26th International Congress on Modelling and Simulation, Adelaide, SA, Australia, 30 Nov–4 Dec 2025
mssanz.org.au/modsim2025

Towards seamless provenance in collaborative modelling: integrating Airflow with Provena

Ross Petridis [a], Jonathan Yu [b], Peter Baker [c], Jordan Munnerley [d], and Fareed Mirza [b]

[a] CSIRO NCMI, Hobart, Tas, Australia, [b] CSIRO Environment, Clayton, Vic, Australia,
[c] CSIRO Environment, Black Mountain, ACT, Australia, [d] CSIRO Environment, Dutton Park, Qld, Australia

Abstract: Provenance capture in collaborative modelling workflows is essential for reproducibility, auditability, integrity, and trust, but often involves manual curation, leading to inconsistent adoption and incomplete or error prone provenance. This paper presents a pilot approach to embed provenance capture within automated computational workflows by coupling an open-source workflow orchestration engine (Apache Airflow) with Provena, an open-source provenance and information system. By enabling passive provenance capture as a byproduct of routine modelling, this architecture supports reproducibility, transparency and standardised provenance practices, particularly as models transition from R&D to operational readiness.

Keywords: *Provenance, automation, workflows, environmental modelling, workflow engines*

1. INTRODUCTION

Collaborative modelling workflows can be complex, requiring robust low-friction provenance capture tools to understand what was done, rerun workflows, and enable reproducibility. Traditional approaches to recording provenance often impose manual overheads on modellers, leading to inconsistent adoption, increased probability for human error, and incomplete capture of activity details, or implementation-specific logs. In environmental modelling, the computational environment in which models are run often lack appropriate interfaces to record provenance to relevant provenance systems, thus requiring post-modelling recording of provenance or manual integration. Provena (Yu et al., 2023), which is an open source provenance system, addresses a gap in provenance tools by enabling provenance recording independently of modelling workflow implementations and standardising how provenance is captured. However, the challenge remains to develop streamlined automated provenance capture for these modelling activities due to the manual nature of curating relevant provenance records, e.g. model run and dataset metadata.

In response to this, we explore a pilot approach to embed provenance capture within modelling workflows in a workflow engine. We propose coupling an open-source workflow orchestration engine (Apache Airflow) with Provena to achieve this. We present a design below that integrates provenance capture directly into workflow execution environments using the Provena Python client to facilitate communication between Apache Airflow (the chief workflow orchestrator) and a Provena provenance system . This passive provenance capture model enables metadata to be recorded as a natural byproduct of routine modelling activities. We explore a use case from the Reef Restoration and Adaptation Program (RRAP) as an implementation of this approach.

2. PASSIVE PROVENANCE VIA WORKFLOW AUTOMATION

Passive provenance refers to the capturing of provenance as a byproduct of workflow execution, rather than as a separate manual process. This approach requires an automated layer between the computational workflow execution and the provenance system. The Provena client library [1] provides this capability by enabling programmatic provenance capture from - or alongside - the workflow execution environment, interfacing directly with a Provena server. Coupled with a workflow orchestration engine like Apache Airflow, this enables fully automated workflow execution with automated provenance capture. Figure 1 shows the proposed passive provenance-enabled workflow system where there is an integration between Provena and Apache airflow, achieved via interfaces provided by the Provena client library. Provena serves as the provenance store allowing seamless inspection, retrieval, and auditing of model workflow activities to provide traceability, auditability, reproducibility, and confidence in modelling activities. Apache Airflow provides an orchestration engine to initiate computational workflow runs automatically, or at least with significant abstractions away from implementation or API-level specifics. With simple scripting either side of workflow executions in Airflow, the Provena client library can be configured to interface automatically with Provena, therefore providing automated Provena capture at workflow

[1] https://provena.github.io/provena-python-client/

execution time. Hence replacing the manual and error prone overhead previously required to record provenance for each additional workflow execution, with a once off code configuration. This enables scalable provenance capture as workflows can be triggered by an orchestrator (or automatically by Airflow or some other event) with provenance capture now embedded in the workflow execution.

