Australia's National Science Agency

Enabling model and simulation provenance and data transparency with ***PROVENA***

Jonathan Yu, Peter Baker, Ross Petridis, Linda Thomas, Parth Kulkarni, Peter Fitch, Sharon Tickell, Xinyu Hou

eResearch 2024

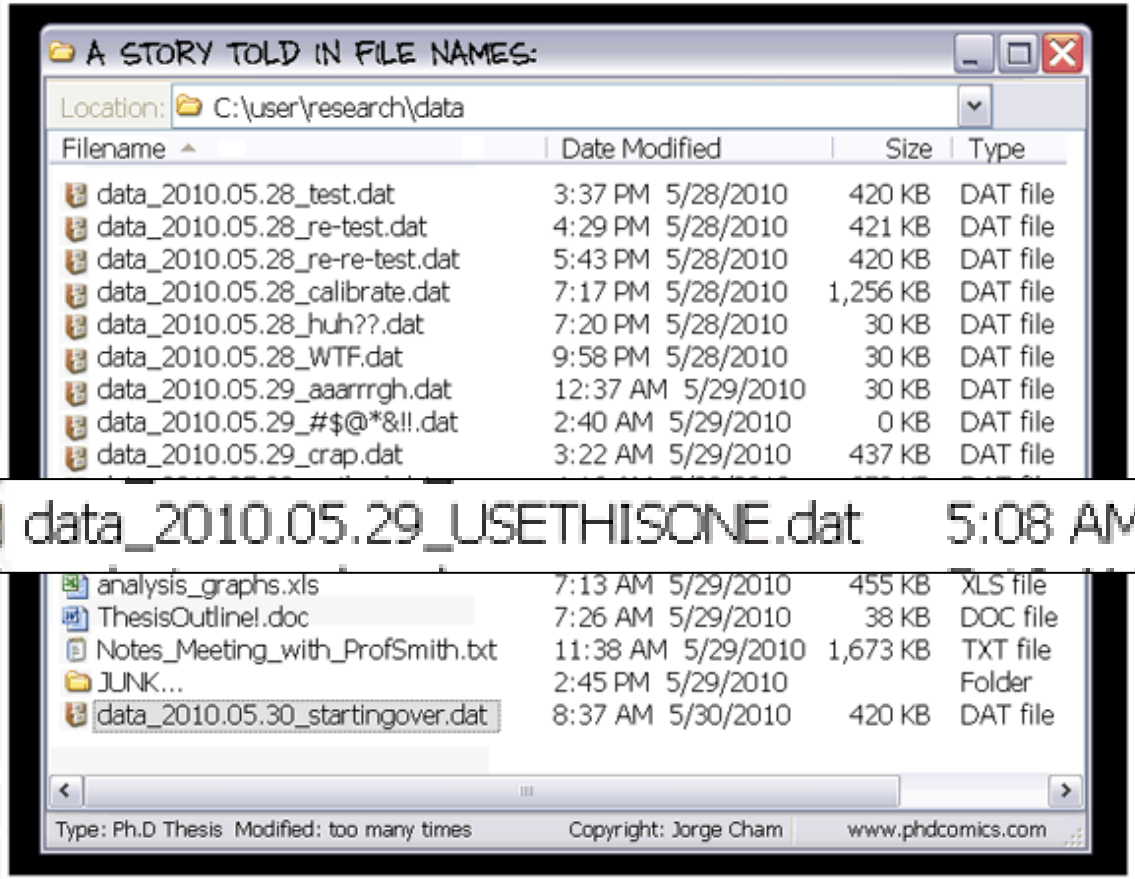
REEF RESTORATION & ADAPTATION PROGRAM

IN PARTNERSHIP

National Bushfire Intelligence Capability

Funded by the Australian Climate Service, led by CSIRO and supported by the National Emergency Management Agency

Where did this data come from?



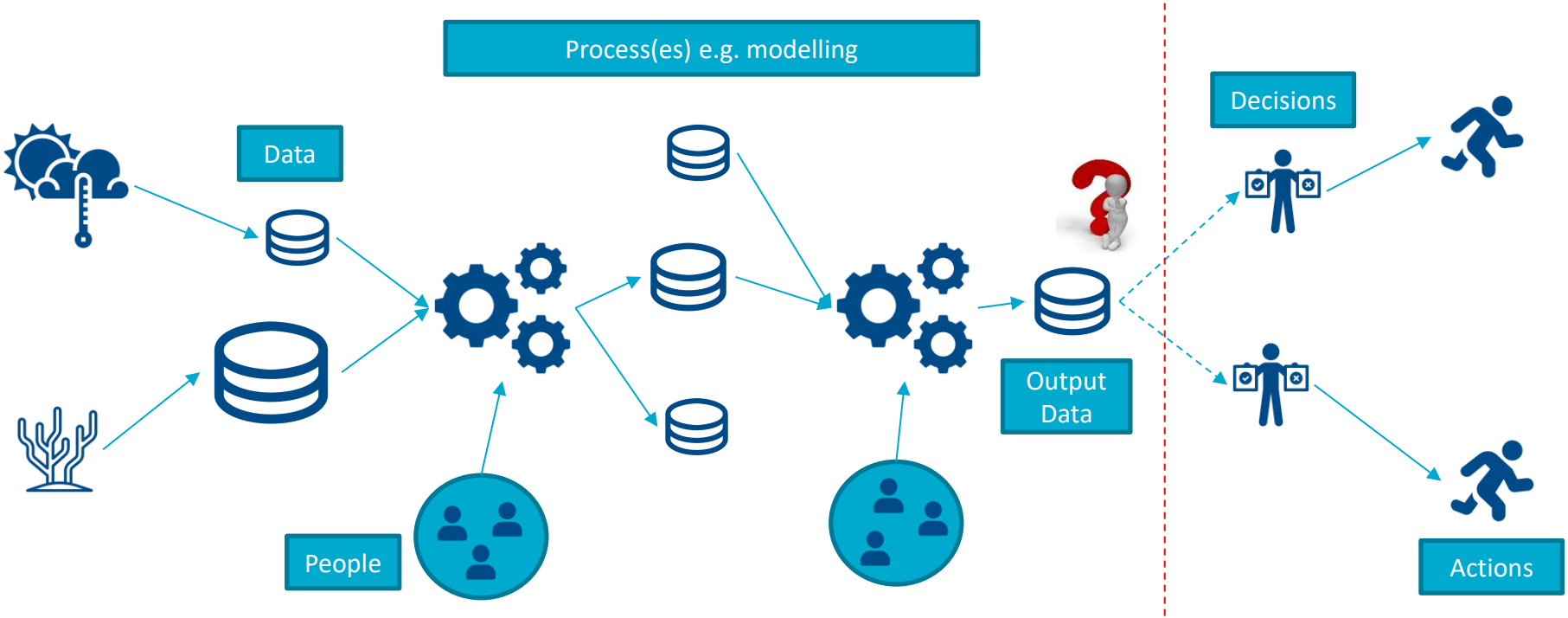
https://phdcomics.com/comics/archive.php?comicid=1323

Where did this data come from?

Trust

Reproducibility, verifiability

"***Provenance*** refers to the sources of information, such as entities and processes, involved in producing or delivering an artifact“ (W3C, Prov)



Reef Restoration and Adaptation Program Modelling and Decision Support (MDS) for the Great Barrier Reef

RRAP MDS Modelling

Modeller

Stakeholders
(e.g. RRAP Program, Reef Managers and Industry Partners)

Decision Maker

Researcher

Fogging

Heat tolerant corals

Rubble stabilisation

Produce data/info/knowledge through modelling

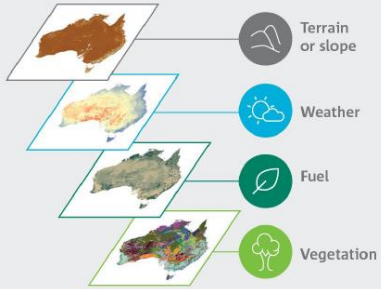
Make decisions

REEF RESTORATION & ADAPTATION PROGRAM

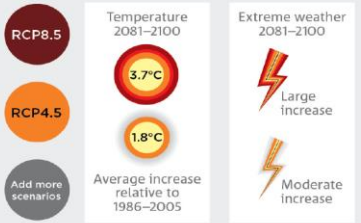
https://gbrrestoration.org/program/modelling-and-decision-support/

