Leibniz-Institut für
Astrophysik Potsdam

Provenance Data Model

Let's keep discussions going on ...

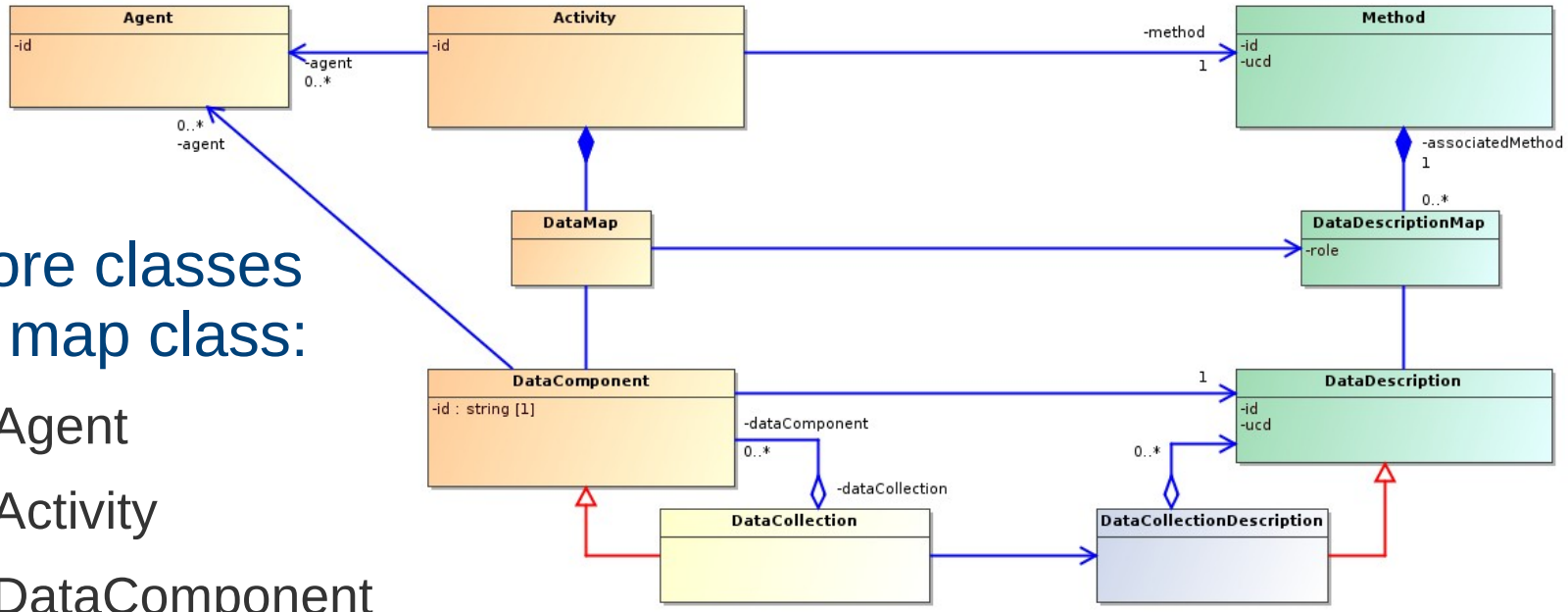
InterOp Sesto, June 2015

Kristin Riebe, GAVO

Status

First attempt: Model with protoypes

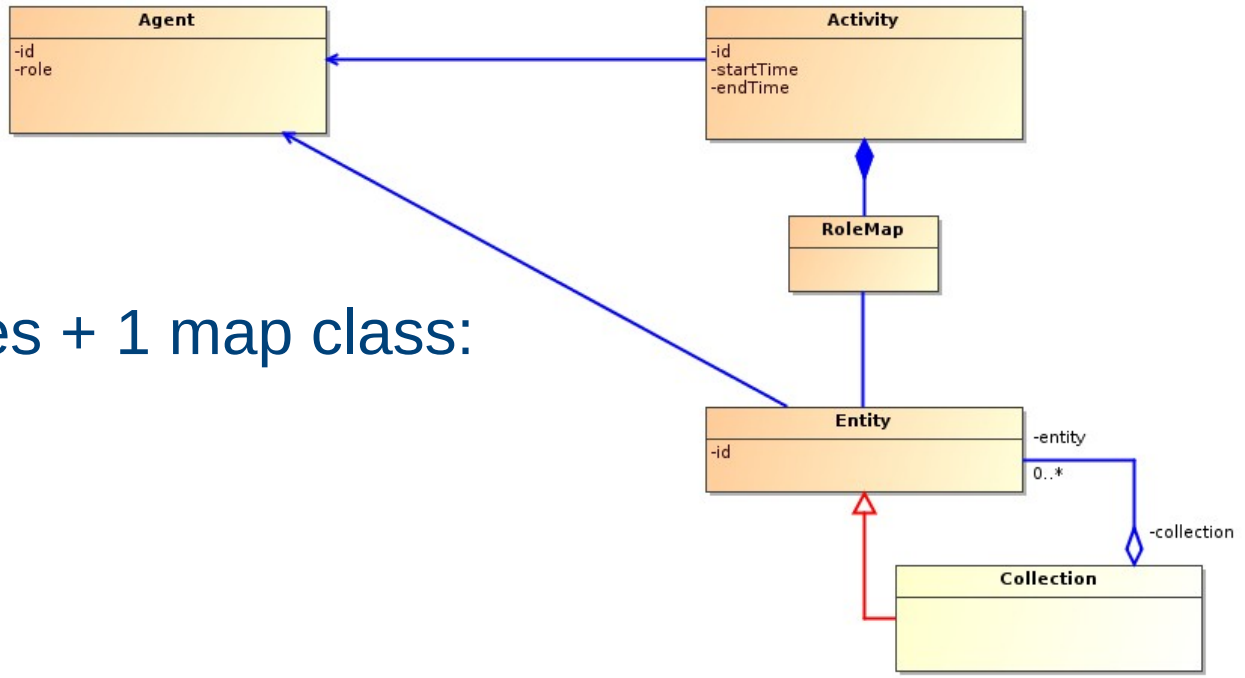
(inspired by SimDM and W3C model)



