Towards automated provenance collection for experimental runs of agent-based models

Doug Salt Gary Polhill and Corran Musk Lorenzo Milazzo Dawn Parker

The James Hutton Institute

September 5, 2023

Structure of the presentation

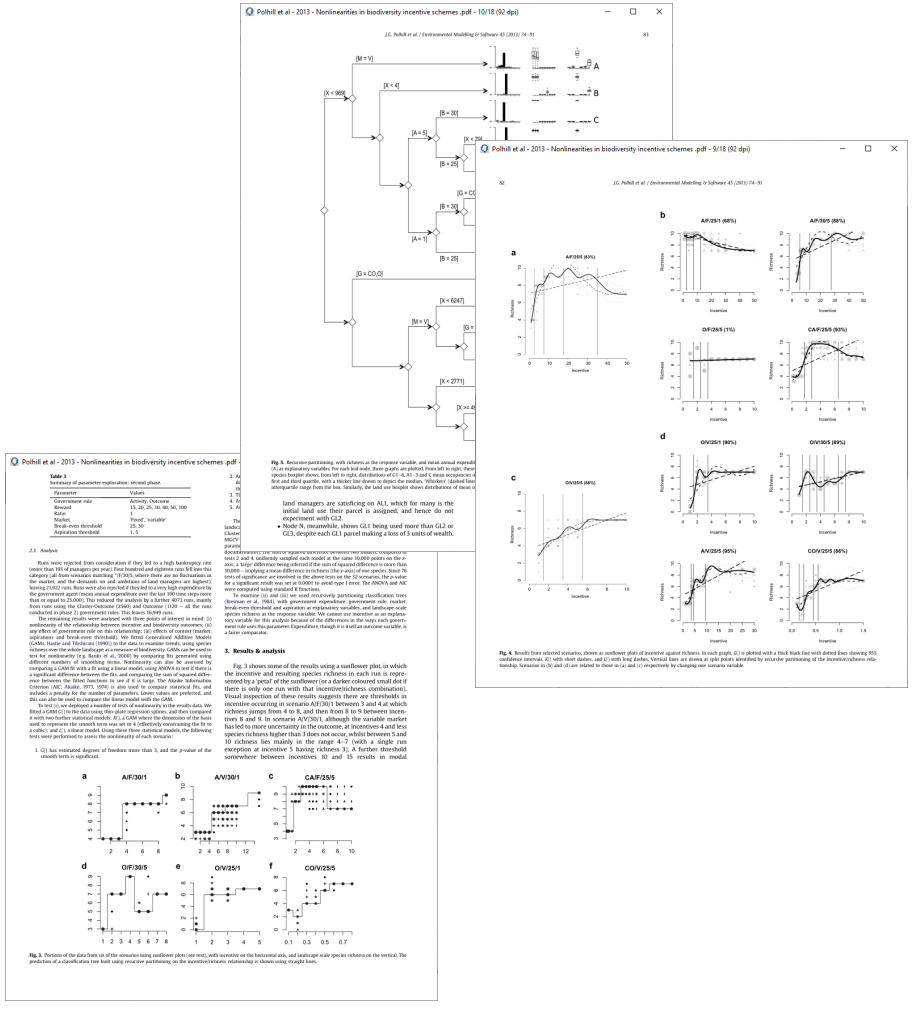
The James Hutton Institute

- Abstract
- Motivation
- The standard approach
- The database
- The scope
- A simple workflow
- Some provenance
- So what?
- But so what?
- Maybe a use?
- What else?

Abstract

We demonstrate a working framework for the automatic recording of provenance and metadata for primarily agent-based models that could easily be adapted to the other modelling environments. We discuss the need for such a framework, the philosophy behind the design we adopted, the implementation, discuss the results and demonstrate a simple tool for for tracing bad data through a provenance graph.