National Bushfire Intelligence Capability

1 For each factor, NBIC generates input data layers or users can provide their own



2 Users choose variables to be applied to the modelling and analyses



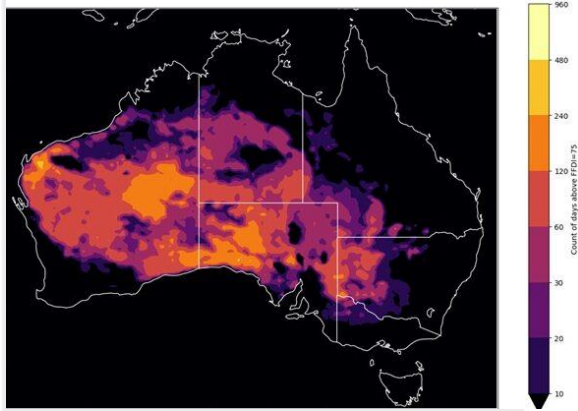
NBIC

repeatable, modular analyses

cloud-computing environments

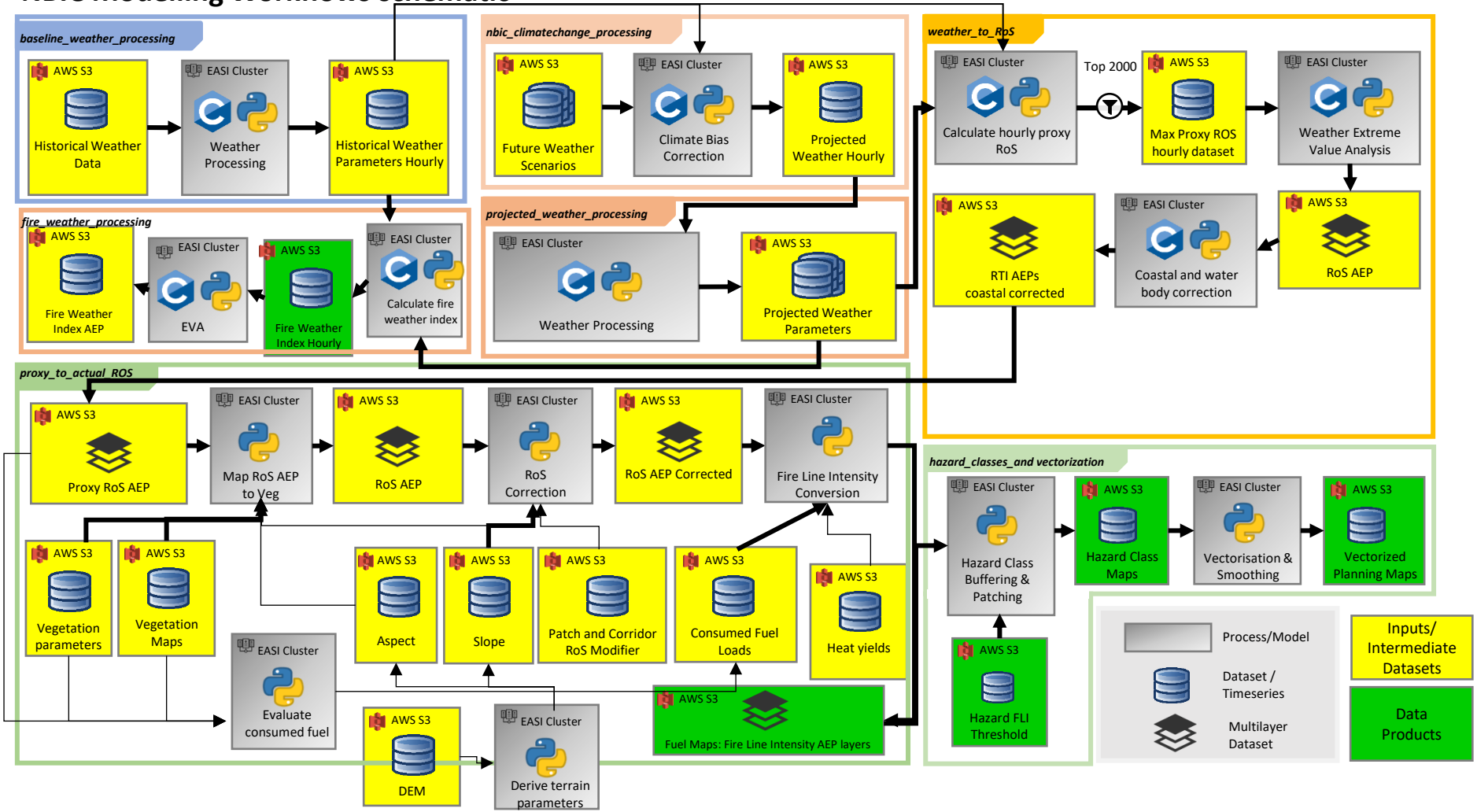
research, innovation and **integration**

3 Bushfire hazard data



https://research.csiro.au/nbic

NBIC Modelling Workflows schematic



REEF RESTORATION & ADAPTATION PROGRAM

National Bushfire Intelligence Capability

People

Modelling approach

Data formats

Why?

Modelling and Decision Support

7 organisations

5 modelling teams, ~35 people total
(25 modellers)

Rigour | Credibility | Transparency

CSIRO

1 organisation

7 sub-teams, ~40 people total
(14 modellers)

Data quality | Communication | Understand dependencies

Data Provenance Challenges

1

Registering, (storing) and accessing data

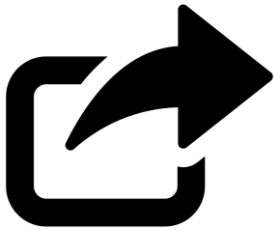
2

Describing the data and its provenance – human and machine readable

3

Finding / querying provenance info (data, model, workflows…)

4

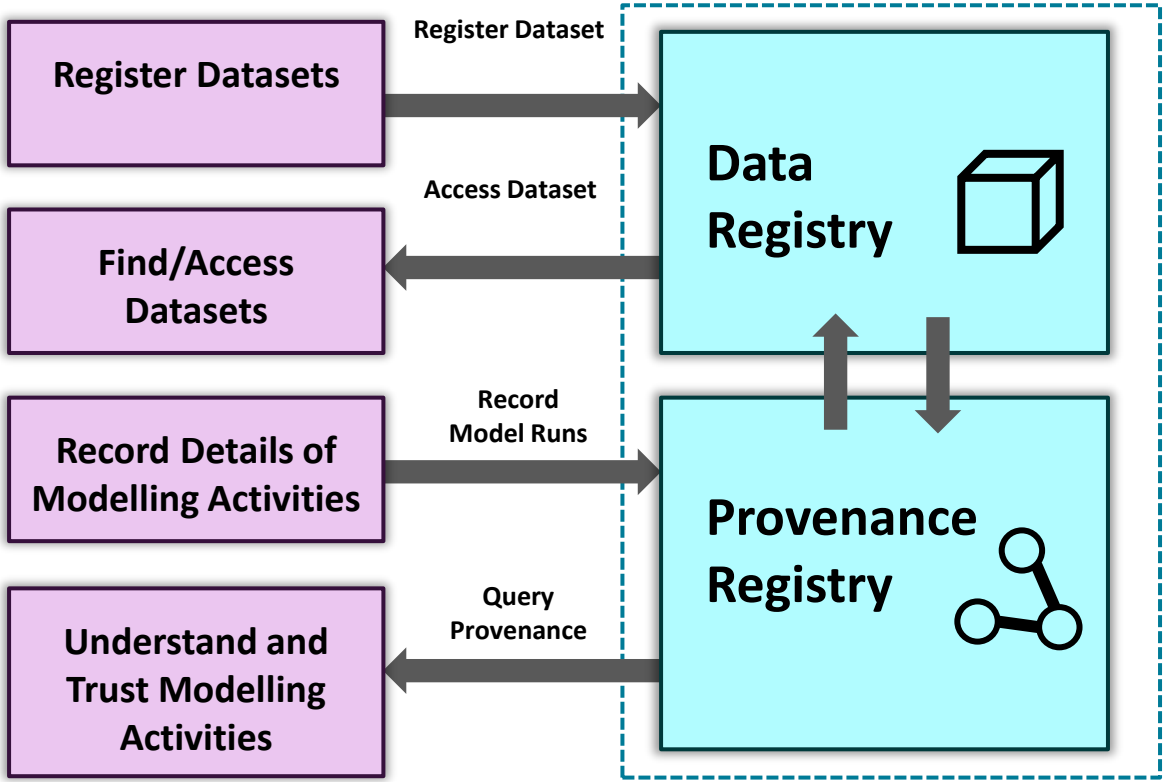


Sharing data/info (with the right people)

5

Integrating this into modelling and simulation processes

High-level Design of a Provenance solution



Some data provenance solutions to date

Embed provenance in a file naming system…

data_2010.05.29_USETHISONE.dat

★☆☆☆☆

Record model runs, config, datasets in Excel

★★☆☆☆

Use a workflow engine

nextflow Snakemake Apache Airflow

★★★☆☆

Write (more) docs about our process

★☆☆☆☆

Let's make copies of inputs, code, and store it with the model outputs

★★★☆☆

Embed the provenance metadata in or with the dataset

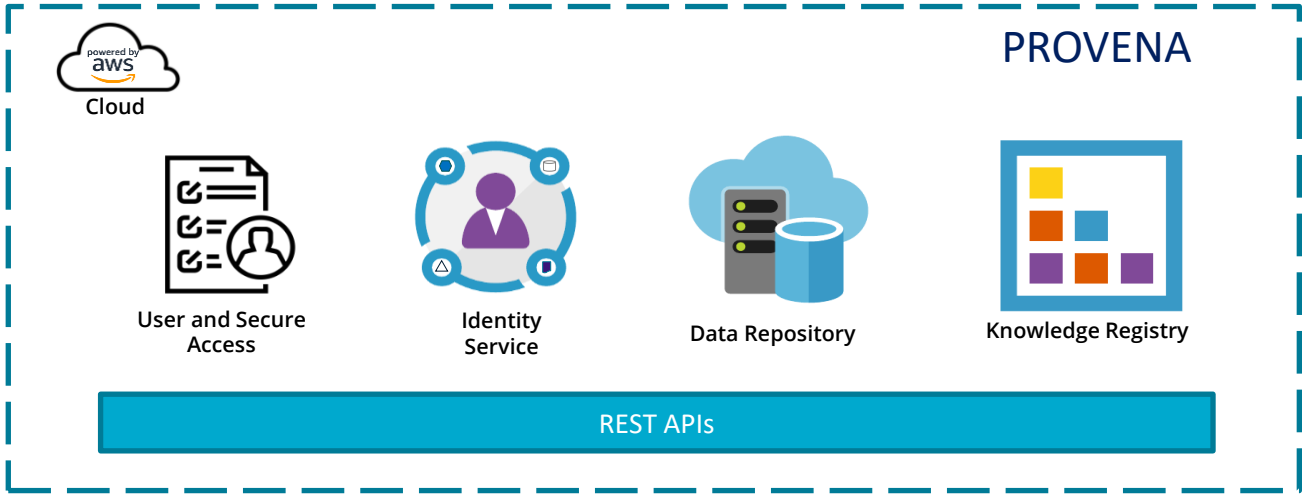
{ }

★★★★☆

Provena – a solution to provenance and data challenges

PROVENA

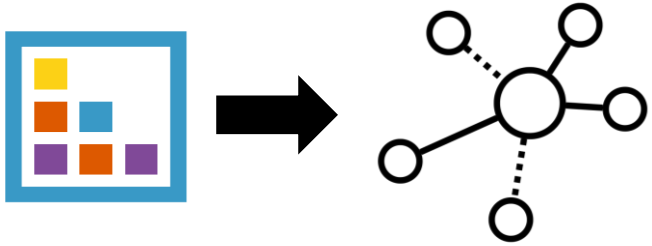
- Cloud-based provenance management system for projects
- Provenance of datasets and details of related modelling and simulation workflows
- (Currently) AWS-based solution



https://provena.io/

CSIRO

The Provena high level features



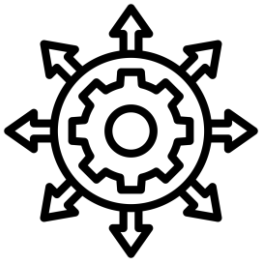
Provenance system – Scalable Registry which records all datasets, modelling activities and related metadata. Records are query-able via graph database.

