An R package for managing data in a multi-stage experimental workflow

Data versioning and provenance considerations in interactive scripting

Philip Eichinski, Paul Roe
philip.eichinski@qut.edu.au
Queensland University of Technology, Brisbane, Australia

Overview

◊ Datatrack R package records and maintains provenance metadata for saved data objects produced in a workflow.

◊ It can display this metadata in R Studio as a directed graph showing dependencies between data objects. This is useful as a tool to assist an experimenter to select the correct input data when running a step of the workflow in isolation.

Motivation

In experimental research using computation, a workflow is a sequence of steps involving some data processing or analysis where the output of one step may be used as the input of another. The processing steps may involve user-supplied parameters, that when modified, result in a new version of input to the downstream steps, in turn generating new versions of their own output. As more experimentation is done, the results of these various steps can become numerous.

It is important to keep track of which data output is dependent on which other generated data, and which parameters were used. In many situations, scientific workflow management systems solve this problem, but these systems are best suited to collaborative, distributed experiments using a variety of services.

This contrasts with exploratory research performed in interactive scripting environments such as R, where the code that performs the analysis is also the code that defines the workflow, and is being constantly modified. This does not always lend itself to integration with a SWfMS.

Even in circumstances where a SWfMS could be used, it may be resisted by the researcher because it may draw the researcher out of the environment that they are comfortable working in and require additional effort to learn a new system. When the research calls for frequent modification to the code that runs the data processing steps, working in a SWfMS means effectively working in two environments possibly with two scripting languages.

Compared with other solutions

such as scientific workflow management systems

SWfMS are excellent tools for tasks such as workflow composition, mapping the workflow onto resources or services, executing the workflow and recording the provenance metadata to allow the final output to be reproduced in the future. However, the disadvantage is the investment of time learning a new system. In the initial experimental stages of creating analysis scripts in R, when the code is being constantly modified and re-run switching to a SWfMS may not be desirable.

The advantage of Datatrack is that it can be integrated easily into the familiar R scripting environment, with minimal changes to the working patterns that the experimenter is already using.

In the R scripts that comprise the workflow, instead of using the native functions to read and write files to disk, use the datatrack functions.

Integration in existing R scripts is simple.

Writing data

```
WriteDataObject(data, name, [parameters],
[dependencies], [annotations])
```

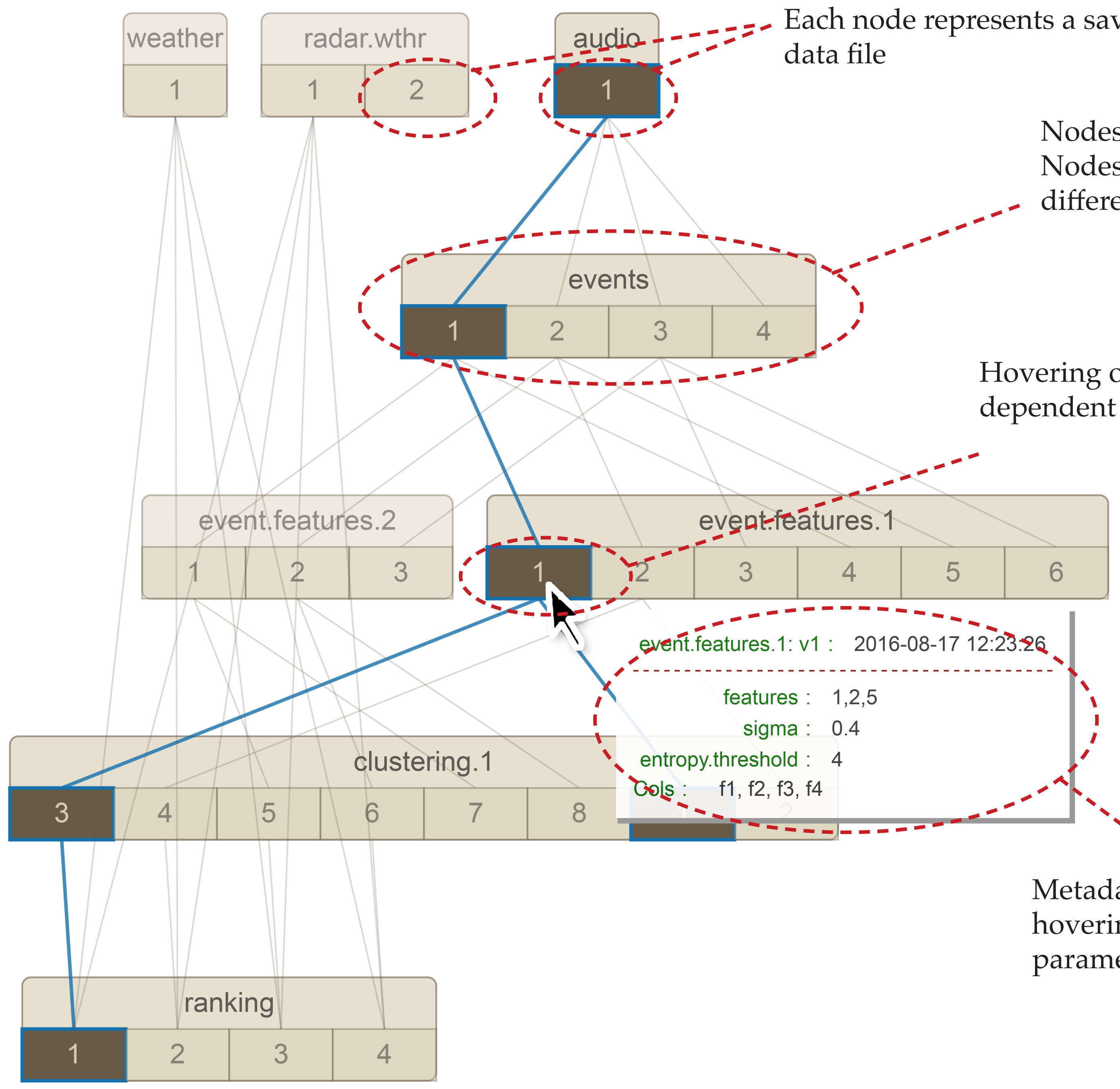
The user can specify (as arguments to the function) the metadata that will be associated with the saved data object.