Motivation


Polhill et al - 2013 - Nonlinearities in biodiversity incentive schemes .pdf - 10/18 (92 dpi)
Polhill et al - 2013 - Nonlinearities in biodiversity incentive schemes .pdf - 9/18 (92 dpi)
Polhill et al - 2013 - Nonlinearities in biodiversity incentive schemes .pdf
land managers are satisficing on AL1, which for many is the initial land use their parcel is assigned, and hence do not experiment with GL2.
Node N, meanwhile, shows GL1 being used more than GL2 or GL3, despite each GL1 parcel making a loss of 3 units of wealth.
3. Results & analysis
A/F/30/1
A/V/30/1
CA/F/25/5
O/F/30/5
O/V/25/1
CO/V/25/5
A/F/25/1 (66%)
A/F/30/5 (90%)
O/F/25/5 (1%)
CA/F/25/5 (93%)
O/V/25/1 (90%)
O/V/30/5 (89%)
A/V/25/5 (95%)
CO/V/25/5 (86%)
Incentive
Richness

The standard approach

The James Hutton Institute

- Scripts
- Storage archives
- Diaries

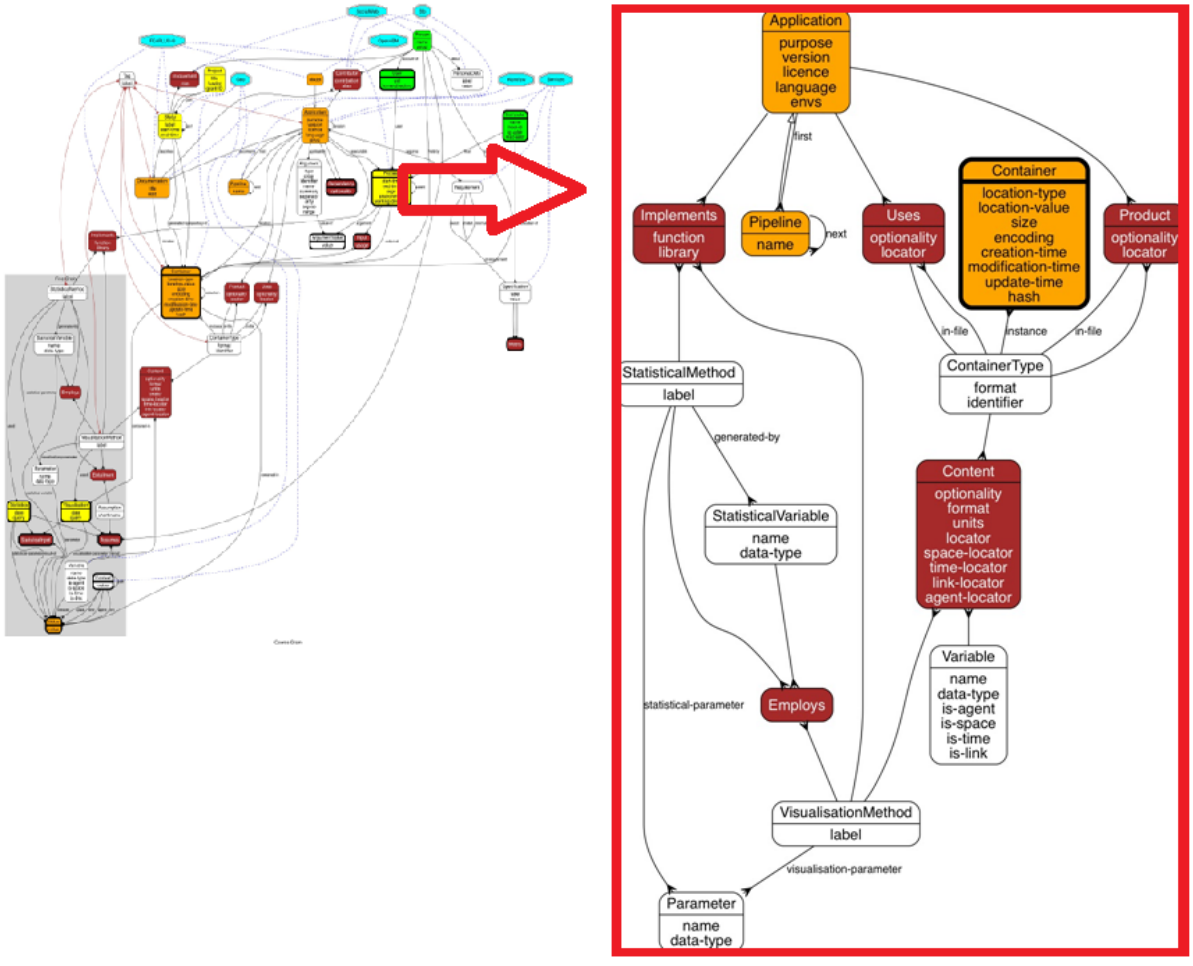
How we did it

```
    (>&2 echo "$0: Required 4G of memory got ${memory}G"
    exit -1
fi

if SSREPI_require_minimum $ME cpus $REQUIRED_NOS_CPUS ${cpus}
then
    (>&2 echo "$0: Minimum requirement for number of cpus failed")
    (>&2 echo "$0: Required $REQUIRED_NOS_CPUS cpus of memory got ${cpus}")
    exit -1
fi

SSREPI_run \
    SSS-StopC2-Cluster-create.sh \
    --purpose="

        Sets up the run for the lower value rewards
        and this is to test out multi-line,
        to see if it works. Which it does but removes the
        carriage returns and tabs. So you can nicely format it."

SSREPI_run \
    SSS-StopC2-Cluster-create2.sh \
    --purpose="Sets up the run for the higher value rewards

        Add some documentary stuff here."


SSREPI_run \
    SSS StopC2 Cluster run.sh \
      purpose="Does the runs for the lower value rewards" \
      add to dependencies=postprocessing.sh.dependencies


SSREPI_run \
    SSS StopC2 Cluster run2.sh \
      purpose="Does the runs for the higher value rewards" \
      add to dependencies=postprocessing.sh.dependencies


SSREPI_run \
    postprocessing.sh \
    --purpose="Post processing to almagamate results and produce pretty diagrams" \
    --with-dependencies=postprocessing.sh.dependencies

SSREPI study \
    $study id \
    --project=$project id \
    --end time=$(date "+%Y%m%dT%H%M%S") > /dev/null
E486: Pattern not found: li                                   244,2-5       99%
```