PID

Leverages persistent identifier system (Handle)

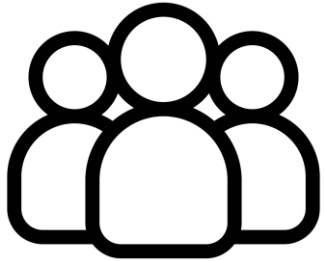
Scalable Data storage backed by AWS S3 with user friendly security implementation

User logins using identity providers such as Australian Access Federation



Documented, automation ready APIs for all system interactions

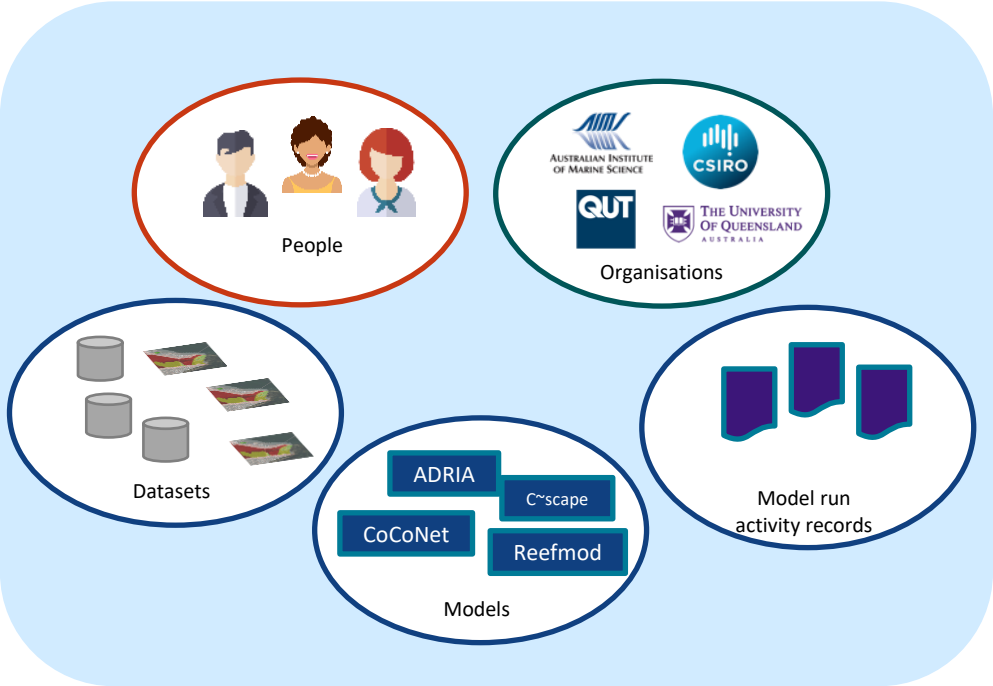
Search index built in



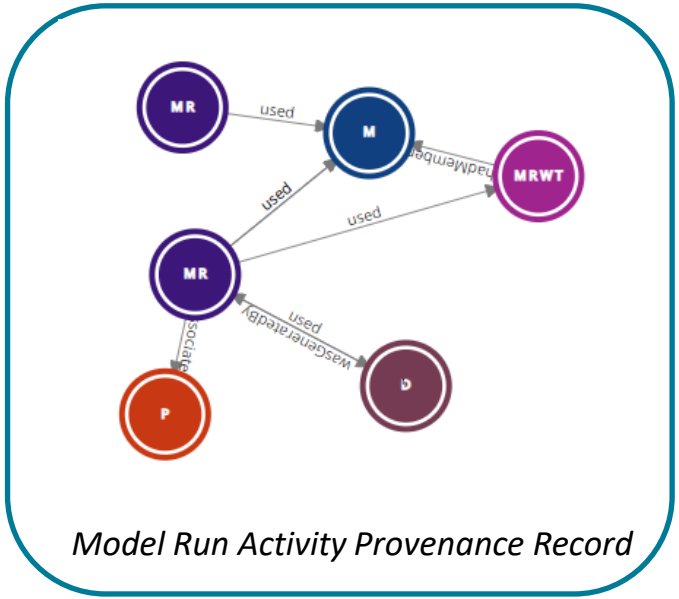
Open source and ready to use operationally

The Provena Registry

The ***Registry*** is a centralised location to register, update, explore and share *persistently identified* resources



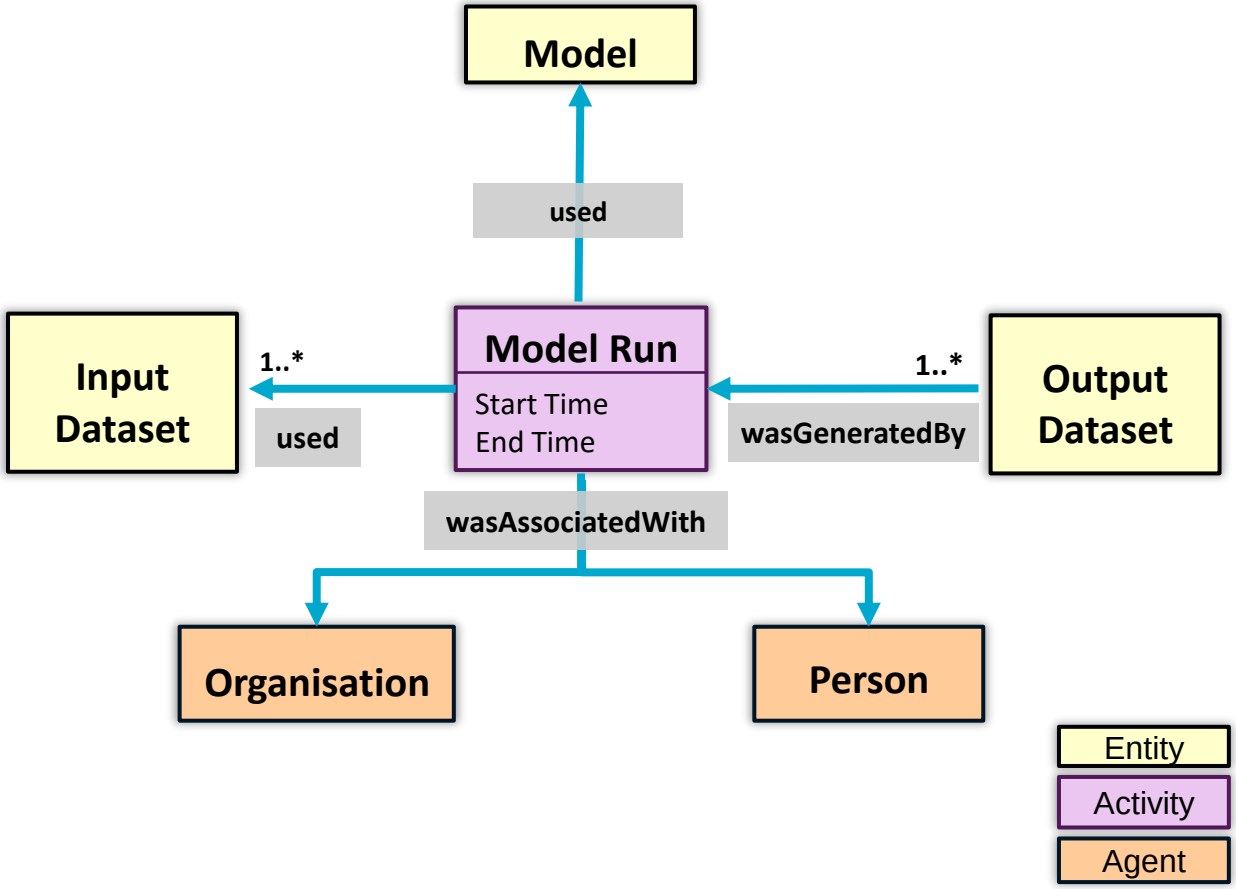
… enables linking things together and the creation of provenance records

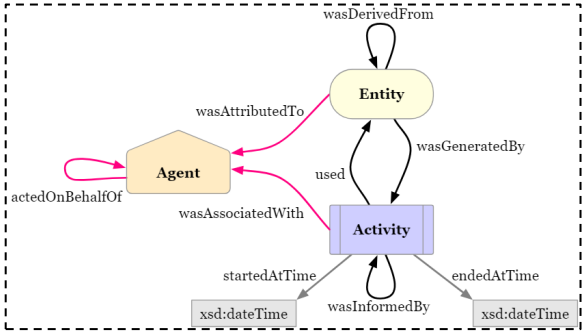


Model Run Activity Provenance Record

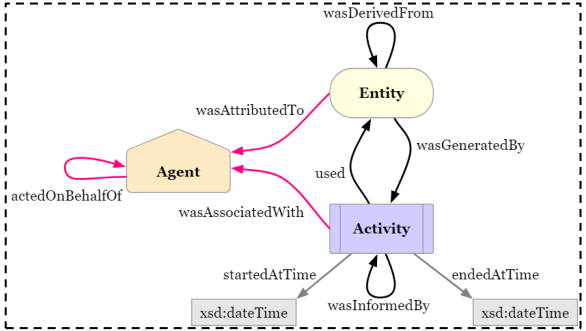
Capturing provenance

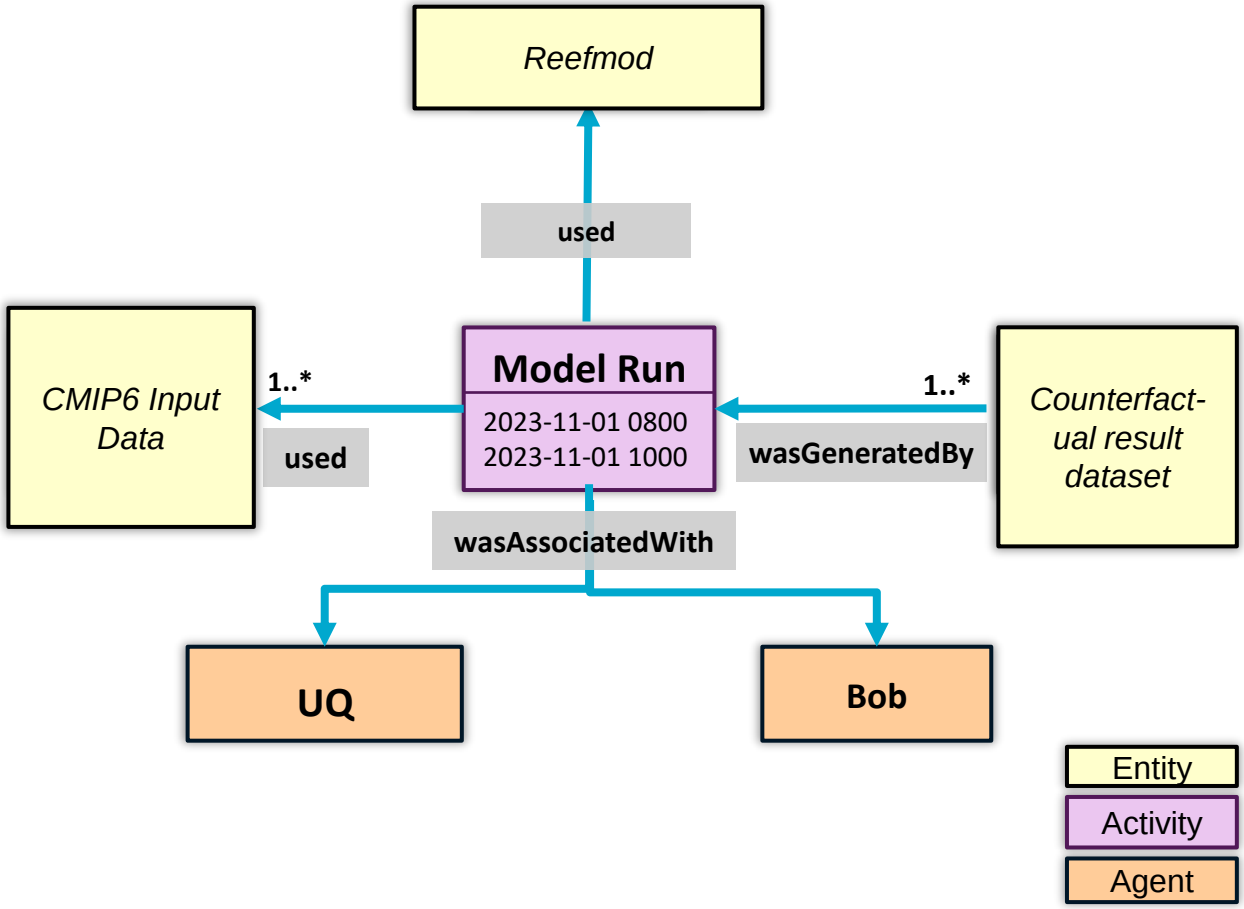
Model Run Data Model (Prov-O extension)





