



Leibniz-Institut für  
Astrophysik Potsdam

# Provenance Data Model

## RAVE Use case

**DM session at IVOA InterOp**

**May 2016, Cape Town**

**Kristin Riebe, AIP, GAVO**



# Example: Reduced RAVE-fits file

- PROV-N notation

- 2 files

```
entity(rave:0645m522I0049.wav.fits, [prov:type = 'std:fits']  
entity(rave:0645m522I0049.fits, [prov:type = 'std:fits']
```

- 2 agents

- 2 activities

- relations

# Example: Reduced RAVE-fits file

- PROV-N notation

- 2 files

```
entity(rave:0645m522I0049.fits, [prov:type = 'std:fits']  
entity(rave:0645m522I0049.wav.fits, [prov:type = 'std:fits']
```

- 2 agents

```
agent(aao:Paul_Cass, [prov:type='prov:Person'])  
agent(rave:Alessandro_Siviero, [prov:type='prov:Person'])
```

- 2 activities

- relations

# Example: Reduced RAVE-fits file

- PROV-N notation

- 2 files

```
entity(rave:0645m522I0049.fits, [prov:type = 'std:fits']  
entity(rave:0645m522I0049.wav.fits, [prov:type = 'std:fits']
```

- 2 agents

```
agent(aao:Paul_Cass, [prov:type='prov:Person'])  
agent(rave:Alessandro_Siviero, [prov:type='prov:Person'])
```

- 2 activities

```
activity(rave:act_observation, 2008-02-16T13:25:24, -,  
        [ prov:type = 'obs:Observation' ] )  
activity(rave:act_irafReduction, 2008-03-04T09:46:57, -,  
        [ prov:type = 'std:reduction' ] )
```

- relations

# Example: Reduced RAVE-fits file

- PROV-N notation

- 2 files

```
entity(rave:0645m522I0049.fits, [prov:type = 'std:fits']  
entity(rave:0645m522I0049.wav.fits, [prov:type = 'std:fits']
```

- 2 agents

```
agent(aao:Paul_Cass, [prov:type='prov:Person']  
agent(rave:Alessandro_Siviero, [prov:type='prov:Person'])
```

- 2 activities

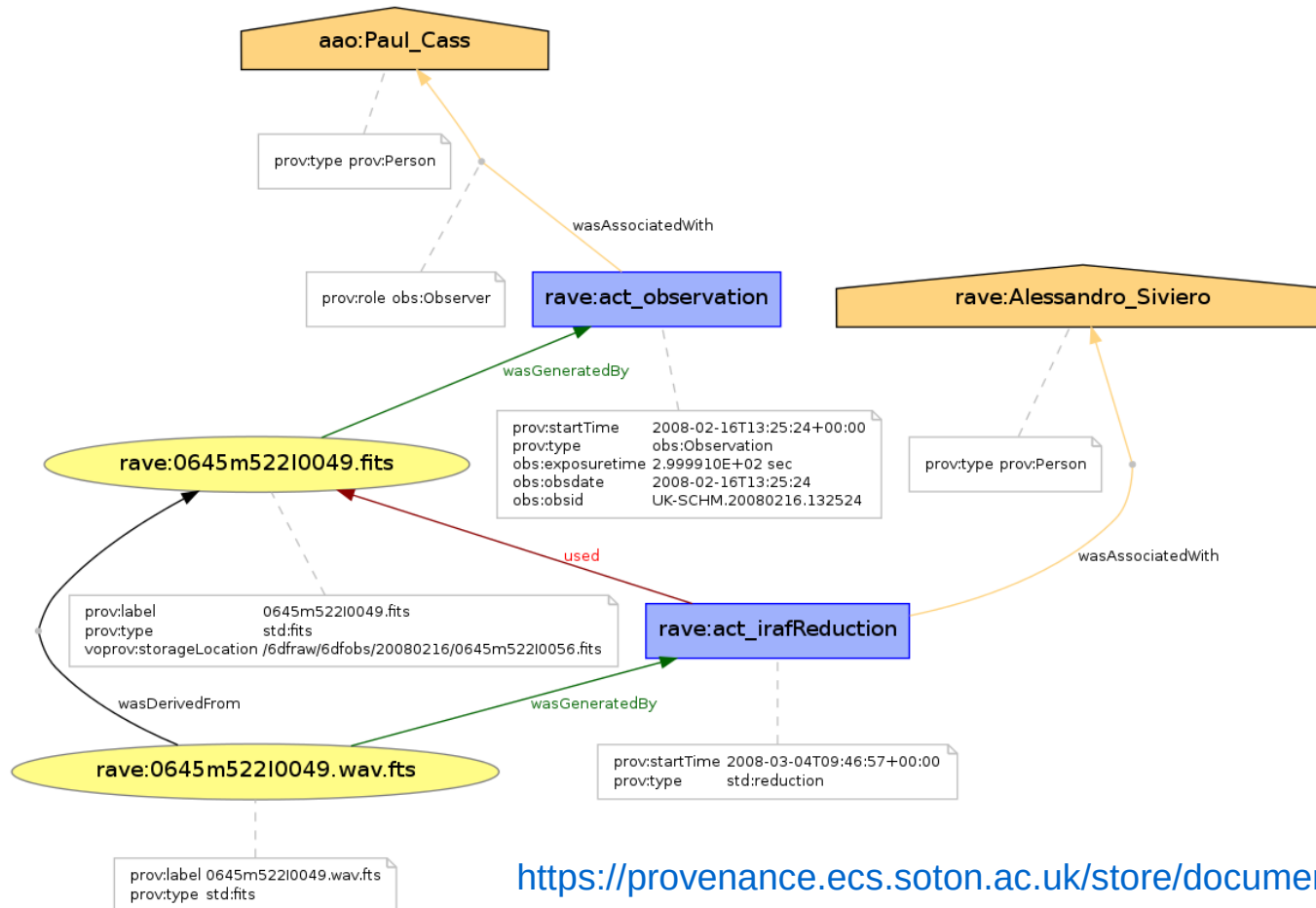
```
activity(rave:act_observation, 2008-02-16T13:25:24, -,  
[ prov:type = 'obs:Observation' ] )  
activity(rave:act_irafReduction, 2008-03-04T09:46:57, -,  
[ prov:type = 'std:reduction' ] )
```