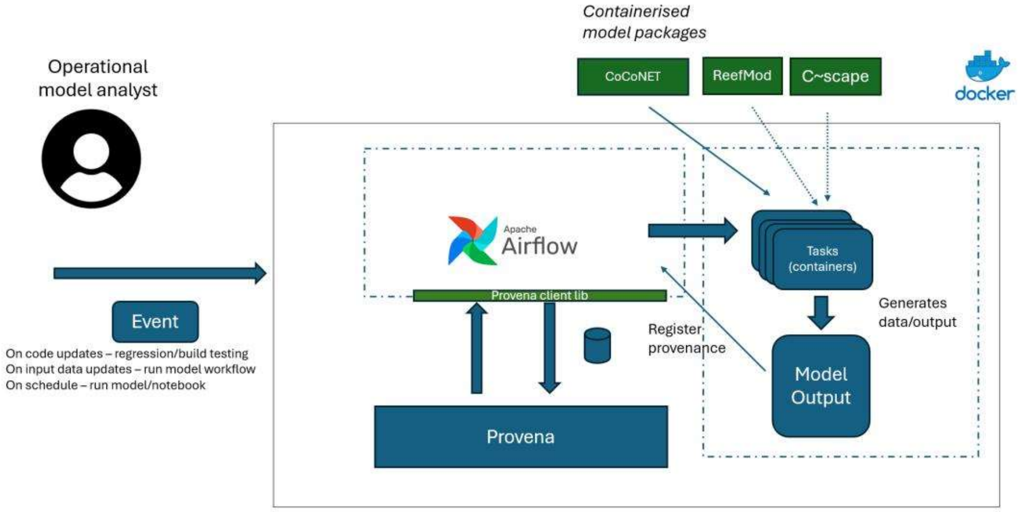

Figure 1. Architecture of a passive provenance-enabled workflow system.

When implemented this architecture provides a simple interface enabling users who are not modelling experts to trigger available workflows without needing to know API-level details of how to interface with Provenance system, or how to orchestrate compute resources. It enables scalability as the need for manual provenance capture is removed and Airflow can trigger multiple workflows or components of workflows in parallel. Airflow also provides the opportunity for computation resources to be parameterised and specified as part of the workflow instantiation. This approach is extensible to other models, however, the pre-requisite requirements are: a) models are required to be containerised with clear input/output protocols; b) compute facilities such as HPC, or access to virtual machines available; c) a deployed Provena server instance is available; d) an Apache Airflow server is deployed with user accounts configured.

We are exploring the deployment of this architecture to support the RRAP Modelling and Decision Support program which runs environmental and ecological modelling of the Great Barrier Reef (GBR) to the year 2100 under a range of possible future climate scenarios utilising the ReefMod, CoCoNet, and C-scape ecological models. Our target users are operational modellers who are outside of the core R&D ecological modelling teams. This frees up the R&D teams to develop models further and enables operational modellers to run these models routinely and as needed. As models mature and are developed further, they can be incorporated into the established workflows.

3. CONCLUSIONS AND FUTURE WORK

Enabling passive provenance capture as discussed above supports reproducibility, transparency and standardized provenance practices for potentially any modelling activity implemented. This approach supports a system for modelling and standardized provenance practices through automation. This is especially useful as models transition from the R&D phase to an increasing operational ready state. While this approach is being applied specifically for the RRAP Modelling and Decision Support program for running environmental and ecological modelling of the GBR, we propose future work in applying this approach to additional models and other domains.

4. ACKNOWLEDGEMENTS

This work was funded by Reef Restoration and Adaptation Program (RRAP), which is funded by the partnership between the Australian Government's Reef Trust and the Great Barrier Reef Foundation. The authors would like to acknowledge RRAP Modelling and Decision Support operational partners at Adaptus, Australian Institute of Marine Science, and University of Queensland for their input into the design process.

5. REFERENCES

Yu, J., Baker, P., Cox, S. J. D., Petridis, R., Freebairn, A. C., Mirza, F., Thomas, L., Tickell, S., Lemon, D., & Rezvani, M. (2023). *Provena: A provenance system for large distributed modelling and simulation workflows*. 14–20. https://doi.org/10.36334/modsim.2023.yu90