- 3 core classes
  + 1 map class:
  - Agent
  - Activity
  - DataComponent
  - DataMap
- + dataDescription side => each item doubled

Status

Model without protoypes

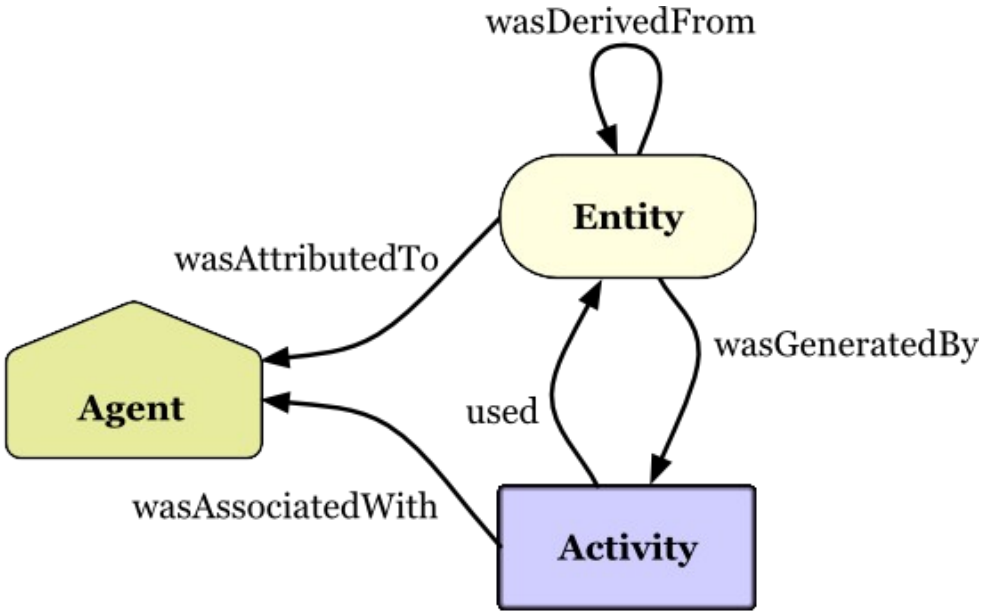


- 3 core classes + 1 map class:
  - Agent
  - Activity
  - Entity
  - RoleMap
- Let's start with this – put descriptions into attributes, common vocabularies and see how far we get

Looking at W3C

http://www.w3.org/TR/prov-dm/

- 3 core classes:
  - Agent
  - Activity
  - Entity
- core relations:
  - used
  - wasGeneratedBy
  - wasDerivedFrom
  - wasAttributedTo
  - wasAssociatedWith
- + many more classes and relations