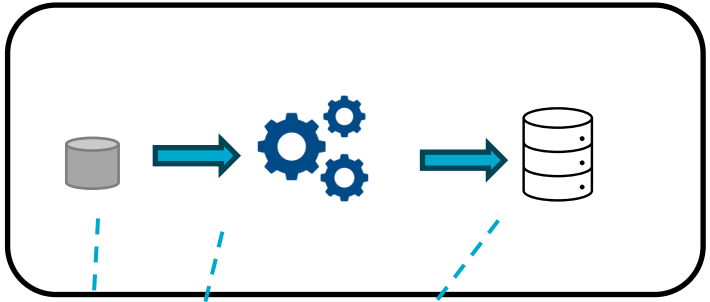
Model Run Data Model (Prov-O extension)





Modelling workflow

Modelling team



Register input datasets

Register model runs

Register output datasets

Modelling team

Find and access datasets

Register / Update dataset record and model run provenance

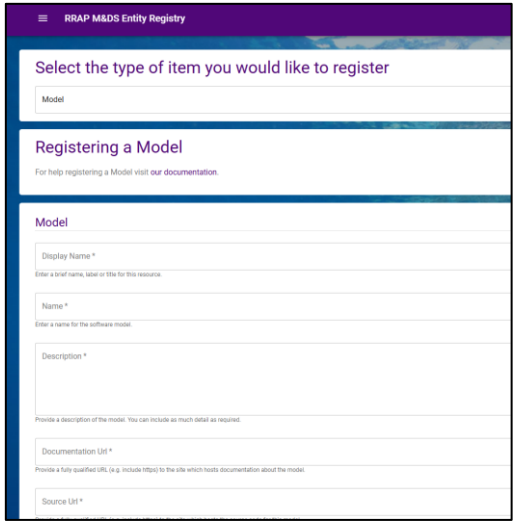
Lookup provenance info

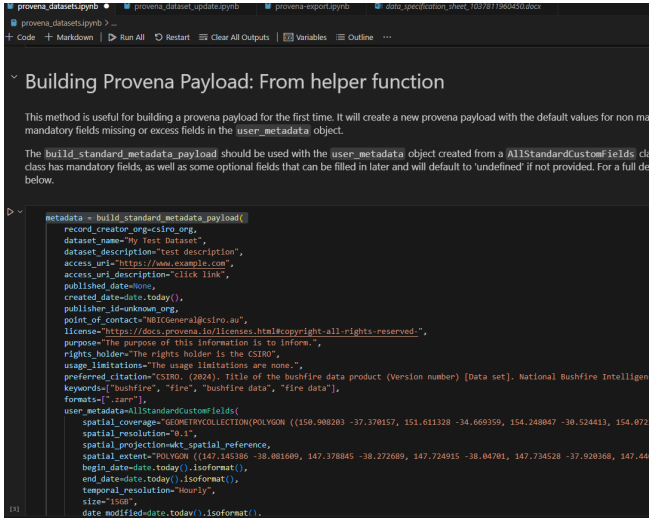
Project team

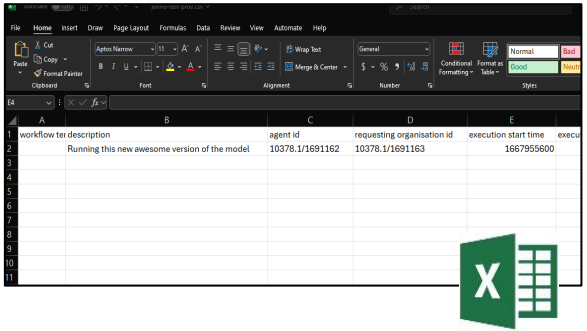
Query info for reporting

Provena

Record model runs, config, datasets in Excel



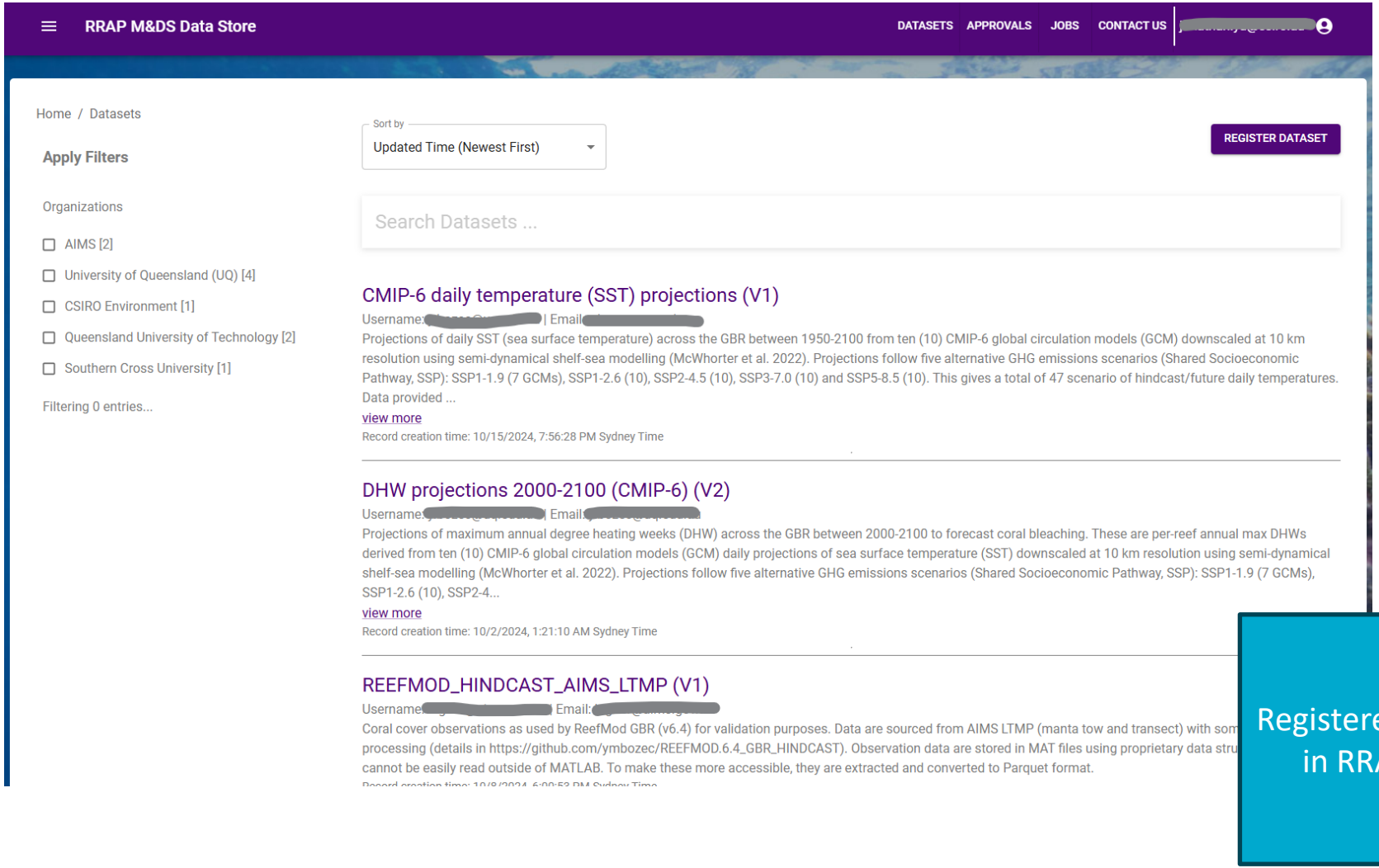


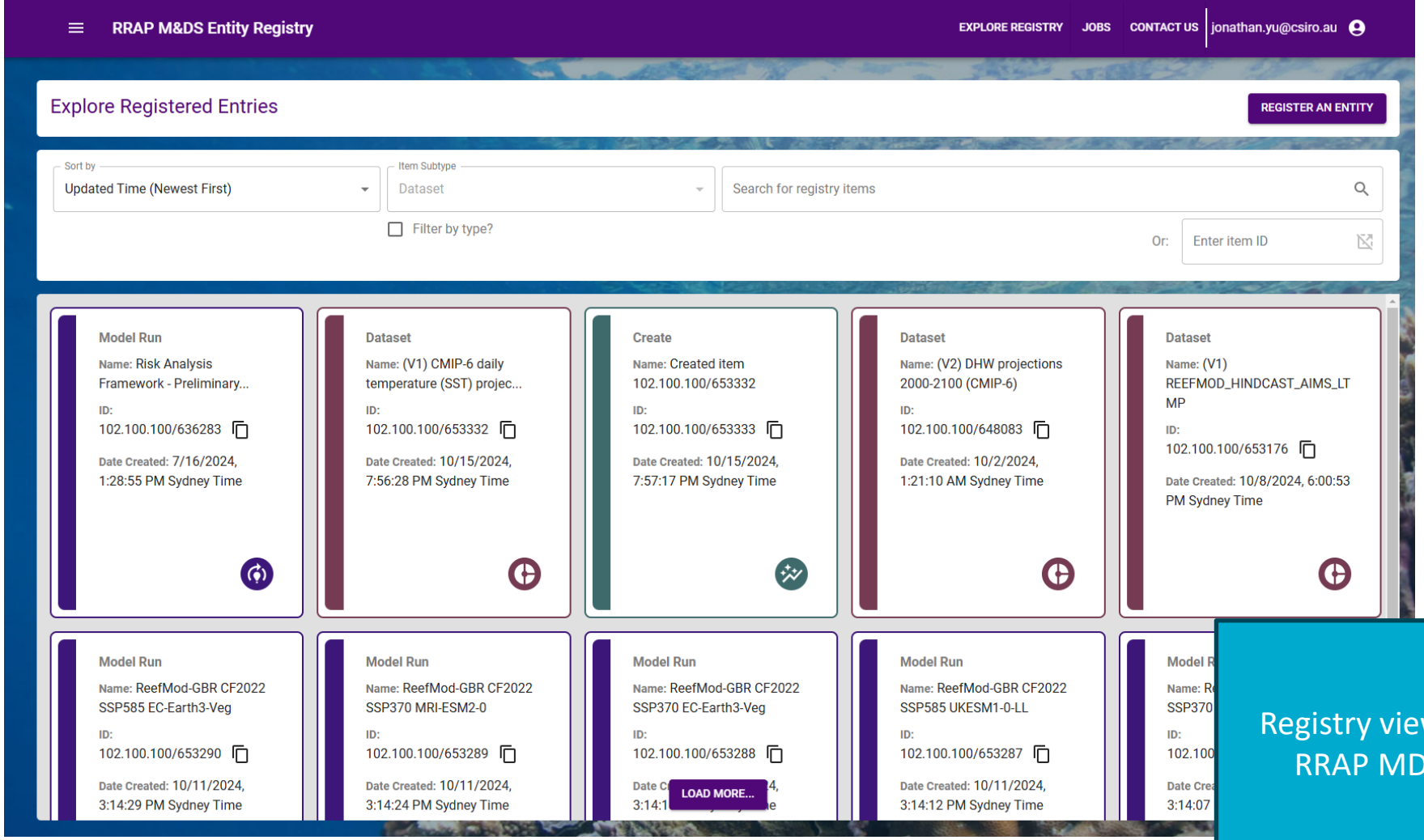


Register datasets, entities, model runs via **Web Browser**

Register datasets, entities and model runs via **Python client**

Register model runs via **Excel**


RRAP M&DS Data Store
DATASETS
APPROVALS
JOBS
CONTACT US
Home / Datasets
Apply Filters
Organizations
AIMS [2]
University of Queensland (UQ) [4]
CSIRO Environment [1]
Queensland University of Technology [2]
Southern Cross University [1]
Filtering 0 entries...
Sort by
Updated Time (Newest First)
REGISTER DATASET
Search Datasets ...
CMIP-6 daily temperature (SST) projections (V1)
Projections of daily SST (sea surface temperature) across the GBR between 1950-2100 from ten (10) CMIP-6 global circulation models (GCM) downscaled at 10 km resolution using semi-dynamical shelf-sea modelling (McWhorter et al. 2022). Projections follow five alternative GHG emissions scenarios (Shared Socioeconomic Pathway, SSP): SSP1-1.9 (7 GCMs), SSP1-2.6 (10), SSP2-4.5 (10), SSP3-7.0 (10) and SSP5-8.5 (10). This gives a total of 47 scenario of hindcast/future daily temperatures. Data provided ...
view more
Record creation time: 10/15/2024, 7:56:28 PM Sydney Time
DHW projections 2000-2100 (CMIP-6) (V2)
Projections of maximum annual degree heating weeks (DHW) across the GBR between 2000-2100 to forecast coral bleaching. These are per-reef annual max DHWs derived from ten (10) CMIP-6 global circulation models (GCM) daily projections of sea surface temperature (SST) downscaled at 10 km resolution using semi-dynamical shelf-sea modelling (McWhorter et al. 2022). Projections follow five alternative GHG emissions scenarios (Shared Socioeconomic Pathway, SSP): SSP1-1.9 (7 GCMs), SSP1-2.6 (10), SSP2-4...
