

- Machine reproducible paper
- Automatically updateable
- Discovery of computational workflows
- Reuse and remixing of analyses



Integration with workflow management tools

- Workflow tools automate a series of computational steps
- Offer improved reproducibility
- May or may not produce provenance metadata
- Some tooling exists to produce RO-Crates from workflow scripts

Implementation	Profile	Version URL/DOI	Example
runcrate	Provenance	runcrate 0.5.0 or later	10.5281/zenodo.7774351
Galaxy	Workflow	Galaxy 23.1.1 or later	10.5281/zenodo.7785861
COMPSs	Workflow	compss 3.2 or later	10.5281/zenodo.7788030
StreamFlow	Provenance	Streamflow 0.2.0.dev10	10.5281/zenodo.7911906
WfExS	Workflow	WfExS 0.10.1 or later	10.5281/zenodo.10091550
Sapporo	Workflow	sapporo-service 1.5.1 or later	10.5281/zenodo.10134581
Autosubmit	Workflow	Autosubmit v4.0.100 or later	10.5281/zenodo.8144612
Nextflow	Provenance	(nf-prov in development)	example

Provenance from federated systems

- Globus provides automated data transfer
- Globus Compute allows for automated computational flows
- The Globus ecosystem potentially enables federated workflows – combining data and computation from disparate sources

- We aim to produce provenance information describing such flows and making them reproducible

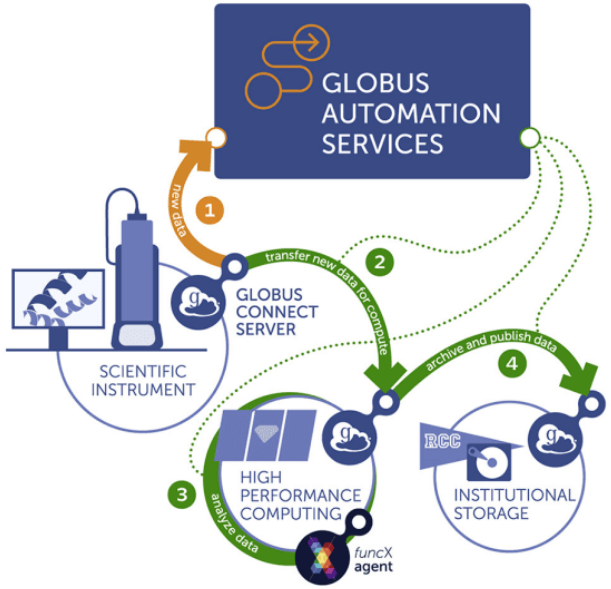
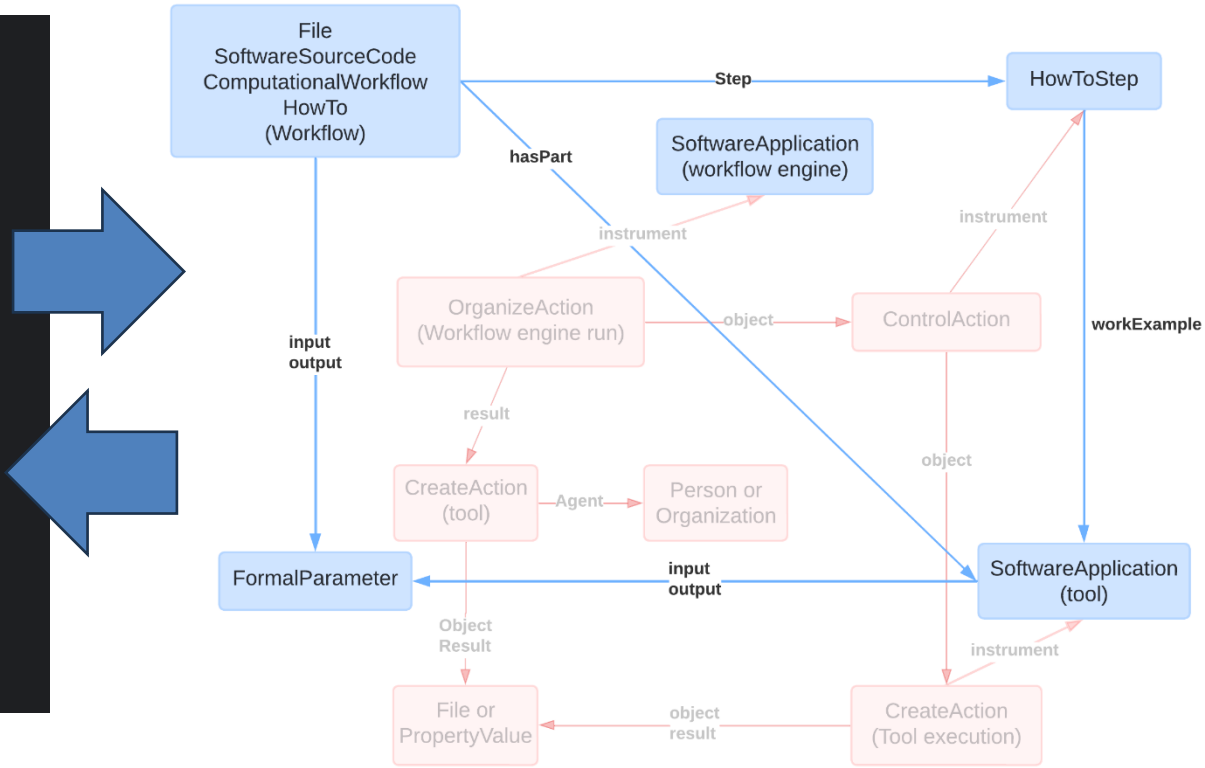