- relations

```
wasAssociatedWith(rave:act_observation, aao:Paul_Cass, -,  
[ prov:role = 'obs:Observer' ] )  
wasAssociatedWith(rave:act_irafReduction, rave:Alessandro_Siviero, -)  
wasGeneratedBy(rave:0645m522I0049.fits, rave:act_observation, -)  
used(rave:act_irafReduction, rave:0645m522I0049.fits, -)  
wasGeneratedBy(rave:0645m522I0049.wav.fits, rave:act_irafReduction, -)  
wasDerivedFrom(rave:0645m522I0049.wav.fits, rave:0645m522I0049.fits)
```

# Example: Reduced RAVE-fits file

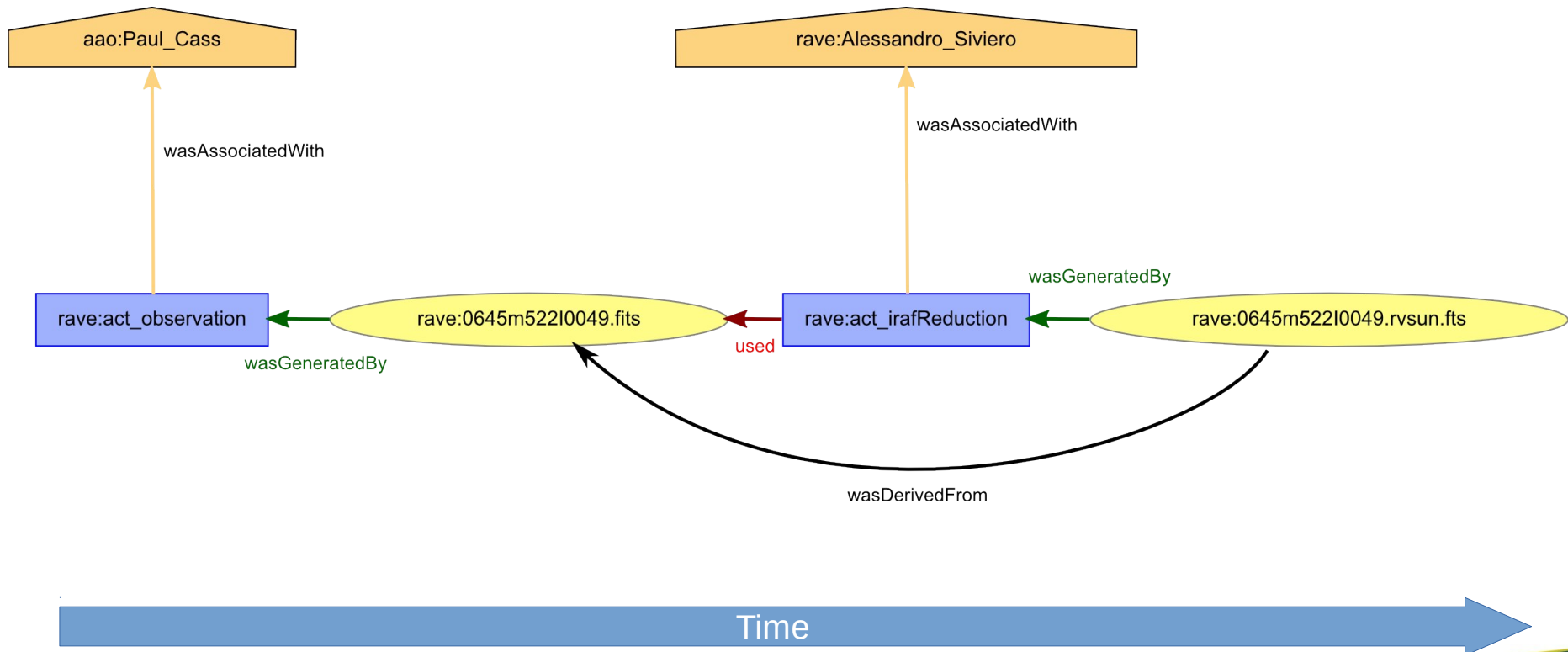
- Graph produced with ProvStore (using GraphViz):



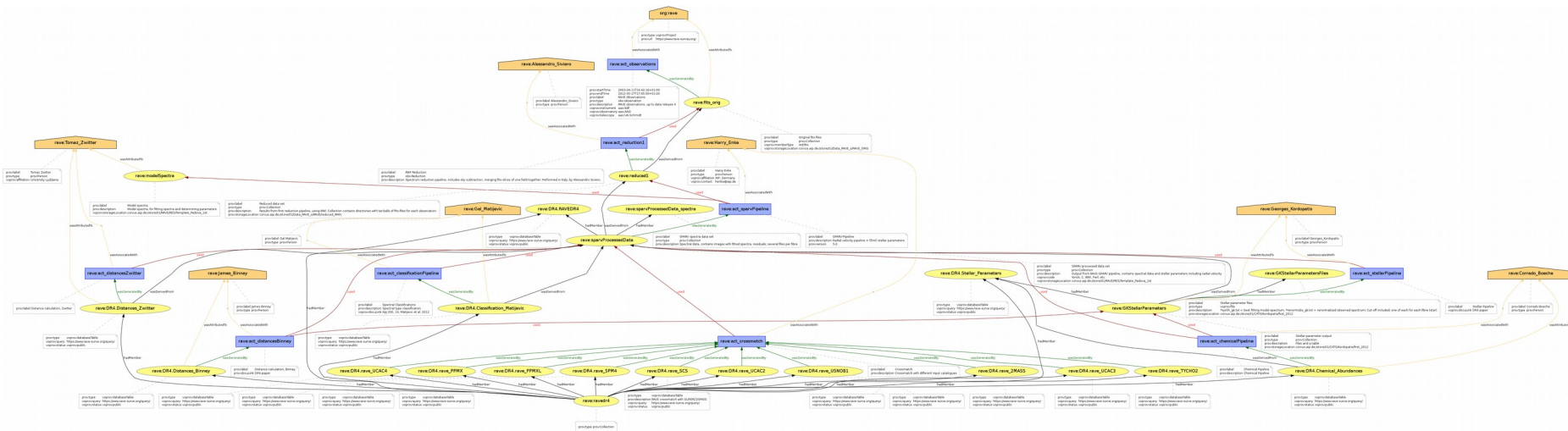
<https://provenance.ecs.soton.ac.uk/store/documents/84420/>

# Example: Reduced RAVE fits-file

- Graph reordered, attributes hidden:



# Example: RAVE database tables (nearly complete history)



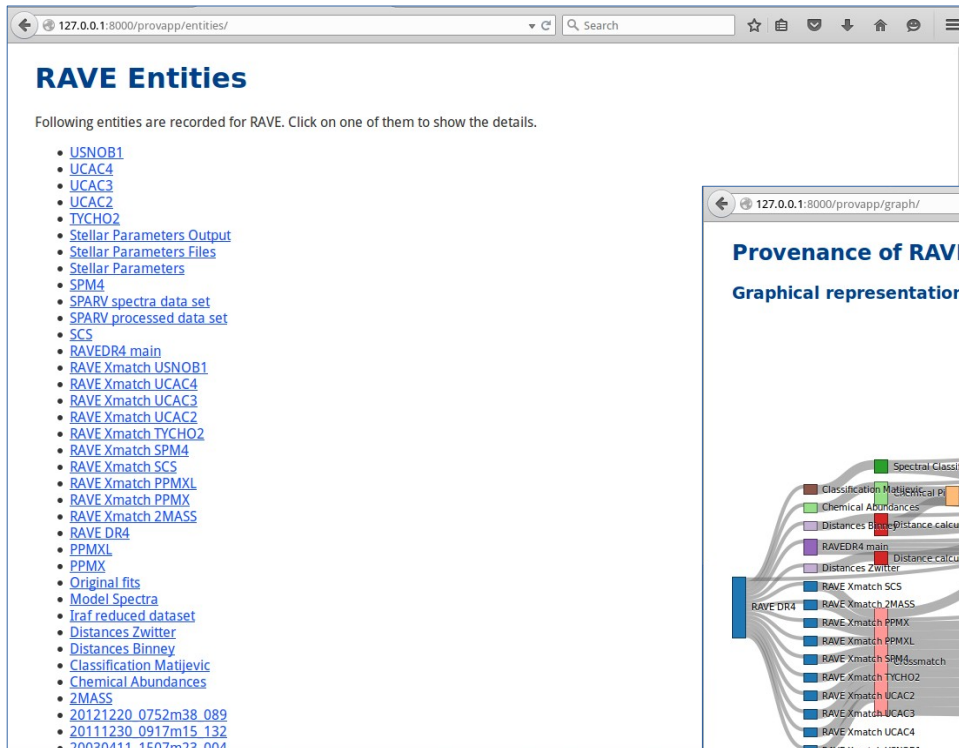
<https://provenance.ecs.soton.ac.uk/store/documents/84064/>



# Webapp for RAVE provenance

- Testing how to implement the data model myself
- Simple setup using Django Framework with SQLite3 database
- Define classes “as is”, main provenance classes, one DB table for each:
  - **entity**
  - **activity**
  - **agent**
  - used -- foreign keys to activity, entity
  - wasGeneratedBy -- foreign keys to entity, activity
  - wasAssociatedWith -- foreign keys to entity, agent
  - hadMember -- foreign keys to entities (one with type collection)
  - wasDerivedFrom -- foreign keys to entities