view more
Record creation time: 10/2/2024, 1:21:10 AM Sydney Time
REEFMOD_HINDCAST_AIMS_LTMP (V1)
Coral cover observations as used by ReefMod GBR (v6.4) for validation purposes. Data are sourced from AIMS LTMP (manta tow and transect) with som
processing (details in https://github.com/ymbozec/REEFMOD.6.4_GBR_HINDCAST). Observation data are stored in MAT files using proprietary data stru
cannot be easily read outside of MATLAB. To make these more accessible, they are extracted and converted to Parquet format.
Registered datasets in RRAP MDS



Registry view in RRAP MDS

Using provenance info

Explore Registered Entries

REGISTER AN ENTITY

Sort by: Updated Time (Newest First)

Item Subtype: Model

Search for registry items

Filter by type?

Or: Enter item ID

Model
Name: (V1) REEFMOD.6.8_GBR
ID: 102.100.100/483650
Date Created: 3/14/2023, 2:29:49 PM Sydney Time

Model
Name: (V1) REEFMOD.7.0_GBR
ID: 102.100.100/648072
Date Created: 10/1/2024, 12:53:34 PM Sydney Time

Model
Name: (V1) ADRIA_10_02_2022_alpha
ID: 102.100.100/636122
Date Created: 7/8/2024, 10:08:42 AM Sydney Time

Model
Name: (V2) CoCoNET
ID: 102.100.100/637875
Date Created: 9/18/2024, 11:29:36 AM Sydney Time

Model
Name: (V1) C~scape - version 3.0.4
ID: 102.100.100/637651
Date Created: 9/10/2024, 12:31:25 PM Sydney Time

Model
Name: (V2) ADRIA Model
ID: 102.100.100/636700
Date Created: 8/1/2024, 3:54:10 PM Sydney Time

Model
Name: (V1) Peter's Test Model
ID: 102.100.100/489364
Date Created: 8/28/2023,

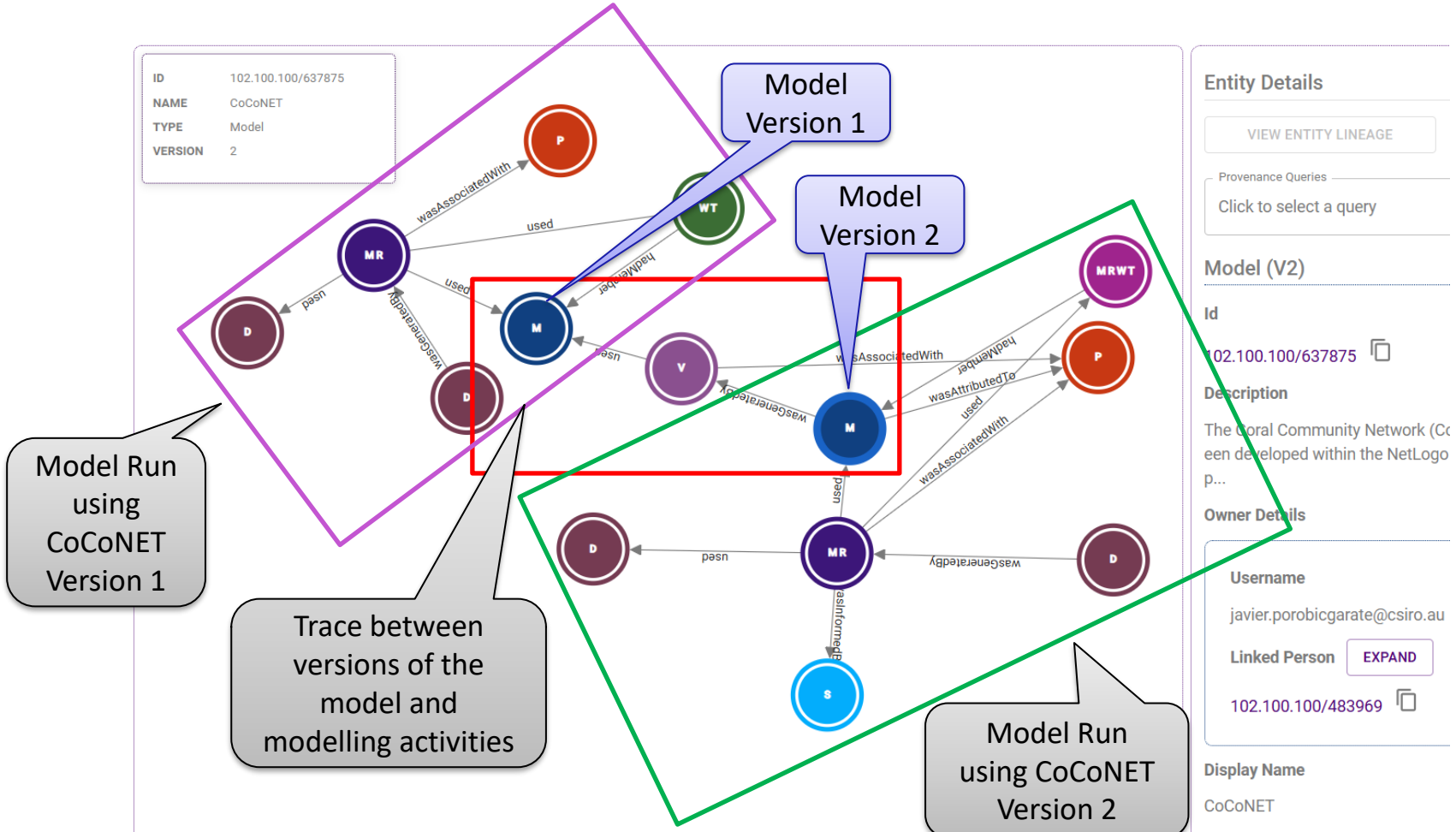
Model
Name: (V1) ReefModEngine.jl
ID: 102.100.100/488715
Date Created: 8/11/2023,

Model
Name: (V1) ReefMod Engine intervention model
ID: 102.100.100/487570

Model
Name: (V1) Production ReefMod intervention reef ...
ID: 102.100.100/487541

LOAD MORE...