Python library

Bash
Library

The database



Social **S**imultion **RE**plication **I**nterface

The scope

The James Hutton Institute

- **Analysis** - fine grain provenance pertaining to statistical and visualisation outputs.
- **Finegrain** - a provenance diagram down to the level of variables.
- **Folksonomy** - a diagram showing annotations against the database, produced and categorised at the discretion of the user doing the annotation
- **Project** - management metadata. Largest granularity of metadata supported
- **Provenance** - provenance diagram at the level of file and parameter
- **Services** - service provided and requirement description
- **Workflow** - the actual workflow

What was needed

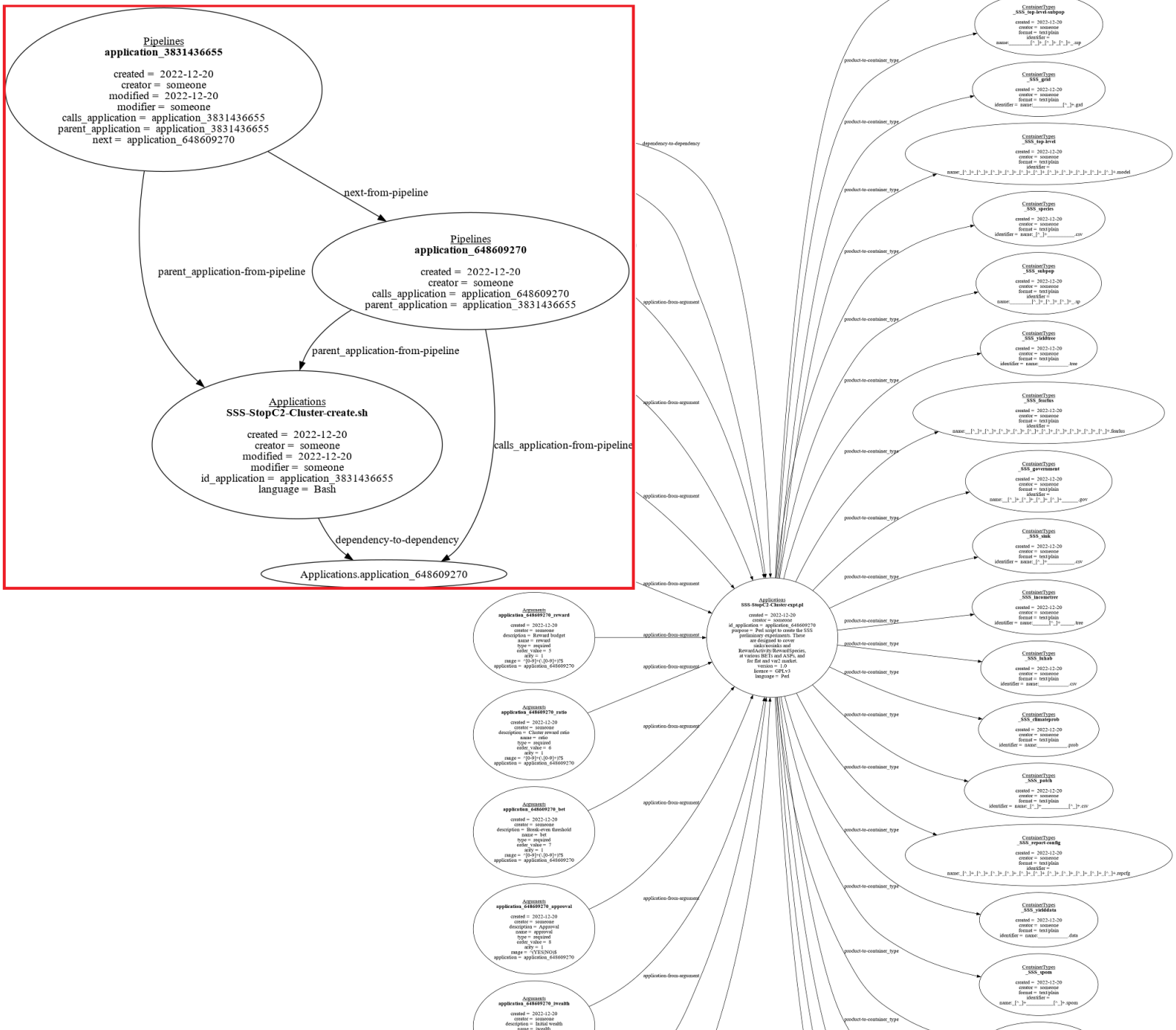
Primitive	Type	Purpose
SSREPI_require_minimum	M	Lower bound on software hardware required
SSREPI_require_exact	M	Exact bound on software hardware required
SSREPI_application	P & M	specifies some executable
SSREPI_me	P & M	Determines executable or reference to executable being run.
SSREPI_argument	P	An argument type to an executable
SSREPI_output	P	An output type from an executable
SSREPI_input	P	An input type for an executable
SSREPI_person	M	Provide metadata for a particular actor within this system
SSREPI_project	M	Specifies a project which contains all studies
SSREPI_study	M	A set of experiments makes up a single study
SSREPI_set	M	Sets the default licence and other metadata
SSREPI_involvement	M	Links personnel to a study
SSREPI_paper	M	A paper associated with this study
SSREPI_make_tag	M	Used for building a folksonomy
SSREPI_tag	M	Used to tag any entity with a folksonomy tag
SSREPI_contributor	M	A person with some kind of relation to an executable or script.
SSREPI_statistical_method	M	Record a statistical method
SSREPI_visualisation	M	Record a method to create an image to depict one or more values.
SSREPI_statistics	M	Record activities that populate the values of statistical variables.
SSREPI_visualisation_method	M	Methods for generating visualisations.
SSREPI_implements	M	Links a statistical or visualisation method to an application
SSREPI_parameter	M	Record the name taken by a statistical or visualisation method.
SSREPI_statistical_variable	M	A name for (one of) the result(s) of a statistical method.
SSREPI_visualisation_variable	P & M	Declares a named variable of interest
SSREPI_variable	M	Names a variable of interest
SSREPI_statistical_variable_value	P & M	Sets an actual value for a named statistical variable
SSREPI_value	P	Sets a value.
SSREPI_content	M	Links a kind of output/input/argument to a variable
SSREPI_person_makes_assumption	M	Links a person to an assumption

What was needed? (cont.)

