Provena: A provenance system for large distributed modelling and simulation workflows

Jonathan Yu [a], Peter Baker [b], Simon J.D. Cox [a], Ross Petridis [a], Andrew C Freebairn [b], Fareed Mirza [a], Linda Thomas [c], Sharon Tickell [d], David Lemon [b] and Mojtaba Rezvani [b]

[a] CSIRO Environment, Clayton, Victoria, Australia
[b] CSIRO Environment, Black Mountain, ACT, Australia
[c] CSIRO Environment, Hobart, TAS, Australia
[d] CSIRO Environment, Brisbane, QLD, Australia
Email: Peter.Baker122@csiro.au

Abstract: The ability to evaluate options and support decision making in the face of uncertainties and future scenarios is a key challenge in many domains. For environmental domains, supporting the decision-making process for evaluating social, economic and environmental pathways relies on many factors. These include information from modelling and simulations using numerous datasets, including climate data, socio-economic trade-off modelling, and ecological effects from environmental events. This information needs to be trusted and defensible. Thus, capturing and recording the provenance of the scientific modelling - their inputs (e.g. datasets and who conducted the modelling, and with what models), and their results (e.g. resultant datasets and information products) - is critical to high-quality decision making. In a large initiative such as the Reef Restoration and Adaptation Program (RRAP), decisions relating to the Great Barrier Reef include prioritising investment in certain interventions over others. The scientific modelling that is required to produce the relevant information is carried out by multiple teams, who may run one or more computational models. Additionally, each team may depend on another team's outputs to carry out their modelling. Therefore, an approach is needed for capturing provenance in a consistent manner across distributed and heterogeneous modelling environments.

To meet this need, we developed a solution (called Provena) for capturing and querying data, and workflow provenance, in a standardised manner across multiple modelling environments. The W3C Provenance conceptual model, called PROV-O, was extended to capture workflow provenance in the Provena implementation. A novel aspect of the Provena is a registry that allows registration of each element in the provenance record in a general purpose metadata registry and minting a persistent identifier for each element, e.g. entities like datasets, model run workflows and people involved in the modelling. Maintaining persistent identifiers for each registered item allows linkages to be created between entities, people, and model workflow activities using standard PROV-O semantics. Linking up each of these elements provides a provenance chain between activities, entities and people which can be queries. Provena uses an optimised query engine implementing a graph database to enable provenance specific query capabilities that supporting queries such as inspecting all upstream lineage activities and inputs from a single output dataset. While the Provena system is general purpose and can be applied to many domains, we applied this to the RRAP modelling and decision support activities and demonstrate its applicability in that context in this paper.

Keywords: *Provenance, workflows, decision support, information systems, graph databases*

1. INTRODUCTION

The field of modelling and computational workflows is rapidly changing. In the environmental domain, researchers are now able to develop more complex modelling packages and explore more climate scenarios in almost real-time by processing larger volumes of cheaply stored data on increasingly powerful hardware. This facilitates the production of data products which describe a wider range of future scenarios at finer resolutions. Downstream use of a dataset in critical processes, such as data driven decision making, requires confidence in its validity and suitability, enabled by trust and transparency. Therefore, the ability to scrutinize the quality, accuracy and history (provenance) of the data, is critical.

Capturing the provenance of data enables users to review, interpret and scrutinize the data more effectively. Provenance tools assist researchers to produce records which describe the lineage of a dataset, usually by capturing the inputs, processes and associations which led to its creation, derivation or modification. “Provenance answers the questions of why and how the data was produced, as well as where, when and by whom” (‘Data Provenance’, 2022). Provenance capture can occur at several abstraction levels, each potentially requiring a different approach. These range from the operating system (or script level) to the experiment or workflow (e.g. inputs and outputs as file or references). For workflow provenance, there are broadly two approaches – white-box and black-box approaches. White-box approaches focus on transparency and reproducibility by capturing the inner workings of processes. These often rely on workflow engines and functions to export to common formats, e.g. (Garijo and Gil, 2011) and (Belhajjame *et al.*, 2015) present methods to capture provenance using the PROV-O model. The alternative is ‘black-box provenance’, which is capturing provenance at the workflow granularity “whose inner workings are not accessible or not relevant” (Ludäscher, 2016). This approach simplifies the capture and description of inputs and outputs to, and details of, the workflow, thereby enabling:

1. Traceability and potential reproducibility of data, model runs, results, workflows and decisions
2. The building of trust through transparency of processes leading to decisions, and
3. Opportunities for identifying benefits, risk and optimisations through analysis of workflows

Capturing workflow provenance is a focus of the Modelling and Decision Support (M&DS) component of the Reef Restoration and Adaptation Program (RRAP) for the reasons listed above. RRAP is a collaborative program of research seeking to explore the possible outcomes of interventions to preserve and restore the Great Barrier Reef (GBR). Examples of interventions include biocontrol to restore coral reef health and marine cloud brightening solutions to relieve heat and light stress on coral reef organisms. Such interventions require rigorous assessment before investment and deployment. In the context of RRAP, decisions about these interventions are highly varied and subject to significant complexity and uncertainty, in terms of 1) the knowledge base pertaining to the ecosystem, 2) the associated socio-economic and cultural system, and 3) the potential future outlook across these dimensions given local and global forces and threats. The M&DS subprogram was initiated to support decision makers, the broader RRAP program, reef managers, and industry partners to determine objectives, evaluate options and generate knowledge around RRAP intervention options. Thus, capturing provenance of inputs to those decisions is critical for transparency and auditability. For RRAP M&DS, using a black-box provenance approach to capture workflow provenance is sufficient. It allows decision makers and auditors to understand the provenance of data which supports a given decision.

While there are many examples of white-box workflow provenance capture tools, there is a lack of tools supporting the black-box approach. We address this gap in this paper with Provena, which provides an architecture and solution for capturing provenance at the workflow granularity to support large, distributed modelling and simulation with RRAP M&DS as a driver for its application. In Section 2, the RRAP M&DS provenance requirements and respective use cases are discussed. In Section 3, an overview of the Provena system is presented. In Section 4, the querying capability of the Provena system is discussed with examples. In Section 5, we discuss the Provena system and related work and potential future work. Lastly, in Section 6, we provide conclusions. While Provena has been designed with the RRAP M&DS in mind, the solution is broadly applicable in other research programs that require the capturing of provenance for transparency.

2. RRAP M&DS PROVENANCE REQUIREMENTS

A key objective of RRAP M&DS is to enable people to make and communicate quality decisions relating to prioritisation and evaluation of interventions in the Great Barrier Reef. Facilitating the registration and storage of data, and capture of workflow provenance is an important component in providing input into the decision-making process as it provides transparency and a knowledge base of entities and activities used in the program. Motivated by the RRAP M&DS goals, user requirements for the design and implementation of a data and provenance architecture are presented in Section 2.1 and a list of accompanying use cases in Section 2.2.

2.1. User requirements

Requirements for capturing provenance and data lineage for each user group, are shown in the Figure 1.

1. M&DS Modelers
- Enable registration, storage and sharing of relevant datasets
- Enable recording details of modelling activities – processes that were run to generate the datasets, or artifacts used in decision making

2. RRAP stakeholders – researchers, decision makers, other stakeholders (reef managers, partners)
- Facilitating access to the relevant data
- Enabling users to understand and trust the modelling activities that are carried out.