Understand modelling activities



Provena features

1

Register, (store) and access data

2

Describe the data and its provenance – human and machine readable

3

Find / query provenance info (data, model, workflows…)

4

Share data/info (with the right people)

5

Integrate this into modelling and simulation processes

Provena can't do

1

Register**-ing**, (store) and access data

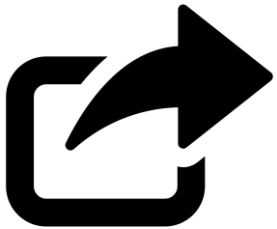
2

Describ**-ing** the data and its provenance

3

Find**-ing** / query**-ing** provenance info (data, model, workflows…)

4



Shar**-ing** data/info (with the right people)

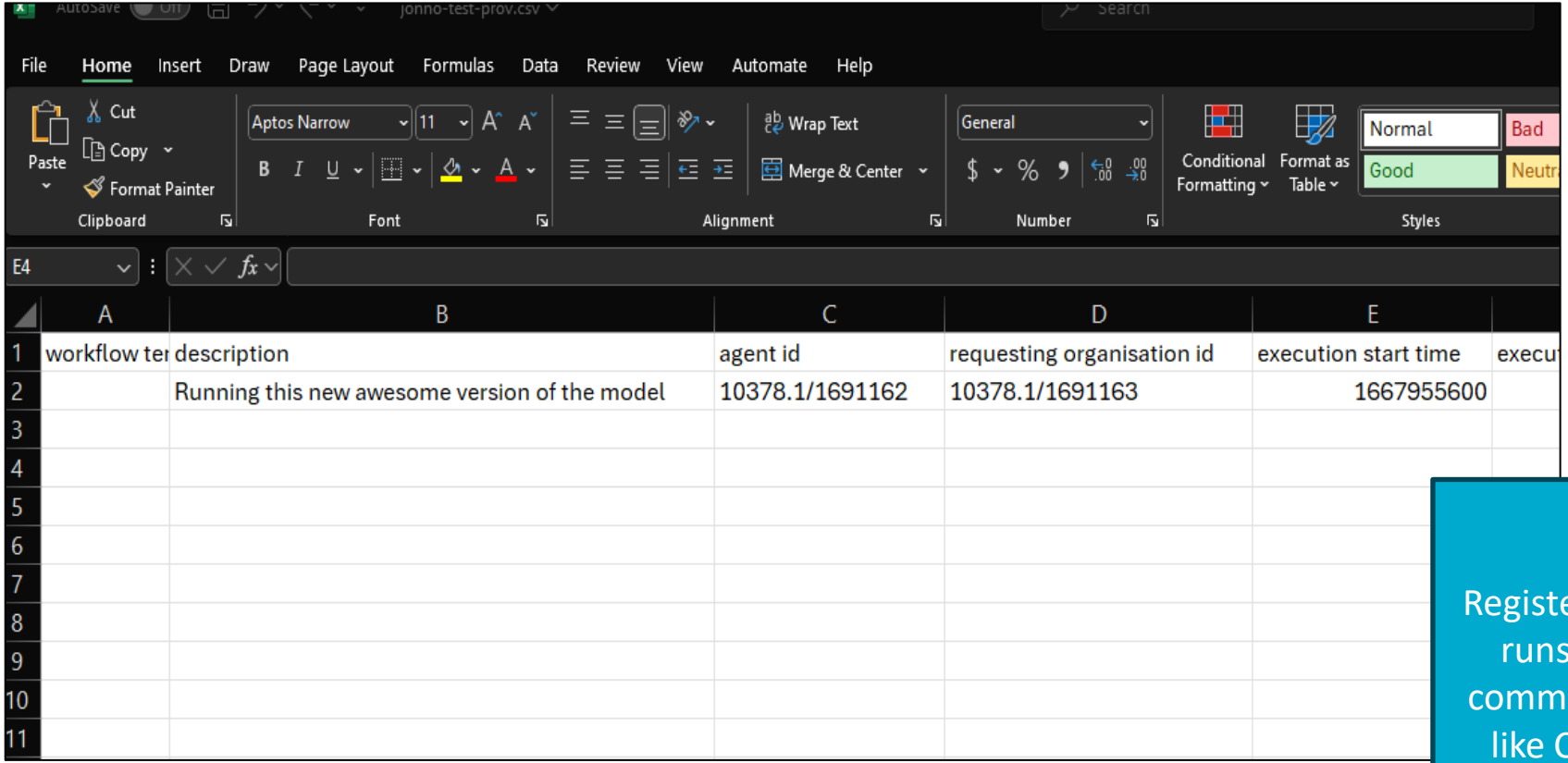
5

Integrat**-ing** this into modelling and simulation processes

Relies on people


I WOULD REGISTER
PROVENANCE DATA ALL DAY
imgflip.com


YOU WANT ME TO REGISTER PROVENANCE?
imgflip.com
TOO HARD... AINT NOBODY GOT TIME FOR THAT...



Register model runs using common tools like CSV file uploads

provena_datasets.ipynb ● provena_dataset_update.ipynb provena-export.ipynb data_specification_sheet_1037811960450.docx

provena_datasets.ipynb > ...

+ Code + Markdown | Run All Restart Clear All Outputs | Variables Outline ···

Building Provena Payload: From helper function

This method is useful for building a provena payload for the first time. It will create a new provena payload with the default values for non ma
mandatory fields missing or excess fields in the `user_metadata` object.

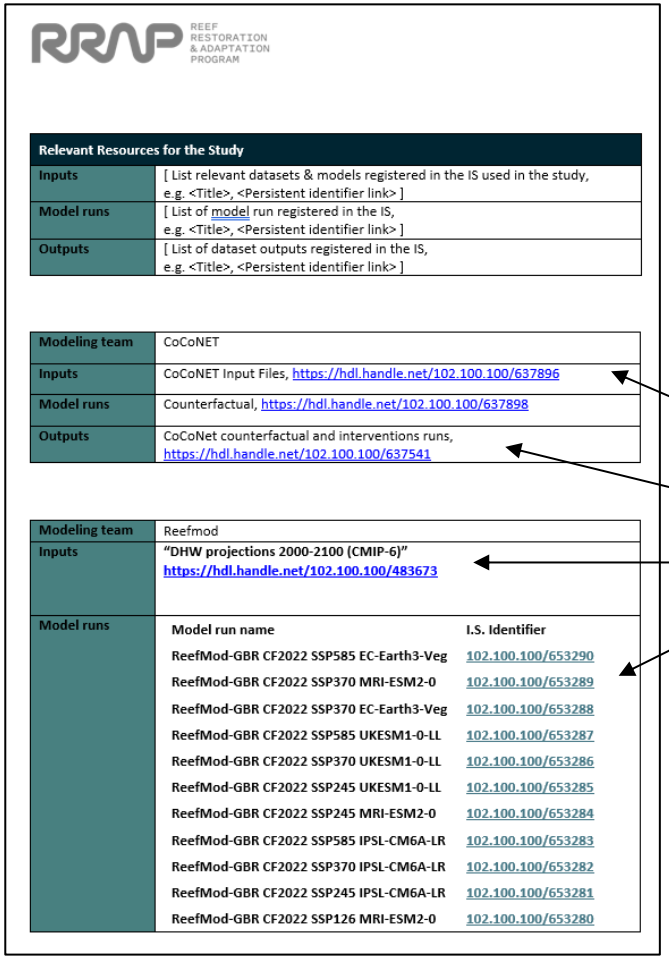
The `build_standard_metadata_payload` should be used with the `user_metadata` object created from a `AllStandardCustomFields` cl
class has mandatory fields, as well as some optional fields that can be filled in later and will default to 'undefined' if not provided. For a full de
below.

```
metadata = build_standard_metadata_payload(
    record_creator_org=csiro_org,
    dataset_name="My Test Dataset",
    dataset_description="test description",
    access_uri="https://www.example.com",
    access_uri_description="click link",
    published_date=None,
    created_date=date.today(),
    publisher_id=unknown_org,
    point_of_contact="NBICGeneral@csiro.au",
    license="https://docs.provena.io/licenses.html#copyright-all-rights-reserved-",
    purpose="The purpose of this information is to inform.",
    rights_holder="The rights holder is the CSIRO",
    usage_limitations="The usage limitations are none.",
    preferred_citation="CSIRO. (2024). Title of the bushfire data product (Version number) [Data set]. National Bushfire Intelligen
    keywords=["bushfire", "fire", "bushfire data", "fire data"],
    formats=[".zarr"],
    user_metadata=AllStandardCustomFields(
        spatial_coverage="GEOMETRYCOLLECTION(POLYGON ((150.908203 -37.370157, 151.611328 -34.669359, 154.248047 -30.524413, 154.072
        spatial_resolution="0.1",
        spatial_projection=wkt_spatial_reference,
        spatial_extent="POLYGON ((147.145386 -38.081609, 147.378845 -38.272689, 147.724915 -38.04701, 147.734528 -37.920368, 147.44
        begin_date=date.today().isoformat(),
        end_date=date.today().isoformat(),
        temporal_resolution="Hourly",
        size="15GB",
        date_modified=date.today().isoformat(),
```

[3]

python™

Register datasets and model runs from python environments



REEF RESTORATION & ADAPTATION PROGRAM

Relevant Resources for the Study	
Inputs	[List relevant datasets & models registered in the IS used in the study, e.g. <Title>, <Persistent identifier link>]
Model runs	[List of model run registered in the IS, e.g. <Title>, <Persistent identifier link>]
Outputs	[List of dataset outputs registered in the IS, e.g. <Title>, <Persistent identifier link>]

Modeling team	CoCoNET
Inputs	CoCoNET Input Files, https://hdl.handle.net/102.100.100/637896
Model runs	Counterfactual, https://hdl.handle.net/102.100.100/637898
Outputs	CoCoNet counterfactual and interventions runs, https://hdl.handle.net/102.100.100/637541

Modeling team	Reefmod	
Inputs	**"DHW projections 2000-2100 (CMIP-6)"** **https://hdl.handle.net/102.100.100/483673**	
Model runs	**Model run name**	**I.S. Identifier**
	ReefMod-GBR CF2022 SSP585 EC-Earth3-Veg	**102.100.100/653290**
	ReefMod-GBR CF2022 SSP370 MRI-ESM2-0	**102.100.100/653289**
	ReefMod-GBR CF2022 SSP370 EC-Earth3-Veg	**102.100.100/653288**
	ReefMod-GBR CF2022 SSP585 UKESM1-0-LL	**102.100.100/653287**
	ReefMod-GBR CF2022 SSP370 UKESM1-0-LL	**102.100.100/653286**
	ReefMod-GBR CF2022 SSP245 UKESM1-0-LL	**102.100.100/653285**
	ReefMod-GBR CF2022 SSP245 MRI-ESM2-0	**102.100.100/653284**
	ReefMod-GBR CF2022 SSP585 IPSL-CM6A-LR	**102.100.100/653283**
	ReefMod-GBR CF2022 SSP370 IPSL-CM6A-LR	**102.100.100/653282**
	ReefMod-GBR CF2022 SSP245 IPSL-CM6A-LR	**102.100.100/653281**
	ReefMod-GBR CF2022 SSP126 MRI-ESM2-0	**102.100.100/653280**

Inputs, model run and output record summaries and links automatically generated

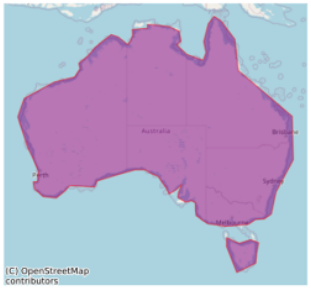
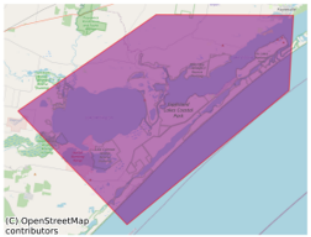
Generate a report of modelling activities and datasets