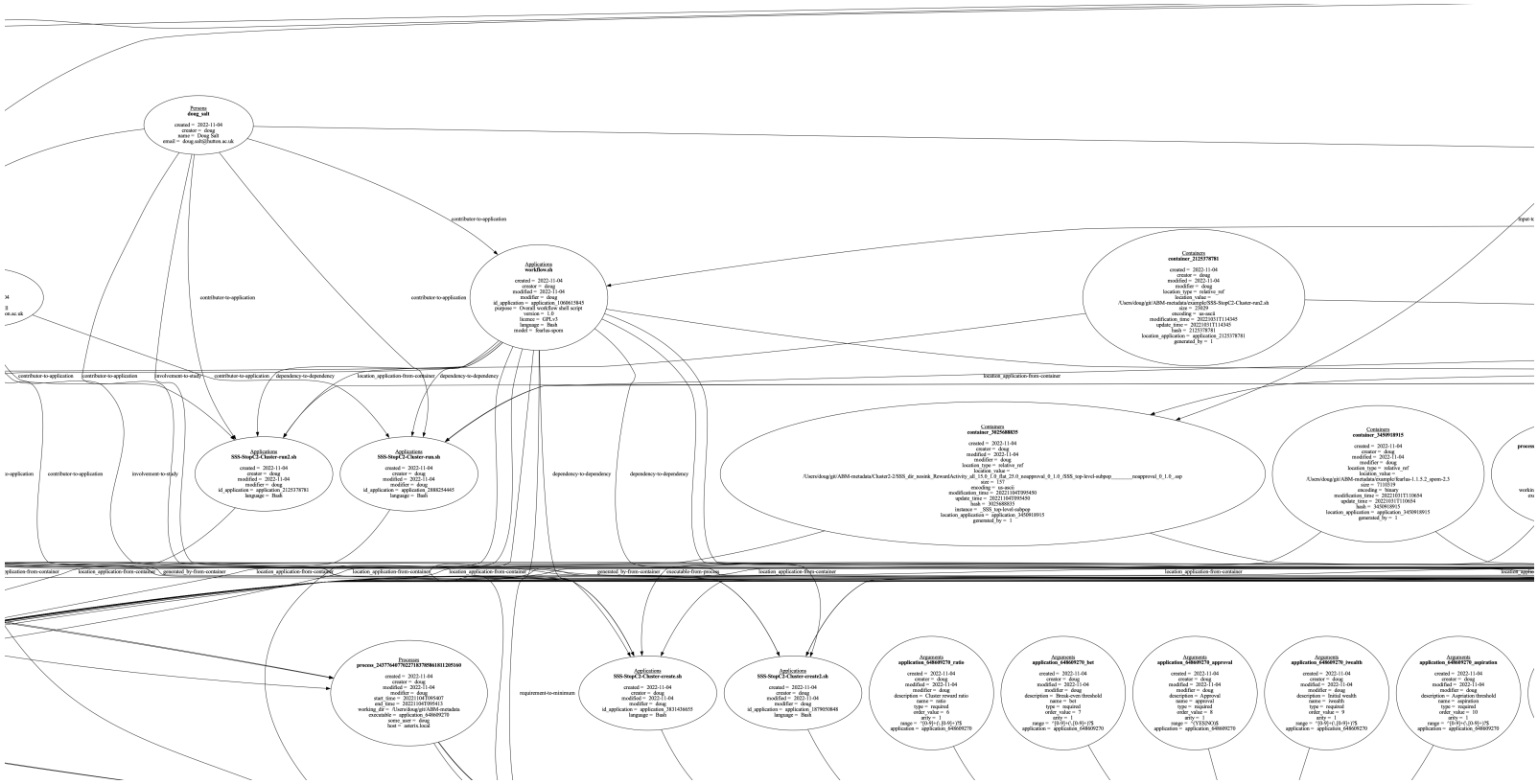
The James Hutton Institute

We encoded the previous actions in Python and then then called these from Bash

- `create_database.py` - creates a database idempotently.
- `exists.py` - checks if a particular row in a table exists.
- `get_value.py` - gets any specified single value from a table given the primary key.
- `get_values.py` - gets one or more rows given the search values.
- `next_study.py` - gets the next available and unique study number.
- `update.py` - idempotently updates a particular row in a table.