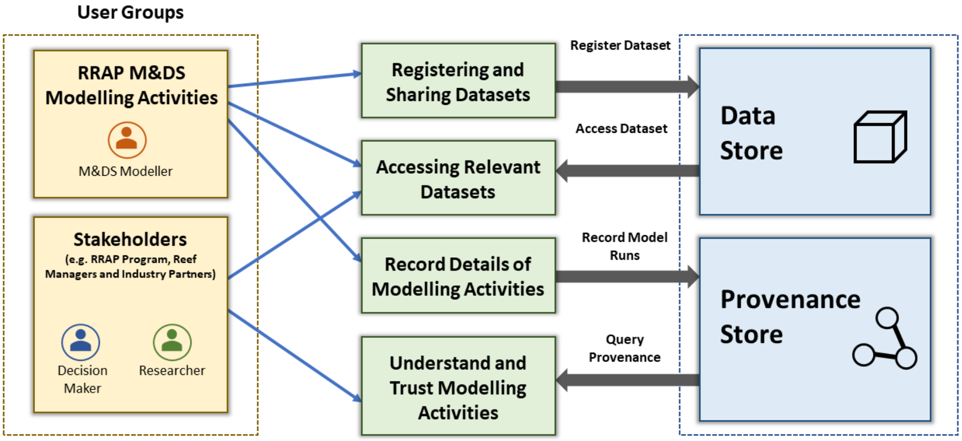
User Groups
RRAP M&DS Modelling Activities
M&DS Modeller
Stakeholders
(e.g. RRAP Program, Reef Managers and Industry Partners)
Decision Maker
Researcher
Registering and Sharing Datasets
Accessing Relevant Datasets
Record Details of Modelling Activities
Understand and Trust Modelling Activities
Register Dataset
Access Dataset
Record Model Runs
Query Provenance
Data Store
Provenance Store

Figure 1. Use cases

The four use cases (shown in green in Figure 1) are critical, as the data and modelling activities require review during a decision-making process or in a retrospective audit in the future, e.g. a person challenging the integrity of the decisions and information used. A system that has the capability to provide that evidence base is important.

2.2. Provenance query use cases

At a workflow level of granularity, a requirement is to be able to query the activities, their inputs and outputs and any related information to help users of the model outputs understand and trust the modelling activities that are carried out. This provides transparency as to what was run, by whom, at what time, and with what data. We provide 7 use cases for provenance queries that are relevant in this context (Table 1). These requirements and provenance use cases have informed the design of a solution, presented in the next section.

Table 1. Provenance query use cases

Query Use case #	Description
1	Find the primary sources of a data/information item
2	Show the summary of how the final outcome of a workflow has been generated
3	Show the lineage of a workflow
4	Find the lineage of a result
5	Find the ancestor data sources to a data object
6	Find the usage of a data object
7	Find all results derived from a dataset

3. THE PROVENA ARCHITECTURE AND IMPLEMENTATION

A key requirement of a provenance system is to facilitate the capture of workflow provenance for the use-cases presented in Section 2. Provenance information can then be visualised, summarised and queried by modellers, decision makers, auditors and other stakeholders. To enable workflow provenance, the workflow and references to its inputs and outputs must be able to be registered, managed and stored. Figure 2 shows the design architecture of the proposed data and provenance system, called Provena. The core component of Provena is the Registry, which provides a central location for all records including dataset metadata, critical agents (people and organisations), model run records, and supporting entities. The Data Store and Provenance Store communicate with the Registry to facilitate the registration and visualisation of datasets and workflow provenance respectively.