NBIC ACS Stage 2

National Bushfire Intelligence Capability (NBIC) data specification sheet

My Test Dataset

Provena dataset identifier: 10378.1/1962484

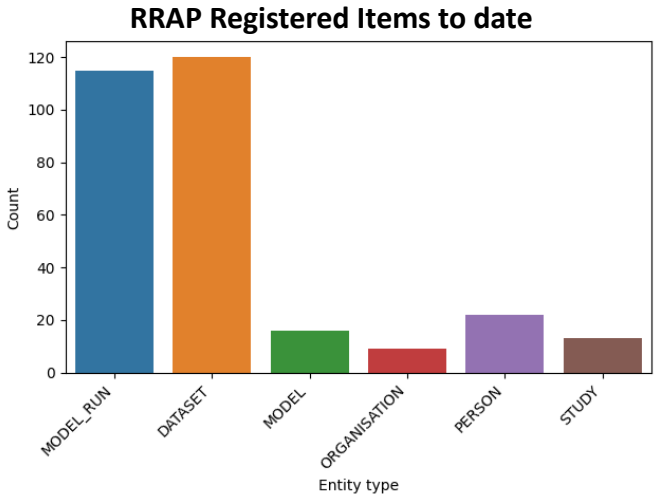
identifier	Undefined
title	My Test Dataset
author(s)	Ross Petridis
abstract	This is still a test dataset
resource purpose	The purpose of this information is to inform.
citation	CSIRO. (2024). Title of the bushfire data product (Version number) [Data set]. National Bushfire Intelligence Capability. URL
source(s)	CSIRO
licence	https://docs.provena.io/licenses.html#copyright-all-rights-reserved-
attribution	Undefined
associated resource	Undefined
additional documentation	Undefined
rights holder	The rights holder is the CSIRO
releasability	Undefined
use limitations	The usage limitations are none.
spatial projection	GEOGCS["WGS 84", DATUM["WGS_1984", SPHEROID["WGS 84",6378137,298.257223563, AUTHORITY["EPSG","7030"]], AUTHORITY["EPSG","6326"]], PRIMEM["Greenwich",0, AUTHORITY["EPSG","8901"]], UNIT["degree",0.0174532925199433, AUTHORITY["EPSG","9122"]],
	AUTHORITY["EPSG","4326"]]
spatial resolution	0.1
spatial coverage	
spatial extent	
begin date	2024-10-08
end date	2024-10-08
temporal resolution	Hourly
lineage	Undefined
browse graphic	Undefined
keyword(s)	bushfire, fire, bushfire data, fire data
resource topic	Undefined
size	15GB
format	.zarr
linkage	Undefined

Exporting Data Specification Sheets from Provena Dataset Records using an ISO-19135 template

Outcomes and Learnings

Highlighted project examples:

- National Bushfire Intelligence Capability – since Jan 2024
- RRAP MDS – since Nov 2022



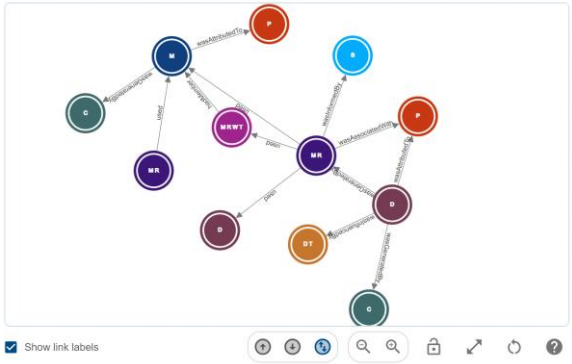
Reflections

- Provena is a tool to enable rigour in science
- Challenge - building tools and iterating with users in mind and making it as easy as possible
- Role of Provena – partnering with projects as data is produced. Opened up many interesting conversations about data reproducibility

Summary

Provena = Cloud-based provenance management system designed to support modelling and simulation activities

Enabling users to do the "basics" – mint identifiers, capture provenance (entities and processes) and support practical uses of provenance info



Provena approach and tools - general purpose and adaptable to different domains and modes of modelling – e.g. NBIC, RRAP MDS

Using Provena to record data provenance

★★★★½

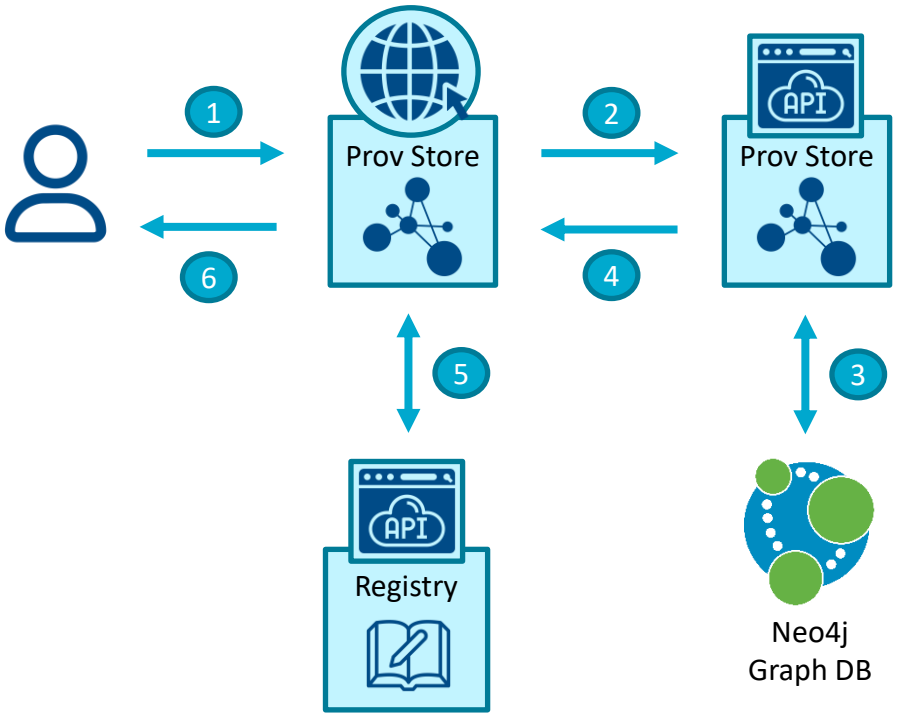
Thank you

CSIRO Environment
Jonathan Yu
jonathan.yu@csiro.au

Website **provena.io**
Github https://github.com/provena/provena
https://provena.github.io/provena-python-client/

Appendix

Provenance Query Workflow

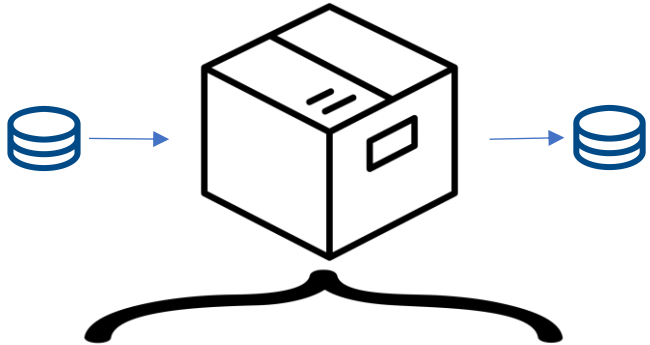


1. User interacts with graph to query provenance data
2. Explore request made to Prov API
3. Prov API runs Cypher query against Neo4j graph
4. Prov API returns synthesised result in graph format
5. Prov Store fetches item information from Registry
6. Expanded and populated graph visualised for user