Comparison with W3C

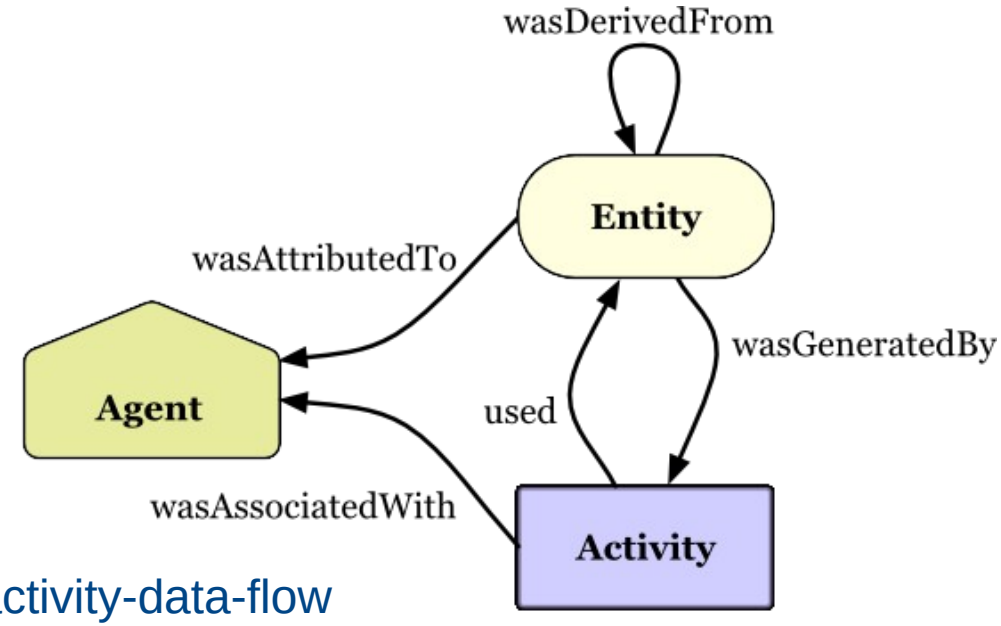
http://www.w3.org/TR/prov-dm/

- 3 core classes:
  - Agent
  - Activity
  - Entity
- core relations:
  - used
  - wasGeneratedBy
  - wasDerivedFrom
  - wasAttributedTo
  - wasAssociatedWith

activity-data-flow

data-flow

responsibility view

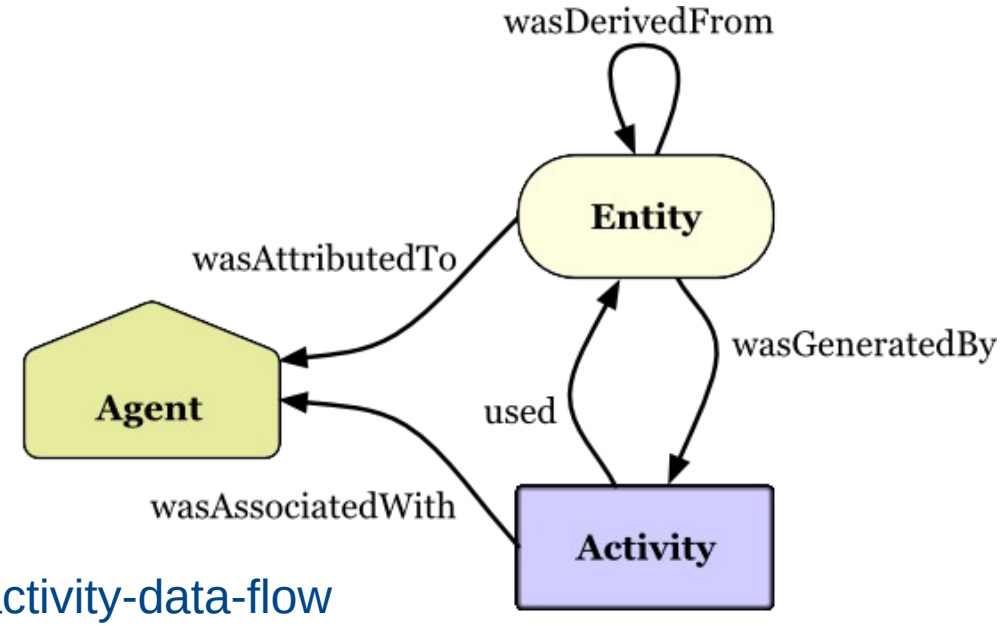


- + many more classes and relations

Comparison with W3C

http://www.w3.org/TR/prov-dm/

- 3 core classes:
  - Agent
  - Activity
  - Entity
- core relations:
  - used
  - wasGeneratedBy
  - wasDerivedFrom
  - wasAttributedTo
  - wasAssociatedWith
- + many more classes and relations



activity-data-flow

data-flow

responsibility view

=> Should be sufficient to give us what we need!

Example: Reduced RAVE-fits file

- PROV-N notation
- 2 files
- 2 agents
- 2 activities

- relations

entity(rave:0645m522l0049.wav.fits, [prov:type = 'std:fits']
entity(rave:0645m522l0049.fits, [prov:type = 'std:fits']

Example: Reduced RAVE-fits file

- PROV-N notation
- 2 files
- 2 agents
- 2 activities

entity(rave:0645m522l0049.fits, [prov:type = 'std:fits']
entity(rave:0645m522l0049.wav.fits, [prov:type = 'std:fits']

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

- relations

Example: Reduced RAVE-fits file

- PROV-N notation
- 2 files
- 2 agents
- 2 activities
- relations

```
entity(rave:0645m522l0049.fits, [prov:type = 'std:fits']
entity(rave:0645m522l0049.wav.fits, [prov:type = 'std:fits']

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

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' ])
```