Datatrack will save a different version of the data for each unique combination of parameters and dependencies.

Reading data

```
ReadDataObject(name)
```

Where multiple versions of the data with the specified name exist, datatrack will prompt the user to select which version after presenting them with an interactive graph of provenance metadata that shows the available choices.



event.features.1: v1 : 2016-08-17 12:23:26

features : 1,2,5
sigma : 0.4
entropy.threshold : 4
Cols : f1, f2, f3, f4

Each node represents a saved data file

Nodes are grouped by name. Nodes in a group represent different versions for that name

Hovering over a node highlights the dependent and dependency data files.

Metadata for a node is revealed by hovering, including date saved, parameters used

Interactive graph of data object dependencies is shown in the R studio viewer to assist selection of the correct input data

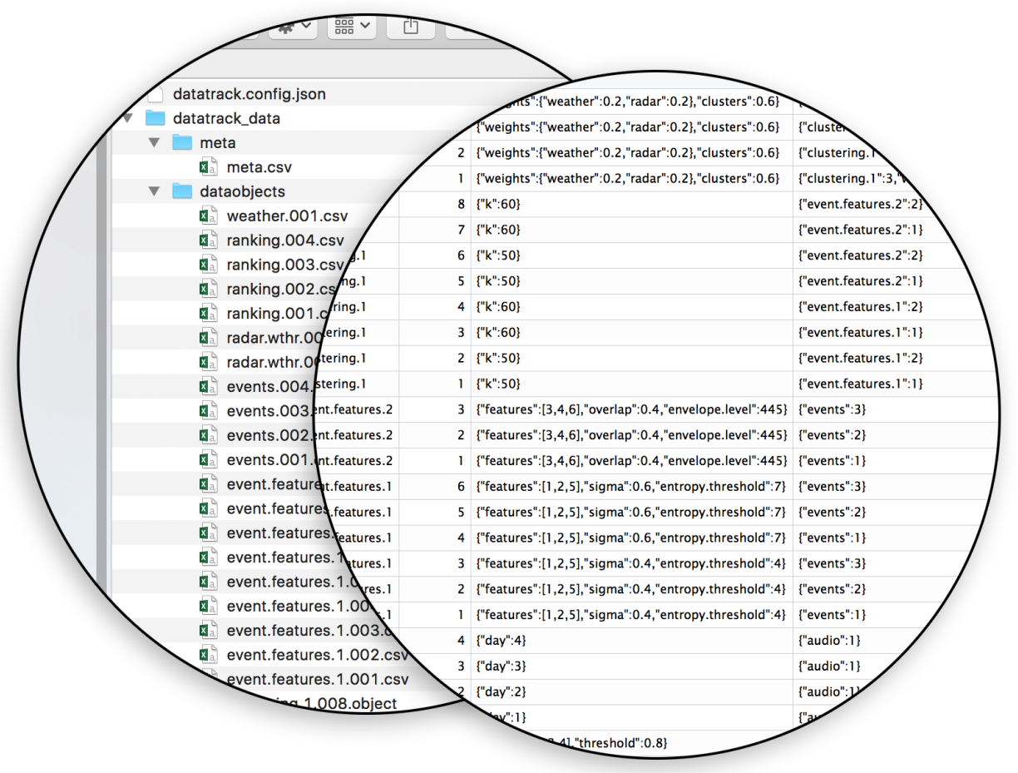
Stored metadata

User specified

- Parameters used when generating the data
- Which data was used as inputs to the process that generated the data
- Annotations

Background

- System information: R version, loaded packages versions, OS version
- Stack trace to the function saving the data
- Date and time



Metadata is stored in a csv on disk along with the data files at a location specified in configuration

Available as a beta release on github https://github.com/peichins/datatrack