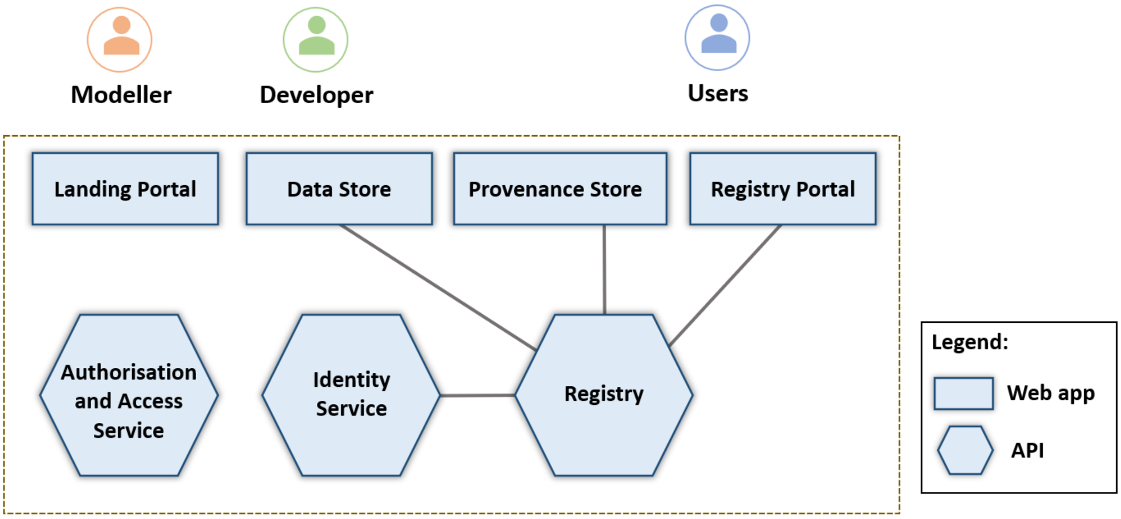

Figure 2. Technical architecture of the Provena Data and Provenance System

Provena supports these functions:

1. **Register and store datasets:** Users can register a dataset, provision a data storage location, and upload and download data. Registered datasets are persistently identified as a resource in the Registry.
2. **Record model runs:** Users can record model runs so that key results are persistently available and auditable. The model run record is registered in the Registry and is assigned a persistent identifier.
3. **Centralised and managed provenance records:** Provenance records are consistent, high quality and discoverable through the registration of and reference to well defined entities in a structured format.
4. **Register supporting items:** Users can register supporting items to reference in a model run provenance record, e.g. people, organisations, and dataset templates. These are assigned a persistent identifier.
5. **Viewing, exploring and querying provenance records:** Users can explore provenance records to find relevant model runs and datasets easily, and query/visualize provenance lineage. This includes tracing the lineage of model run and dataset provenance entries to explore the audit trail of related records.

A key principle of the Provena system is that each item in a provenance record is assigned a persistent identifier (PID). PIDs enable linkages to be created within the system in a robust way, i.e. identities and their linkages persist even if the underlying IT infrastructure changes. This is critical for creating well-maintained workflow provenance records and ensuring their longevity over time. The Australian Research Data Commons (ARDC) Handle Service is used to generate PIDs in Provena and persistence is guaranteed as long as the underlying infrastructure continues to be operational, e.g. ARDC’s Handle Registry, the Handle.net global infrastructure, and the resolution to Provena registered items. In Figure 3, we depict how a PID is used in the Provena system.

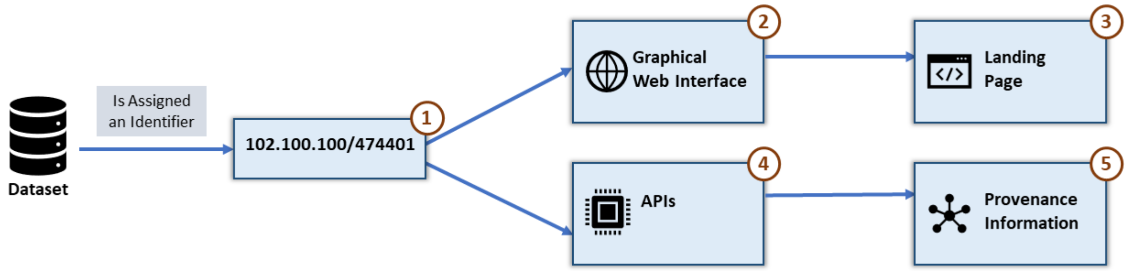

Figure 3. Use of PIDs in the Provena system - *(1) Dataset is registered and assigned a PID; (2) PID is looked up online via a web browser, which (3) automatically redirects to a landing webpage describing the resource; (4) The PID can be used in the Provena system to query provenance information and links to other resources (5).*