Pipelines
application_3831436655
created = 2022-12-20
creator = someone
modified = 2022-12-20
modifier = someone
calls_application = application_3831436655
parent_application = application_3831436655
next = application_648609270
next-from-pipeline
Pipelines
application_648609270
created = 2022-12-20
creator = someone
calls_application = application_648609270
parent_application = application_3831436655
parent_application-from-pipeline
parent_application-from-pipeline
Applications
SSS-StopC2-Cluster-create.sh
created = 2022-12-20
creator = someone
modified = 2022-12-20
modifier = someone
id_application = application_3831436655
language = Bash
calls_application-from-pipeline
dependency-to-dependency
Applications.application_648609270
dependency-to-dependency
application-from-argument
application-from-argument
application-from-argument
application-from-argument
application-from-argument
application-from-argument
application-from-argument
application-from-argument
application-from-argument
Arguments
application_648609270_reward
created = 2022-12-20
creator = someone
description = Reward budget
name = reward
type = required
order_value = 5
arity = 1
range = ^[0-9]+(\.[0-9]+)?$
application = application_648609270
Arguments
application_648609270_ratio
created = 2022-12-20
creator = someone
description = Cluster reward ratio
name = ratio
type = required
order_value = 6
arity = 1
range = ^[0-9]+(\.[0-9]+)?$
application = application_648609270
Arguments
application_648609270_bet
created = 2022-12-20
creator = someone
description = Break-even threshold
name = bet
type = required
order_value = 7
arity = 1
range = ^[0-9]+(\.[0-9]+)?$
application = application_648609270
Arguments
application_648609270_approval
created = 2022-12-20
creator = someone
description = Approval
name = approval
type = required
order_value = 8
arity = 1
range = ^(YES|NO)$
application = application_648609270
Arguments
application_648609270_iwealth
created = 2022-12-20
creator = someone
description = Initial wealth
name = iwealth
Applications
SSS-StopC2-Cluster-expt.pl
created = 2022-12-20
creator = someone
id_application = application_648609270
purpose = Perl script to create the SSS
preliminary experiments. These
are designed to cover
sakko-nesiaku and
RewardActivity/RewardSpecies,
at various BETs and ASPs, and
for flat and var2 market.
version = 1.0
licence = GPLv3
language = Perl
product-to-containter_type
product-to-containter_type
product-to-containter_type
product-to-containter_type
product-to-containter_type
product-to-containter_type
product-to-containter_type
product-to-containter_type
product-to-containter_type
product-to-containter_type
product-to-containter_type
product-to-containter_type
product-to-containter_type
product-to-containter_type
product-to-containter_type
product-to-containter_type
ContainerTypes
_SSS_ecoanaystate
created = 2022-12-20
creator = someone
format = text/plain
identifier = name:_____[^_]+_____state
ContainerTypes
_SSS_top-level-subpop
created = 2022-12-20
creator = someone
format = text/plain
identifier =
name:_____[^_]+_[^_]+_[^_]+_sup
ContainerTypes
_SSS_grid
created = 2022-12-20
creator = someone
format = text/plain
identifier = name:_____[^_]+.gsd
ContainerTypes
_SSS_top-level
created = 2022-12-20
creator = someone
format = text/plain
identifier =
name:_[^_]+_[^_]+_[^_]+_[^_]+_[^_]+_[^_]+_[^_]+_[^_]+_[^_]+_[^_]+_[^_]+_[^_]+_[^_]+_[^_]+.model
ContainerTypes
_SSS_species
created = 2022-12-20
creator = someone
format = text/plain
identifier = name:_[^_]+_____.csv
ContainerTypes
_SSS_subpop
created = 2022-12-20
creator = someone
format = text/plain
identifier =
name:_____[^_]+_[^_]+_[^_]+_sp
ContainerTypes
_SSS_yieldtree
created = 2022-12-20
creator = someone
format = text/plain
identifier = name:_____tree
ContainerTypes
_SSS_fourfits
created = 2022-12-20
creator = someone
format = text/plain
identifier =
name:__[^_]+_[^_]+_[^_]+_[^_]+_[^_]+_[^_]+_[^_]+_[^_]+_[^_]+_[^_]+_[^_]+_[^_]+_[^_]+.fourfits
ContainerTypes
_SSS_government
created = 2022-12-20
creator = someone
format = text/plain
identifier =
name:__[^_]+_[^_]+_[^_]+_[^_]+_[^_]+_____.gov
ContainerTypes
_SSS_sink
created = 2022-12-20
creator = someone
format = text/plain
identifier = name:_[^_]+_____.csv
ContainerTypes
_SSS_incometree
created = 2022-12-20
creator = someone
format = text/plain
identifier = name:_____[^_]+_____tree
ContainerTypes
_SSS_hihab
created = 2022-12-20
creator = someone
format = text/plain
identifier = name:_____.csv
ContainerTypes
_SSS_climateprob
created = 2022-12-20
creator = someone
format = text/plain
identifier = name:_____prob
ContainerTypes
_SSS_patch
created = 2022-12-20
creator = someone
format = text/plain
identifier = name:_[^_]+_____[^_]+.csv
ContainerTypes
_SSS_repact-config
created = 2022-12-20
creator = someone
format = text/plain
identifier =
name:_[^_]+_[^_]+_[^_]+_[^_]+_[^_]+_[^_]+_[^_]+_[^_]+_[^_]+_[^_]+_[^_]+.repcfg
ContainerTypes
_SSS_yieldata
created = 2022-12-20
creator = someone
format = text/plain
identifier = name:_____data
ContainerTypes
_SSS_spom
created = 2022-12-20
creator = someone
format = text/plain
identifier =
name:_[^_]+_____[^_]+.spom