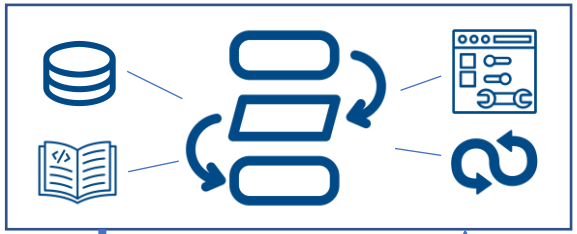
Levels of Provenance

Provena

Black Box Provenance



Workflow Level



Operating System / IO Level

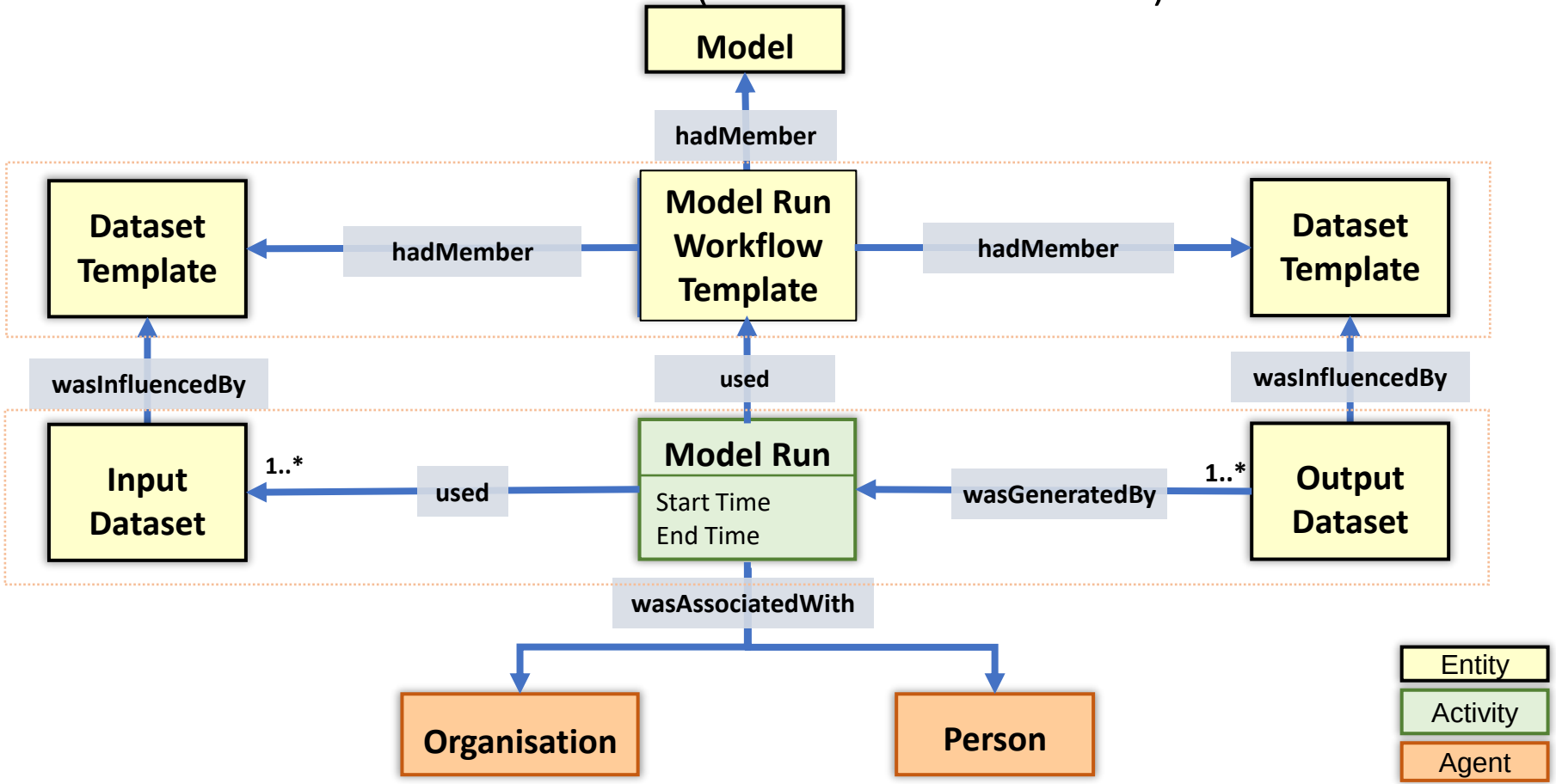
Interpretability

Domain specificity

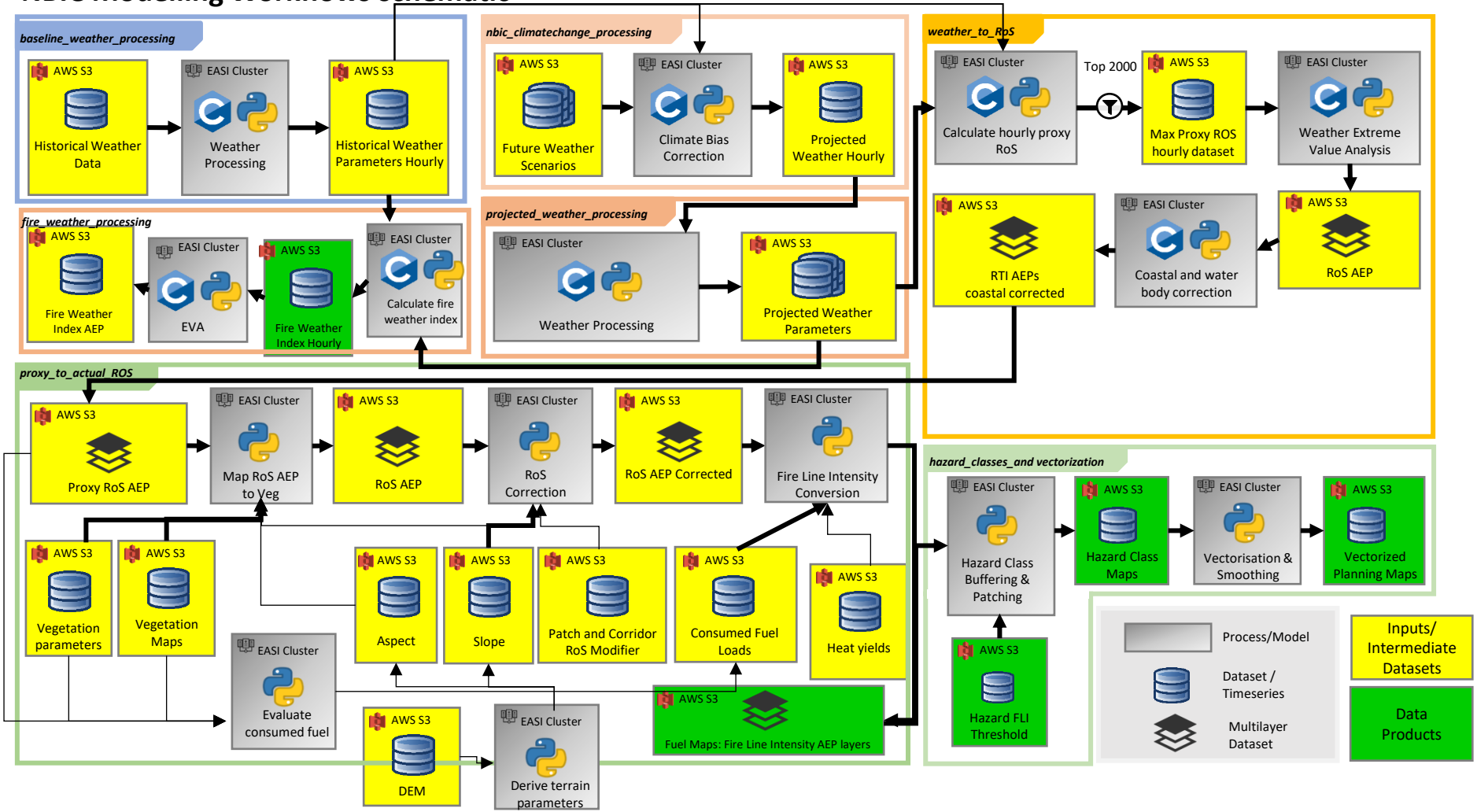
Greater Data Volume

More Detail

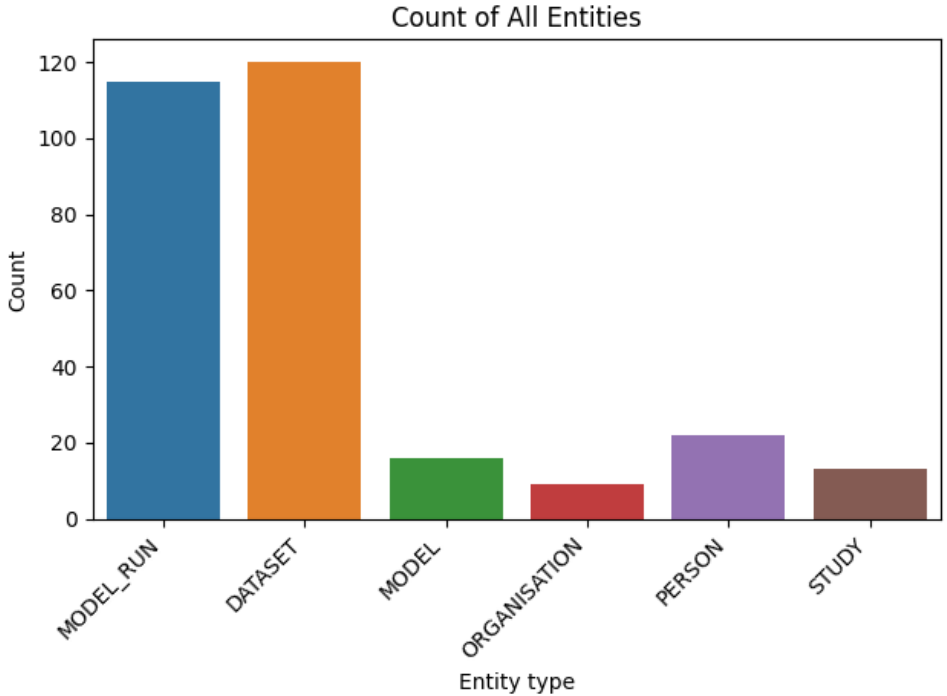
Model Run Data Model (Prov-O extension)



NBIC Modelling Workflows schematic



Data Provenance for RRAP Modelling and Decision Support

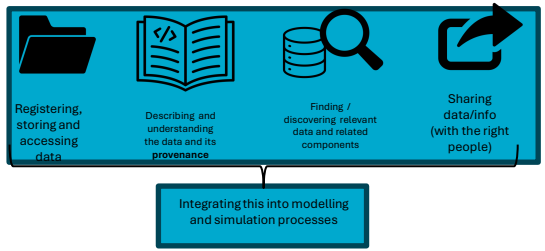


Since November 2022

Total Datasets: 25.6 TB

MODEL_RUN	115
DATASET	120
MODEL	16
ORGANISATION	9
PERSON	22
STUDY	13

NBIC



NBIC Team

Modelling team

NBIC Data stewards

NBIC System

New dataset

Store dataset

NBIC AWS S3 bucket

D

Create/update DAP record

DAP Formatted Metadata

Pre-filled Dataset Specification sheets (ISO 19115)

Update

IP Register

Register / Update dataset record and provenance

Lookup Dataset info

Provena

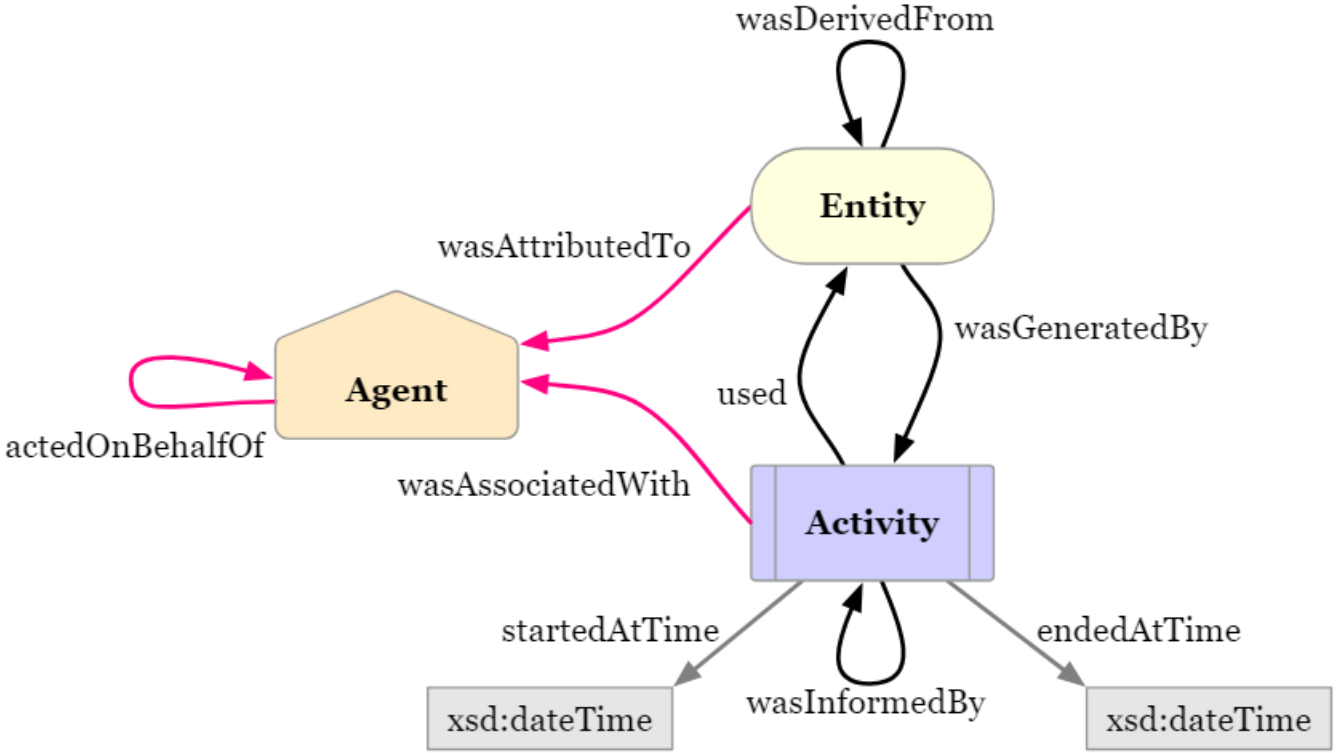
CSIRO

Data Access Portal

Download official NBIC dataset

NBIC Stakeholders

Prov-O Data Model



Candidate solution – operational modelling

DRAFT