To enable the capture of workflow provenance in Provena, we have designed a structured schema for creating and registering items. The Provena schema is based on the W3C PROV-O ontology (Lebo, Sahoo and McGuinness, 2013). PROV-O defines a process-flow model, in which all the elements are classified as either entities (i.e. continuants), activities (i.e. processes or occurrents) or agents (i.e. people, organisations, instruments or software) that are involved in producing a piece of information or a thing. These components are related by four generic property types: *wasGeneratedBy*, *wasAssociatedWith*, *wasAttributedTo* and *used*. Specifically, the Provena schema specialises PROV-O concepts to enable the recording of workflow provenance, which is shown in Figure 4. Specialisations include:

- Entity classes (in blue): Dataset, Dataset Template, Workflow Template and Model
- Agent classes (in orange): Organisation and Person.
- Activity classes (in green): Model Run

Enforcing a schema *with more specialised classes than* PROV-O also led to reusability and knowledge sharing. For example, by having users describe a Workflow Template, it enables reuse and the rapid registration of multiple model runs. Furthermore, the description of expected inputs and output datasets as Dataset Templates enables standardisation within and between teams about expected data formats, domains and structures. During initial user testing, a benefit we observed was that conversation occurred within and between teams during the process of encoding modelling workflows using the Provena schema, that would otherwise remain implicit.

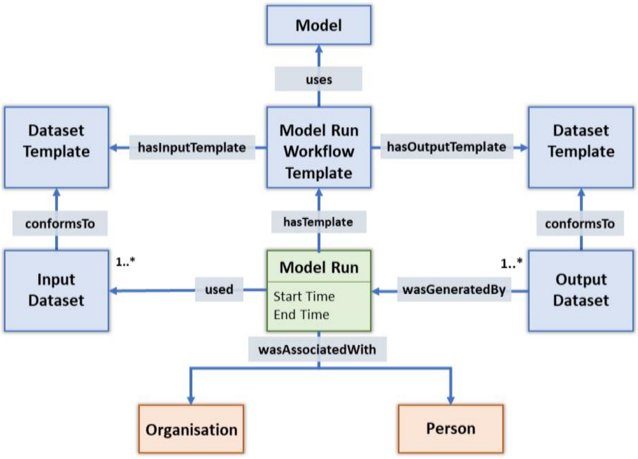

Figure 4. Model run provenance conceptual model

Secondly, provenance record consistency and quality are bolstered, increasing the trustworthiness and transparency of its described artifacts. Consider an alternative system in which the user directly encodes activities into registered PROV-O records. The ability to predict, comprehend, maintain, and reason about the provenance records is significantly diminished due to the variability in structure, content, and quality. High quality, semantically rich and consistent PROV-O records are generated due to a quality and consistency gate facilitated by the Provena schema. The schema also provides querying and analysis consistency.

Figure 5 shows an example provenance record capturing a model run workflow and other related information (input data, associated modelers, outputs data, modelling software/processes) using the example of the CoCoNet model. In this example, a Model Run Workflow Template, called the *CoCoNet counterfactual model template*, is defined specified with expected inputs and outputs via the respective Dataset Templates, and details about which model was used. The templates allow model run records to be created with expected information to be included. Details about the shape of the actual datasets (expected files) are specified in the Dataset Template instance and conformance can be validated.

The method for capturing model run provenance is non-prescriptive. In RRAP M&DS, each modelling team develops their model and their model runs as independent units, often using different software implementations