Some provenance!!!



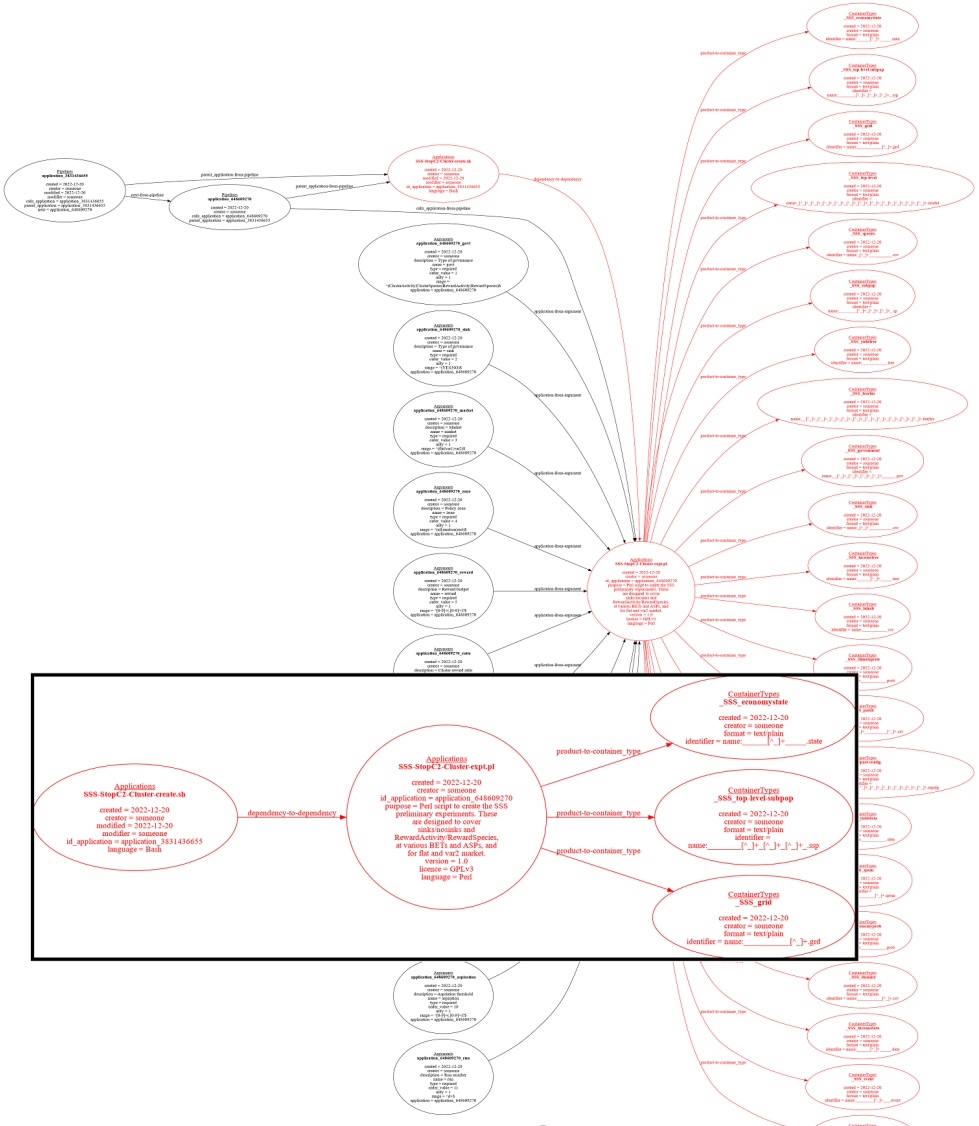
So what?