Image from “Linking Scientific Instruments and Computation: Patterns, Technologies, and Experiences.” Patterns, vol. 3, no.

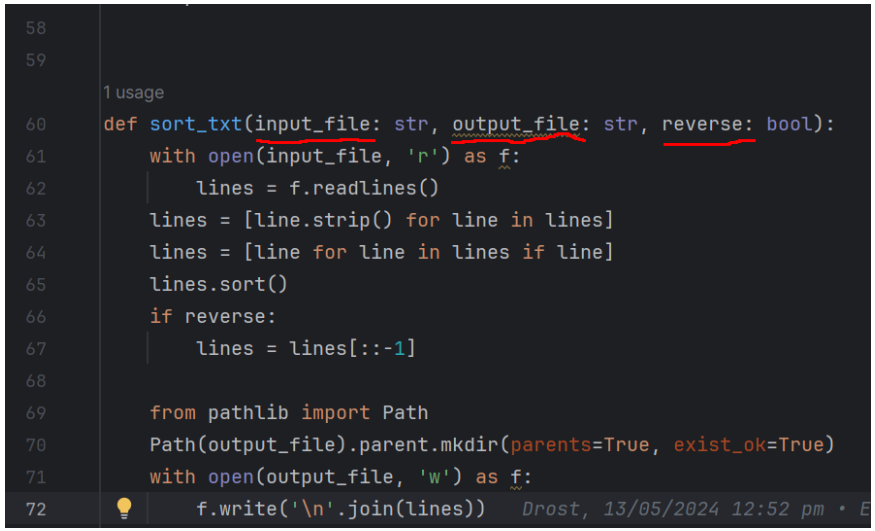
The challenge – lossless recording of workflow execution plan

```
"RevTxt": {
    "Comment": null,
    "Type": "Action",
    "ActionUrl": "https://compute.actions.globus.org",
    "ExceptionOnActionFailure": false,
    "Parameters": {
        "tasks": [
            {
                "endpoint.$": "$.input.compute_endpoint",
                "function.$": "$.input.rev_txt_function_id",
                "payload.$": "$.input.RevTxt"
            }
        ]
    },
    "ResultPath": "$.RevTxt",
    "WaitTime": 300,
    "Next": "Transfer_provenance_rev_txt"
},
```



Building on Gladier – capturing formal parameters

- Gladier simplifies the creation of computational workflows
- We can extend it to provide the additional information needed
- i.e.: Formal Parameters

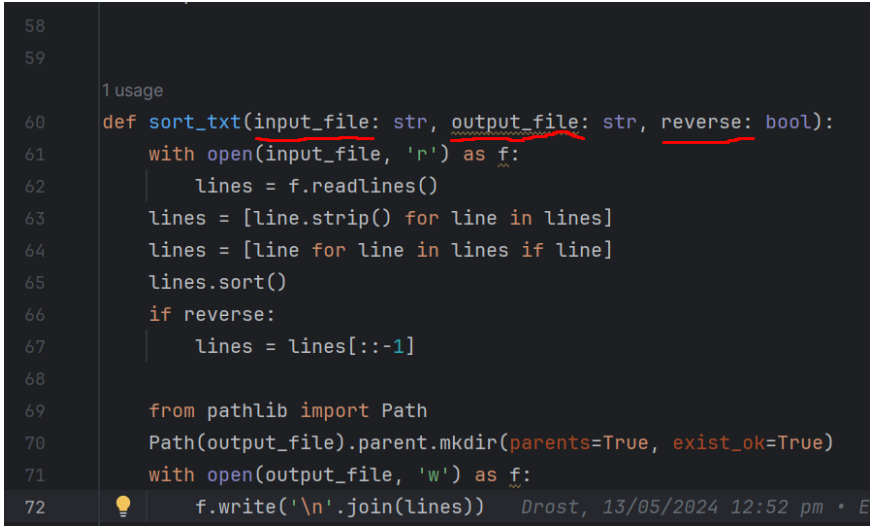


```
58
59
    1 usage
60  def sort_txt(input_file: str, output_file: str, reverse: bool):
61      with open(input_file, 'r') as f:
62          lines = f.readlines()
63      lines = [line.strip() for line in lines]
64      lines = [line for line in lines if line]
65      lines.sort()
66      if reverse:
67          lines = lines[::-1]
68
69      from pathlib import Path
70      Path(output_file).parent.mkdir(parents=True, exist_ok=True)
71      with open(output_file, 'w') as f:
72          f.write('\n'.join(lines))    Drost, 13/05/2024 12:52 pm • E
```