Example: Reduced RAVE-fits file

- PROV-N notation
- 2 files
- 2 agents
- 2 activities
- relations

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

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

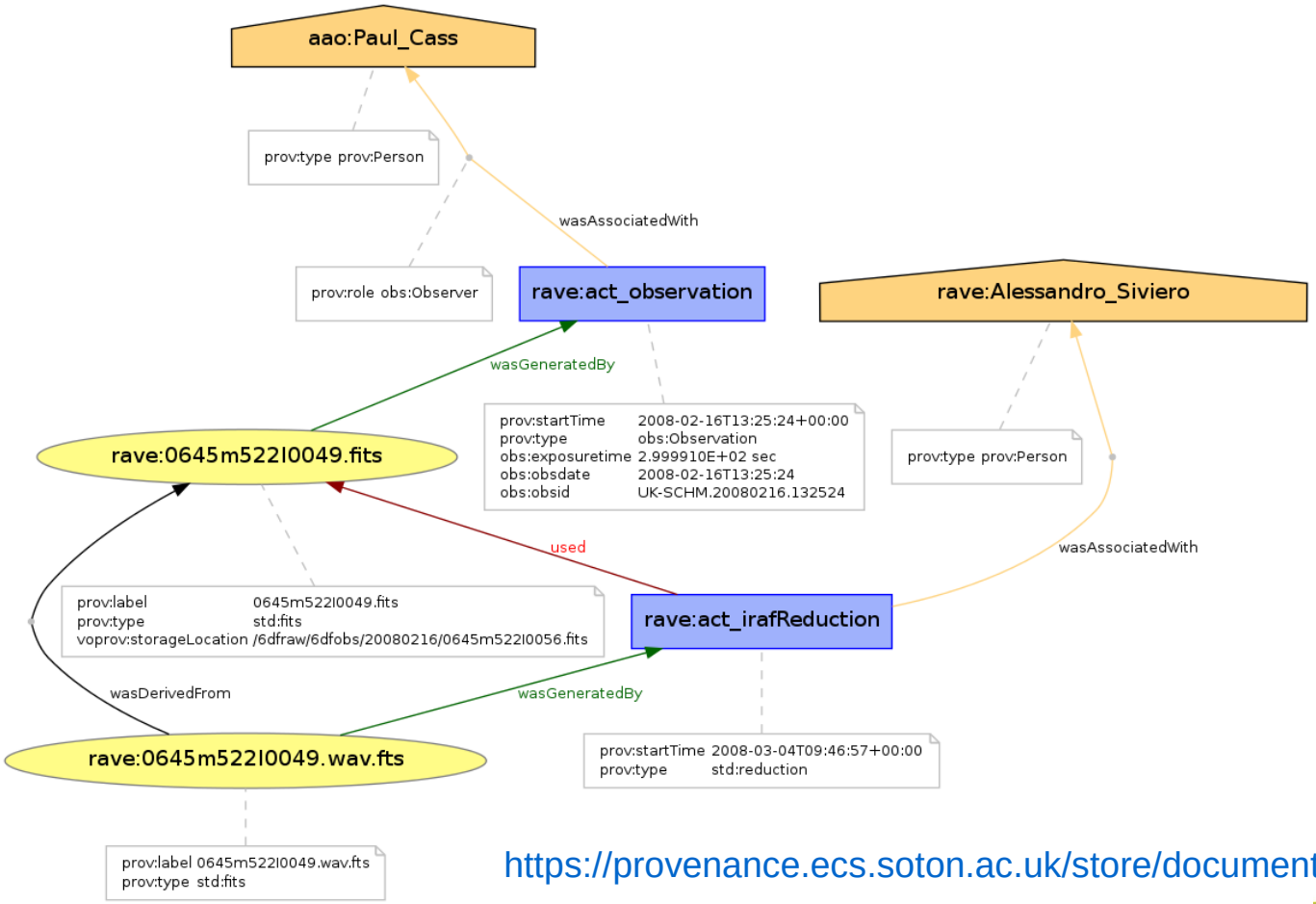
activity(rave:act_observation, 2008-02-16T13:25:24, -,
[prov:type = 'obs:Observation'])
activity(rave:act_irafReduction, 2008-03-04T09:46:57, -,
[prov:type = 'std:reduction'])

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.fts, rave:act_irafReduction, -)
wasDerivedFrom(rave:0645m522I0049.wav.fts, rave:0645m522I0049.fits)