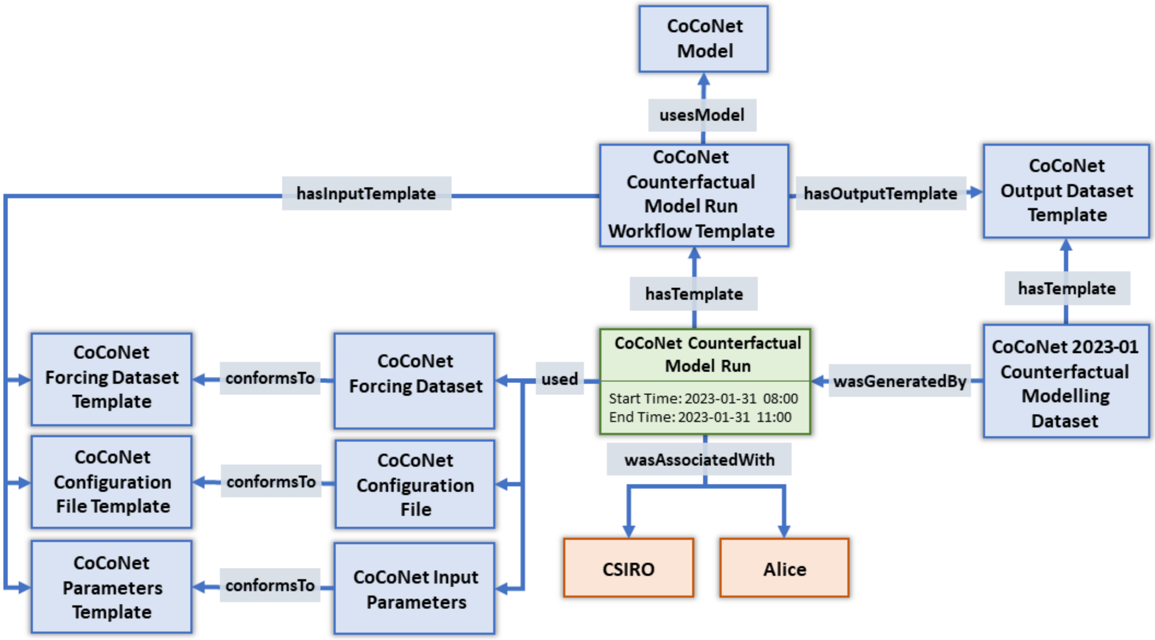
CoCoNet Model
usesModel
hasInputTemplate
CoCoNet Counterfactual Model Run Workflow Template
hasOutputTemplate
CoCoNet Output Dataset Template
hasTemplate
hasTemplate
CoCoNet Forcing Dataset Template
conformsTo
CoCoNet Forcing Dataset
used
CoCoNet Counterfactual Model Run
Start Time: 2023-01-31 08:00
End Time: 2023-01-31 11:00
wasGeneratedBy
CoCoNet 2023-01 Counterfactual Modelling Dataset
CoCoNet Configuration File Template
conformsTo
CoCoNet Configuration File
wasAssociatedWith
CoCoNet Parameters Template
conformsTo
CoCoNet Input Parameters
CSIRO
Alice

Figure 5. Provenance record example using CoCoNet

(e.g. R, Matlab, Julia, Python) run on different operating systems (e.g. Mac, Linux, Windows) and in different computational environments (e.g. HPC, cloud, local computer). As such, we present a web Application Programmatic Interface (API)-based design that is generalised enough to accommodate the different environments using adapters and client libraries/applications that call the relevant APIs for registering provenance. In our pilot implementation of this design, we have developed a few options for interfacing with Provena which expose helpful abstraction layers above the core system REST APIs:

1. Client libraries or wrappers: See https://github.com/orgs/gbrrestoration/repositories?q=wrapper. The user can create scripts to create model run records and registry items as models are run as part of an automated process. Currently Python and Bash clients are supported.
2. Provenance model run uploader application: A simple web application that enables users to upload model run records via a templated CSV format. Users can configure their modelling scripts to write entries in a CSV file which has been structured according to the corresponding model run workflow template. The pre-requisite is a model run workflow template to generate the CSV template.
3. Users instrument the modelling workflows by interfacing with the REST APIs directly.