Building on Gladier – capturing formal parameters

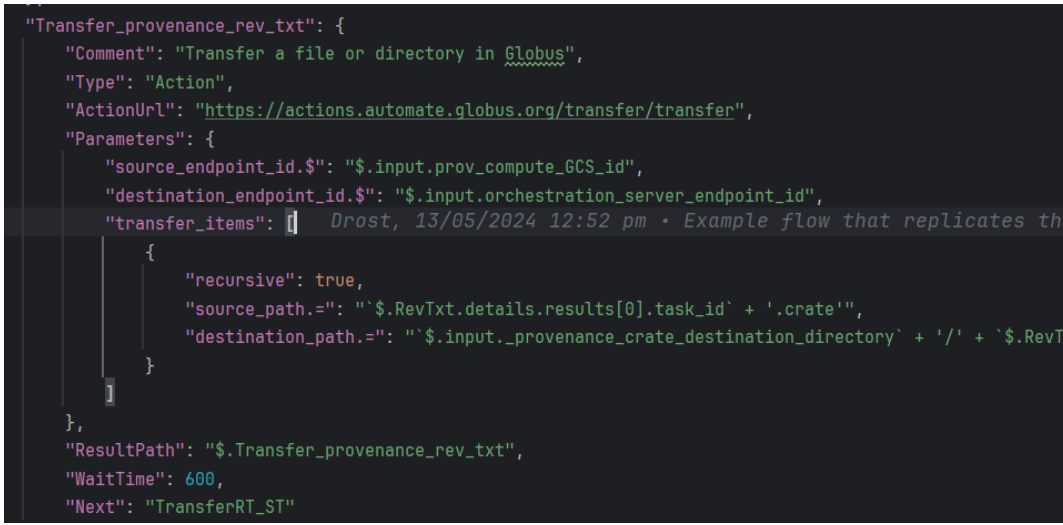
- Gladier simplifies the creation of computational workflows
- We can extend it to provide the additional information needed
- i.e.: Formal Parameters

- Lesson: be more opinionated



Distributed Step Crates

- Capture run time, *retrospective*, information about:
  - Tool execution
  - Hardware/software environment
- Record file creation
- Returned to the orchestration server between steps



Distributed Step Crates

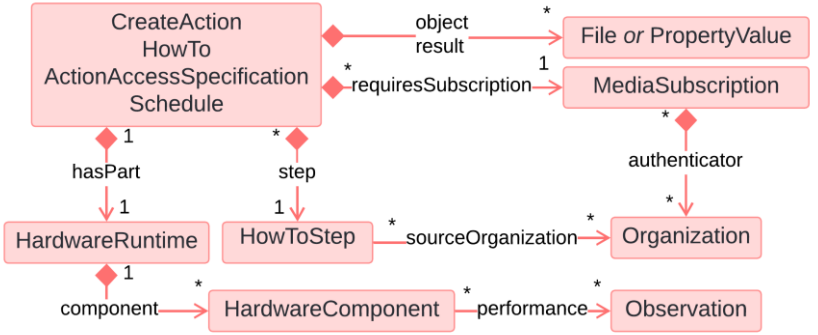
- Capture run time, *retrospective*, information about:
  - Tool execution
  - Hardware/software environment
- Record file creation
- Returned to the orchestration server between steps

- Lesson: none?

```
"Transfer_provenance_rev_txt": {
    "Comment": "Transfer a file or directory in Globus",
    "Type": "Action",
    "ActionUrl": "https://actions.automate.globus.org/transfer/transfer",
    "Parameters": {
        "source_endpoint_id.$": "$.input.prov_compute_GCS_id",
        "destination_endpoint_id.$": "$.input.orchestration_server_endpoint_id",
        "transfer_items": [   Drost, 13/05/2024 12:52 pm • Example flow that replicates th
            {
                "recursive": true,
                "source_path.=": "`$.RevTxt.details.results[0].task_id` + '.crate'",
                "destination_path.=": "`$.input._provenance_crate_destination_directory` + '/' + `$.RevT
            }
        ]
    },
    "ResultPath": "$.Transfer_provenance_rev_txt",
    "WaitTime": 600,
    "Next": "TransferRT_ST"
```