Example: Reduced RAVE-fits file

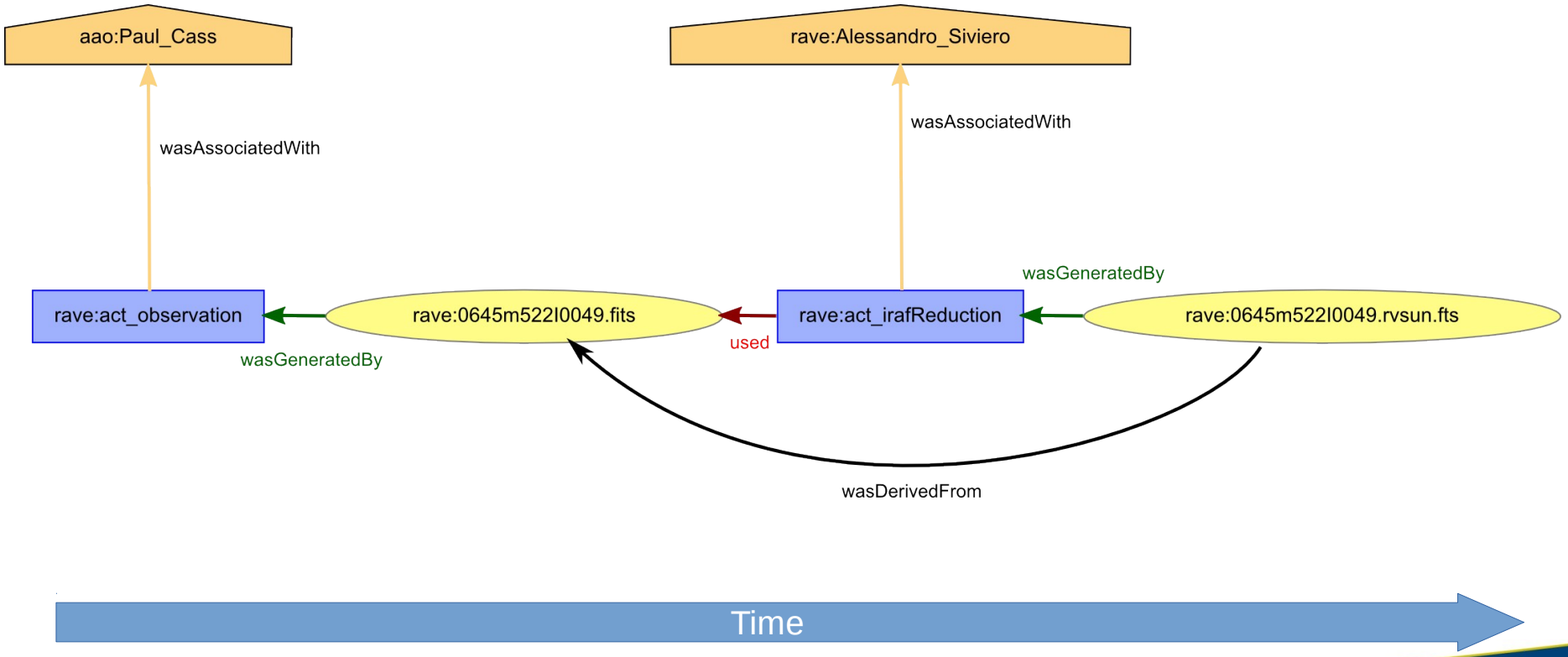
- Graph produced with ProvStore (using GraphViz):