4. PROVENANCE QUERYING

To satisfy the provenance query use cases presented in Section 2.1, the Provena system integrates a query engine to provide the necessary results. For version 1.0 of the Provena system, the community edition of the Neo4j Graph Database (https://neo4j.com) is used as the backend provenance query engine, as provenance records lend themselves to being represented as a graph, e.g. outputs of a model run or workflow can be inputs to other model runs and workflows. Furthermore, native graph database implementations often implement query engines which offer tailored query syntax and performance suitable for making deeply nested lineage and relationship oriented queries (which are often challenging or impossible for traditional relational database engines). The Provena system decouples the persistence of registered items (the Registry satisfies this requirement) with the database engine for storing and querying provenance. This allows APIs to be implemented and allows interfacing with any other underlying query engine/database in the future.

The ProvDB Connector plugin for neo4j (Bieliauskas *et al.*, 2022) is used to ingest the model run records which are represented using PROV-O. By loading the provenance of model runs and related items, a provenance graph is created spanning the modelling activities of RRAP M&DS. The Neo4j's OpenCypher query engine (http://opencypher.org/) is used to perform queries to explore all relationships from a starting node (e.g. show me all nodes that link to a selected node) as well as specify queries using standard PROV-O predicates (e.g. show me items that are related by the *wasGeneratedBy* predicate from a selected node).

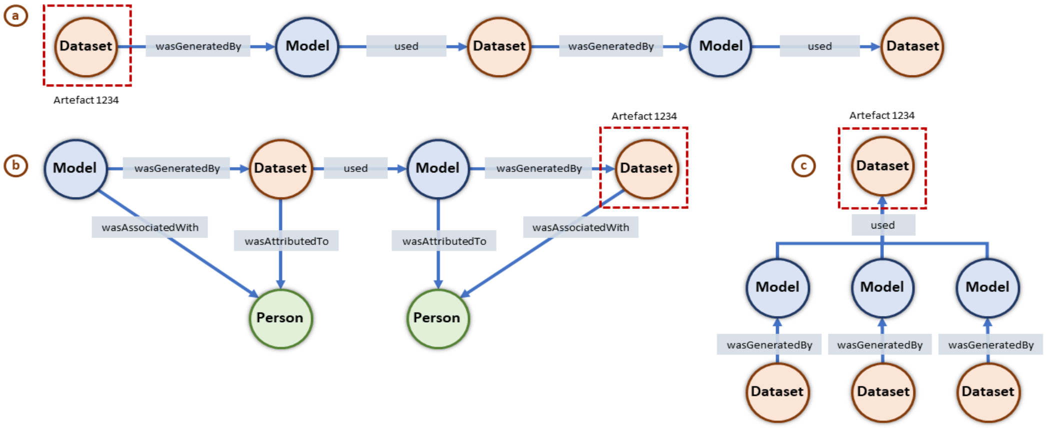

Figure 6. *(a) Query result: Which datasets were used to produce an artifact; (b) Query result: Who was involved in creating the resultant artifact; and (c) Query result: Downstream artifacts of model runs from a source dataset.*

Example queries are listed below with the corresponding Neo4j query, and the visualisation of query results shown in Figure 6:

- What data was used to produce selected artifact (id = 1234)? (Figure 6a)
  ```
  (o:Entity {`item_subtype`:'DATASET'}) <-[*1..5]- (n :Entity {`identifier` : '1234'})
  ```
- Who was involved in creating the selected artifact (id = 1234)? (Figure 6b)
  ```
  (o:Agent {`item_subtype`:'PERSON'}) <-[*1..5]- (n :Entity {`identifier` : '1234'})
  ```
- What artifacts are affected if a result from selected artifact (id = 1234) is invalidated? (Figure 6c)
  ```
  (o:Entity{item_subtype:'DATASET'})-[*1..3]->(n:Entity{`identifier` :'1234'})
  ```