# Webapp for RAVE provenance

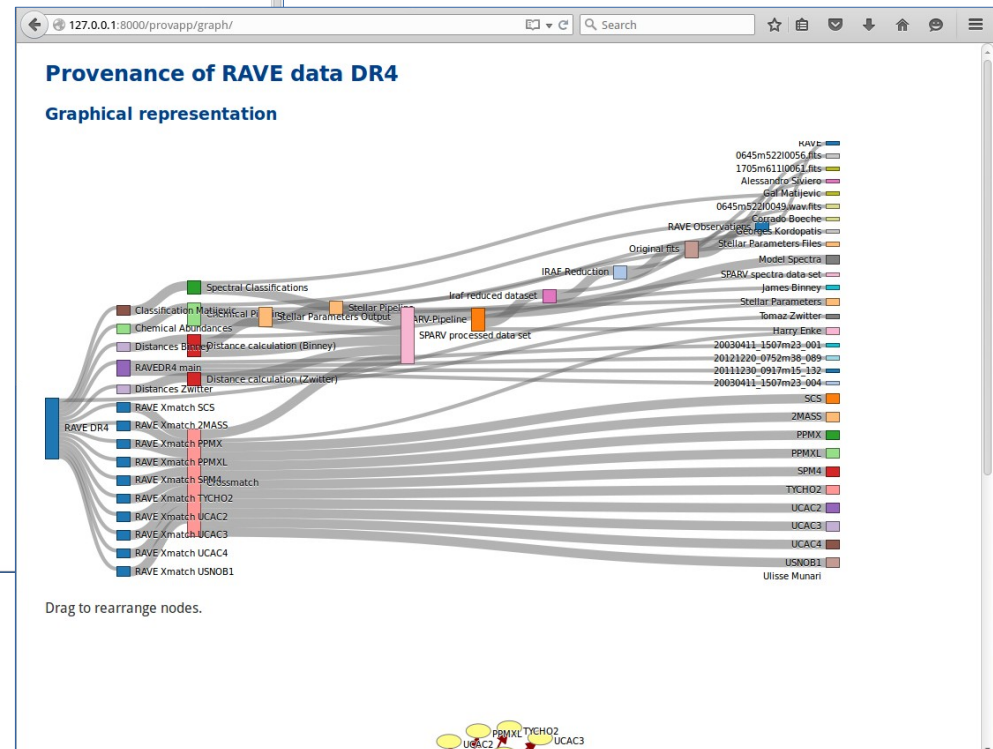


127.0.0.1:8000/provapp/entities/

## RAVE Entities

Following entities are recorded for RAVE. Click on one of them to show the details.

- USNOB1
- UCAC4
- UCAC3
- UCAC2
- TYCHO2
- Stellar Parameters Output
- Stellar Parameters Files
- Stellar Parameters
- SPM4
- SPARV\_spectra\_data\_set
- SPARV\_processed\_data\_set
- SCS
- RAVEDR4\_main
- RAVE Xmatch USNOB1
- RAVE Xmatch UCAC4
- RAVE Xmatch UCAC3
- RAVE Xmatch UCAC2
- RAVE Xmatch TYCHO2
- RAVE Xmatch SPM4
- RAVE Xmatch SCS
- RAVE Xmatch PPMXL
- RAVE Xmatch PPMX
- RAVE Xmatch 2MASS
- RAVE DR4
- PPMXL
- PPMX
- Original fits
- Model Spectra
- Iraf\_reduced\_dataset
- Distances Zwitter
- Distances Binney
- Classification Matijevic
- Chemical Abundances
- 2MASS
- 20121220\_0752m38\_089
- 20111230\_0917m15\_132
- 20030411\_1507m23\_004



# Webapp for RAVE provenance

- Create views to show e.g.
  - provn-serialisation of the complete provenance
  - graph-representation
    - could also divide into 3 different views:
      - data flow (entities)
      - process flow (activities)
      - responsibility view (agents)
  - list of activities, entities, agents
  - details for individual elements
- Provide detailed information for individual observations
  - given an obsId, return file names and locations of intermediate and raw files
- More use cases?

# Use cases

- There are different types of users with different needs:
  - “project manager”
    - interested mainly in coarse data flow, involved processes (activities), not very detailed
  - “pipeline writer”
    - e.g. scientist from the project (internal scientist)
    - interested in redoing parts of the pipeline, using different algorithms, testing influence of different parameters
  - “other scientist”
    - usually interested in science-ready data only (no need for raw observation files), quality assessment, error-bars, applicability of data, error tracking

# Use cases

- **Example: Project management**
  - Give me a visualisation of the data flow and the work flow, showing all involved activities, agents and resulting entities.
    - Interesting for PI of the project, someone writing a report, a funding agency
- **Example: Pipeline analysis**
  - Where are the raw fits-files? The flat-fields?
  - Can I get the extracted spectrum for each fiber? (How?)
  - Which processes were involved and where are they described?

# Use cases

- Examples: Scientist

- Who created the stellar\_parameters-table?
  - i.e.: get the agent associated with this entity, thus: retrieve details for this entity
- Where do the values in column Teff\_K come from? In which paper are the methods described? The uncertainties?
  - errors are in additional columns "e..."-something
- Are intermediate files (spectrum png/ascii) for a given obsId available? How could I get them?
  - Or: who do I need to ask for them?
  - Need: permission/accessibility flag, contact details
-

# Use cases

- Examples: Scientist (continued)
  - How are values (for a given star) changing for each data release? What's the difference in processing?
    - First part can be answered with published data alone, provenance only needed for second question.
  - Are there multiple observations of the same star? If the derived heliocentric radial velocity differs more than the error bars suggest: what was causing this difference? (Which processing step(s)?)
  - What is the coverage of this survey? Compare intended/actual coverage for studies of completeness/selection effects.
    - Needs additional information on failed fibers per field