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

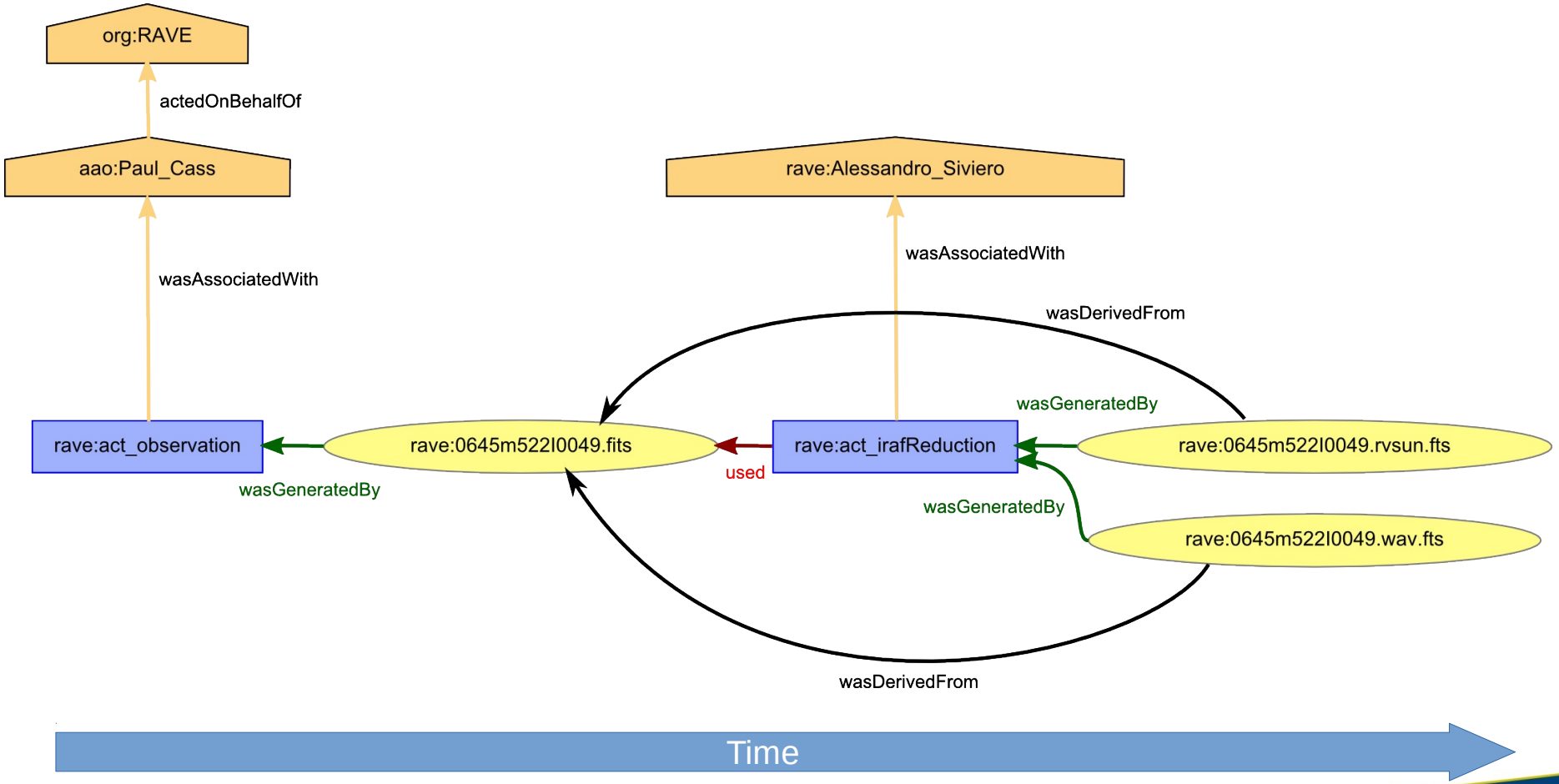
Example: Reduced RAVE fits-file

- Graph reordered, attributes hidden:

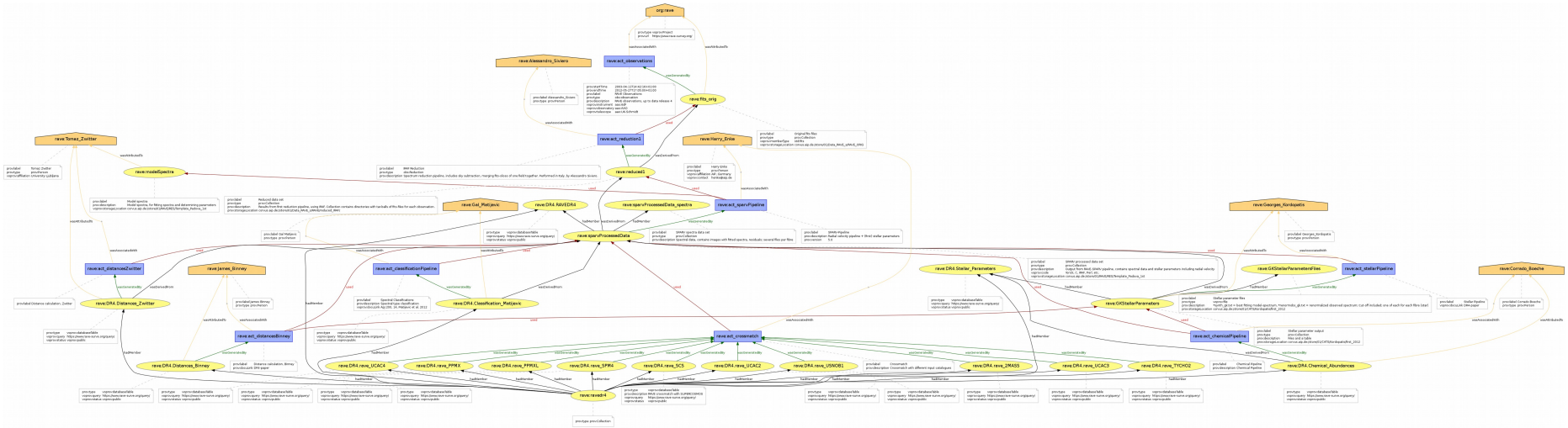


Example: Reduced RAVE fits-file

- Graph reordered, attributes hidden:

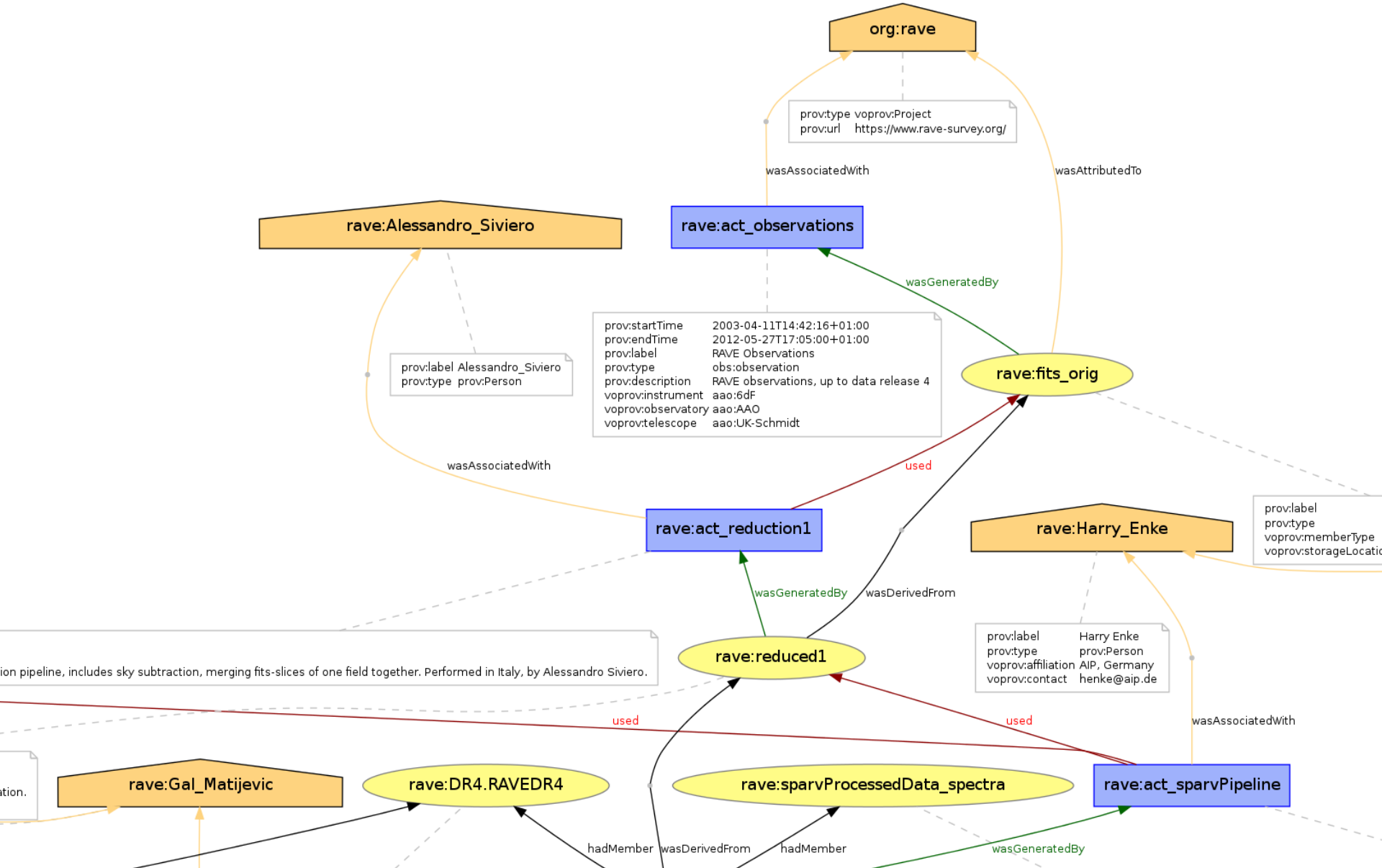


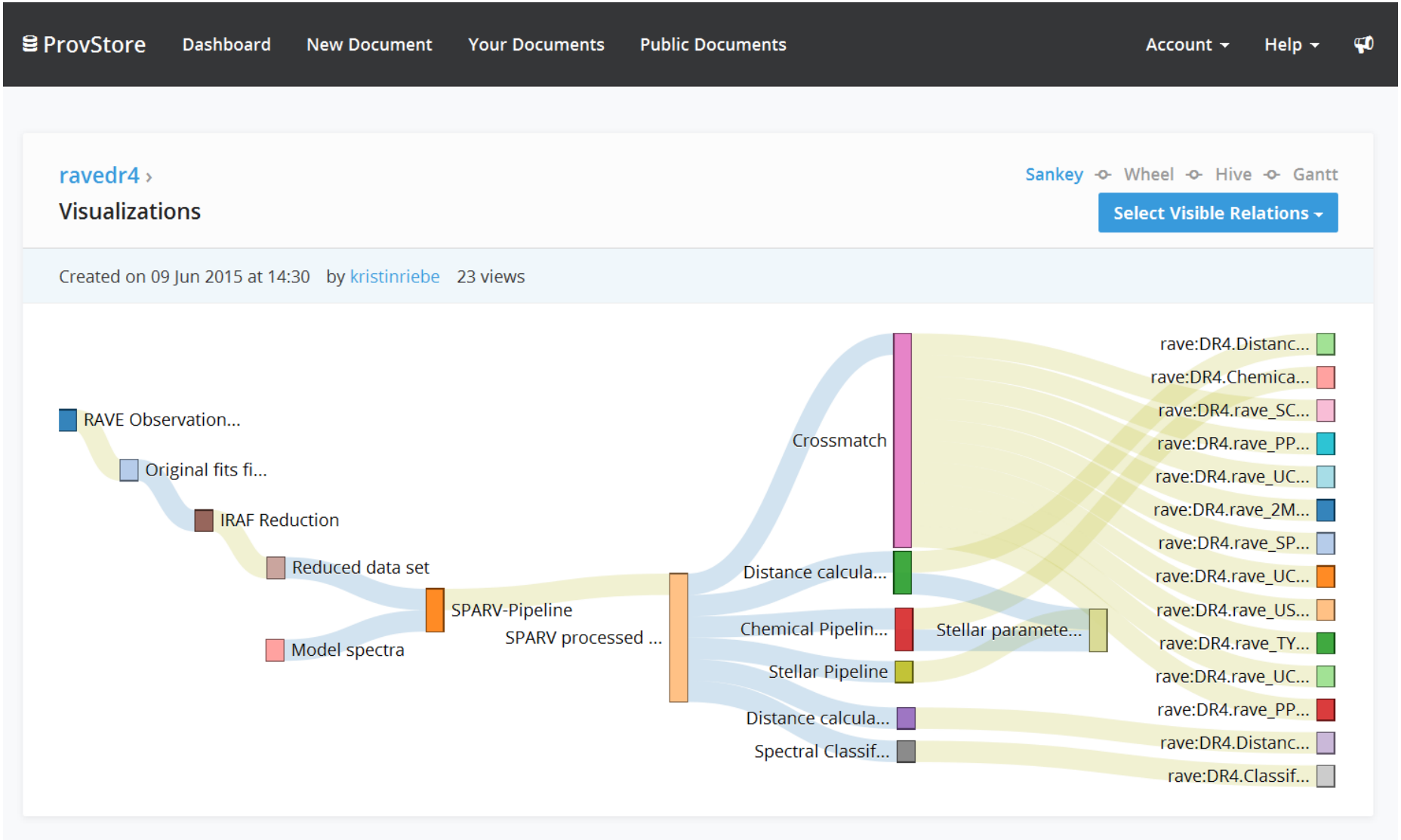
Example: RAVE database tables
(nearly complete history)