The James Hutton Institute

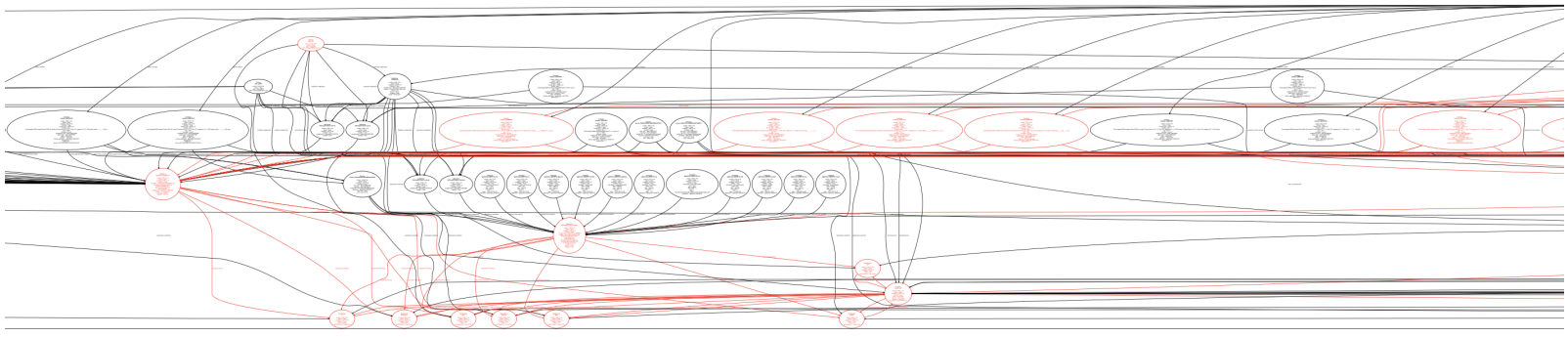
It now takes me the following lines of code to reproduce an experiment

1. Create new directory
2. Enter new directory
3. `mkdir bin`
4. `cp /mnt/storage/doug/git/ABM-metadata/example/*.sh bin`
5. `cp /mnt/storage/doug/git/ABM-metadata/example/*.pl bin`
6. `cp /mnt/storage/doug/git/ABM-metadata/example/*.R bin`
7. `ln -s /mnt/storage/doug/git/ABM-metadata/cfg`
8. `ln -s /mnt/storage/doug/git/ABM-metadata/lib`
9. `ln -s /mnt/storage/doug/git/ABM-metadata/doc`
10. Amend `bin/path.sh` to `bin/path.`*directory_name*`.sh`
11. source `bin/path.`*directory_name*`.sh`
12. `psql`
13. `create database third`
14. `\q`
15. profit!

But so what?


Applications
SSS-StopC2-Cluster-create.sh
created = 2022-12-20
creator = someone
modified = 2022-12-20
modifier = someone
id_application = application_3831436655
language = Bash
dependency-to-dependency
Applications
SSS-StopC2-Cluster-expt.pl
created = 2022-12-20
creator = someone
id_application = application_648609270
purpose = Perl script to create the SSS preliminary experiments. These are designed to cover sinks/zooinks and RewardActivity/RewardSpecies, at various BETA and ASPs, and for flat and var2 market.
version = 1.0
licence = GPLv3
language = Perl
product-to-container_type
product-to-container_type
product-to-container_type
ContainerTypes
_SSS_economystate
created = 2022-12-20
creator = someone
format = text/plain
identifier = name:_____[^_]+_____state
ContainerTypes
_SSS_top_level_subpop
created = 2022-12-20
creator = someone
format = text/plain
identifier = name:_____[^_]+_[^_]+_[^_]+_spp
ContainerTypes
_SSS_grid
created = 2022-12-20
creator = someone
format = text/plain
identifier = name:_____[^_]+_grd

Finally a use?



What else?

Thank you very much