Crate extensions

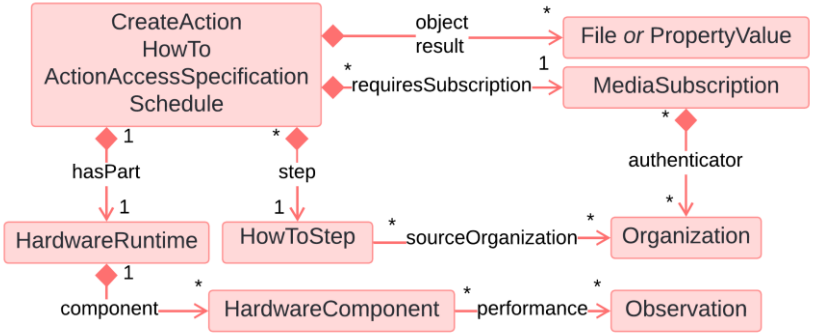
- Hardware specs
  - Captured via a custom compute node
  - Subset captured by wrapped compute function
  - Returned as part of the Distributed Step Crate
- Access/authentication information



Crate extensions

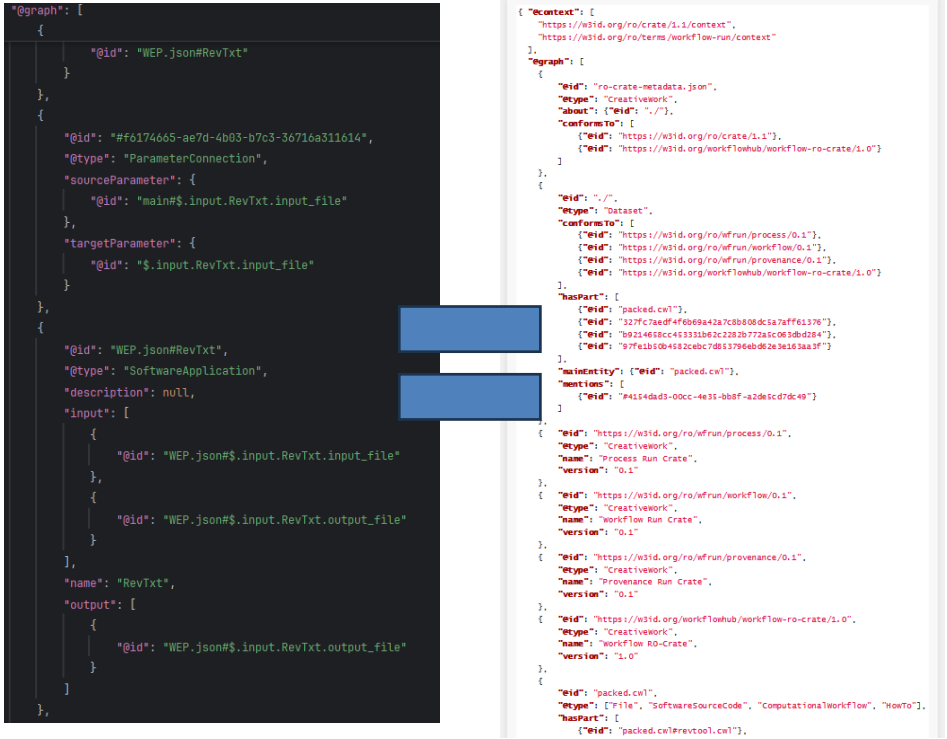
- Hardware specs
  - Captured via a custom compute node
  - Subset captured by wrapped compute function
  - Returned as part of the Distributed Step Crate
- Access/authentication information

- Lesson: don’t shoehorn everything into schema.org



Checking everything works

- Checking RO-Crates by hand is tedious and error prone
- Diff tools don't handle linked json
- *Retrospective* types may not be present – including links
- TDD works well for iterative development



Checking everything works

- Checking RO-Crates by hand is tedious and error prone
- Diff tools don't handle linked json
- *Retrospective* types may not be present – including links
- TDD works well for iterative development

- Lesson: build validation + comparison tools

Lessons learned

- Start with the scaffolding:
  - Validation
  - Gladier
- Complexity is unpredictable
- Be opinionated
- Follow standards where possible, but not at all costs

- Start smaller

Any Questions?

- nelis.drost@auckland.ac.nz
- gus.ellerm@auckland.ac.nz
- https://github.com/LivePublication