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

Example: RAVE database tables




ProvStore
Dashboard
New Document
Your Documents
Public Documents
Account
Help
ravedr4 ›
Visualizations
Sankey
Wheel
Hive
Gantt
Select Visible Relations
Created on 09 Jun 2015 at 14:30 by kristinriebe 23 views
RAVE Observation...
Original fits fi...
IRAF Reduction
Reduced data set
Model spectra
SPARV-Pipeline
SPARV processed ...
Crossmatch
Distance calcula...
Chemical Pipelin...
Stellar Pipeline
Distance calcula...
Spectral Classif...
Stellar paramete...
rave:DR4.Distanc...
rave:DR4.Chemica...
rave:DR4.rave_SC...
rave:DR4.rave_PP...
rave:DR4.rave_UC...
rave:DR4.rave_2M...
rave:DR4.rave_SP...
rave:DR4.rave_UC...
rave:DR4.rave_US...
rave:DR4.rave_TY...
rave:DR4.rave_UC...
rave:DR4.rave_PP...
rave:DR4.Distanc...
rave:DR4.Classif...

Example: RAVE

- RAVE data flow can be modeled in principle using W3C Prov
- RAVE data providers could provide web service:
  - takes id of entry in database table
  - returns complete history of that entry, tracing back through the different steps until the original files from the observation
  - using id as parameter

=> could be queryable via VO services as well?

Discussion

- Is W3C enough? Is it too abstract?
  - In short: Do you agree that we should adopt this?
  - Many implementations already exist, also see:
    - Southampton Provenance Suite, https://provenance.ecs.soton.ac.uk/ includes validator, converter, visualisation tools
    - Prov Implementation report: http://www.w3.org/TR/prov-implementations/
- VO:
  - know most common processes
  - => could predefine input/output of activities (roles)
    e.g. image stacking needs *n* fits-images as input, one fits-image as output
  - => could predefine standard entities (fits-files, VO-tables, …)
  - => use PROV-Template system or similar?

Discussion

- Which non-core relations should we include?
  - e.g. actedOnBehalfOf (between agents), wasInformedBy (between activities), …?
- Collections
  - Can we allow entities to belong to >1 collection?
- What about accessibility of intermediate data products?
  - Ignore non-public data? Flag them?
  - Provide only their header/metadata, more details on request?
- Instrument characteristics + ambient conditions
  - should be modeled elsewhere
  - use provenance data model only for relations between data/activities/agents
- More test cases!

Discussion

- Which non-core relations should we include?
  - e.g. actedOnBehalfOf (between agents), wasInformedBy (between activities), …?
- Collections
  - Can we allow entities to belong to >1 collection?
- What about accessibility of intermediate data products?
  - Ignore non-public data? Flag them?
  - Provide only their header/metadata, more details on request?
- Instrument characteristics + ambient conditions
  - should be modeled elsewhere
  - use provenance data model only for relations between data/activities/agents
- More test cases!

More discussions

- Discussion session on Wednesday afternoon, 4 pm
- More on W3C: TaPP workshop in Edinburgh, http://workshops.inf.ed.ac.uk/tapp2015/