5. DISCUSSION AND FUTURE WORK

There are other approaches to capturing workflow provenance using a similar architecture to Provena, where provenance records are based on the PROV-O ontology, namely PROMS (Car, 2013; Car, Stenson and Hartcher, 2014) and ProvStore (Huynh and Moreau, 2015). Similarities between PROMS and Provena include: a) registration of provenance via APIs via client software; b) persistent identifiers are used for the creation of provenance records for model workflows including the registration of datasets; and c) datasets are minted identifiers with associated metadata and linked in the provenance records (PROMS features an additional Data ID system (DID) however, implementations differ between that and the Provena Data Store). A difference between Provena and PROMS is that PROMS combines the idea of a registry and query engine implemented as a Resource Description Framework (RDF) triple store, whereas the Provena system separates the registration function from the querying capability as a separate component in the architecture. The advantage is a level of flexibility for future Provena implementations which are able to replace the query engine depending on use cases, and reindex registered items from the Provena Registry, while maintaining a separation of concern regarding the point of truth for registered records.

The Provena system is not coupled to any specific workflow engine or domain specific modelling software. This allows provenance records to be captured across systems via adapters in existing workflow engines or modelling software, e.g. Apache Airflow, Snakemake, and Galaxy. We propose future work in exploring the development of adapters and integrations in a selection of these workflow management systems to enable workflow provenance capture using the Provena system.

6. CONCLUSIONS

Workflow provenance is critical to enable process and outcome transparency, build trust with users accessing information used in decision making, and increase potential reproducibility of data via repeatable model runs and analysis. We presented the Provena system and architecture that allows research modelling and simulation workflow provenance to be captured in a standardised manner. The Provena system facilitates cross-system provenance information to be shared across different research teams. Based on use cases from the RRAP M&DS program, we presented a design and implementation in the Provena system to facilitate provenance capture and querying, which allows users to inspect, discover and understand data lineage and data usage within the set of registered items across a set of workflow provenance records. This functionality provides users with the ability to explore registered workflow provenance across teams and navigate to relevant information. We propose several avenues for future work, such as developing Provena system integrations with existing workflow management systems (e.g. Airflow, Galaxy, Flyte) and extending applications in other domains beyond reef restoration and adaptation modelling and simulation workflows.

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.

REFERENCES

Belhajjame, K. *et al.* (2015) 'Using A Suite Of Ontologies For Preserving Workflow-centric Research Objects', *Journal of Web Semantics*, 32, pp. 16–42. Available at: https://doi.org/10.1016/j.websem.2015.01.003.

Bieliauskas, S. *et al.* (2022) 'prov-db-connector'. Zenodo. Available at: https://doi.org/10.5281/zenodo.5821531.

Car, N.J. (2013) 'A Method And Example System For Managing Provenance Information In A Heterogeneous Process Environment', In *Proc. 20th Intl. Congress Modelling & Simulation (MODSIM)*. Adelaide, Australia.

Car, N.J., Stenson, M.P. and Hartcher, M. (2014) 'A Provenance Methodology And Architecture For Scientific Projects Containing Automated And Manual Processes', In *Proc. 13th Intl. Conf. Hydroinformatics*.

'Data Provenance' (2022). ARDC. https://ardc.edu.au/resource/data-provenance/ (Accessed: 22 November 2022).

Garijo, D. and Gil, Y. (2011) 'A New Approach For Publishing Workflows: Abstractions, Standards, And Linked Data', in *WORKS '11: Proc. 6th workshop on Workflows in support of large-scale science*, pp. 47–56.

Huynh, T.D. and Moreau, L. (2015) 'ProvStore: A Public Provenance Repository', in *5th Intl. Provenance and Annotation Workshop, IPAW 2014, Cologne, Germany, June 9-13, 2014.* Springer, pp. 275–277.

Lebo, T., Sahoo, S. and McGuinness, D.L. (2013) *PROV-O: The PROV Ontology*. W3C Recommendation. Cambridge, Mass. USA: World Wide Web Consortium. Available at: http://www.w3.org/TR/prov-o/.

Ludäscher, B. (2016) 'A Brief Tour Through Provenance In Scientific Workflows And Databases', in *Building trust in information: Perspectives on the frontiers of provenance*. Springer, pp. 103–126.