Provenance of astronomical data

The IVOA Provenance Working Group:

Catherine Boisson
François Bonnarel
Johan Bregeon
Pierre Le Sidaner
Julien Lefaucheur
Mireille Louys
Markus Nullmeier
Ana Palacios
Kristin Riebe
Michèle Sanguillon
Mathieu Servillat

What is provenance?

- In general: tracking the history, origin of something:
  - art
  - food industry
  - information (data vis) on news webpage
  - scientific data!

- In astronomy: explain how data sets were produced:
  - Who created the data?
  - Which algorithm was used to produce it?
  - Which steps were undertaken to process the image?
  - Can I get access to the original, uncalibrated files from the observation?

Goals

- For a given data set, provenance should help to …
  - Discover steps of production
    Which processing steps have been done already?
  - Give attribution
    Who was involved in the project? Who can I ask about these data?
  - Aid in reprocessing
    But not necessarily: allow reprocessing on keypress
  - Aid in debugging
    Find possible error sources, e.g. check version of processing software, ambient conditions, telescope configuration, parameter settings, …
  - Allow to assess the quality of the data
  - Search in structured provenance metadata

What is provenance?

- From W3C, Prov-Overview:

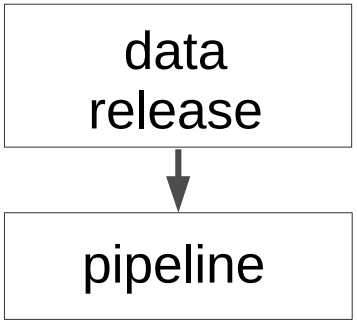
> Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness.

Example in astronomy

data release

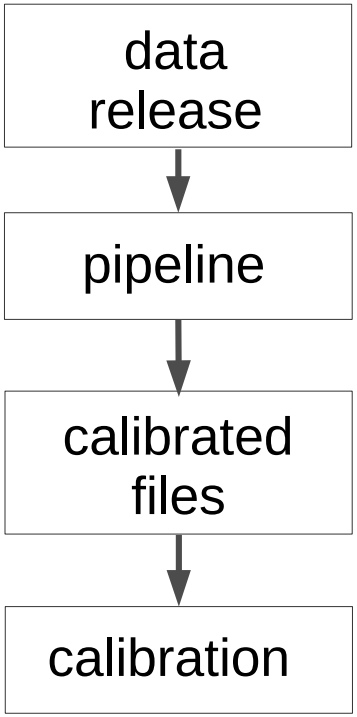
- Where is the data coming from?

Example in astronomy



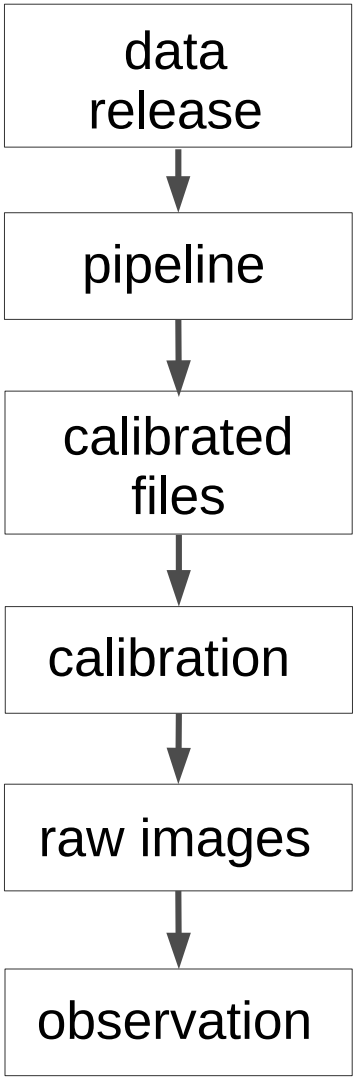
- Where is the data coming from?
- What were the input files for the pipeline?

Example in astronomy



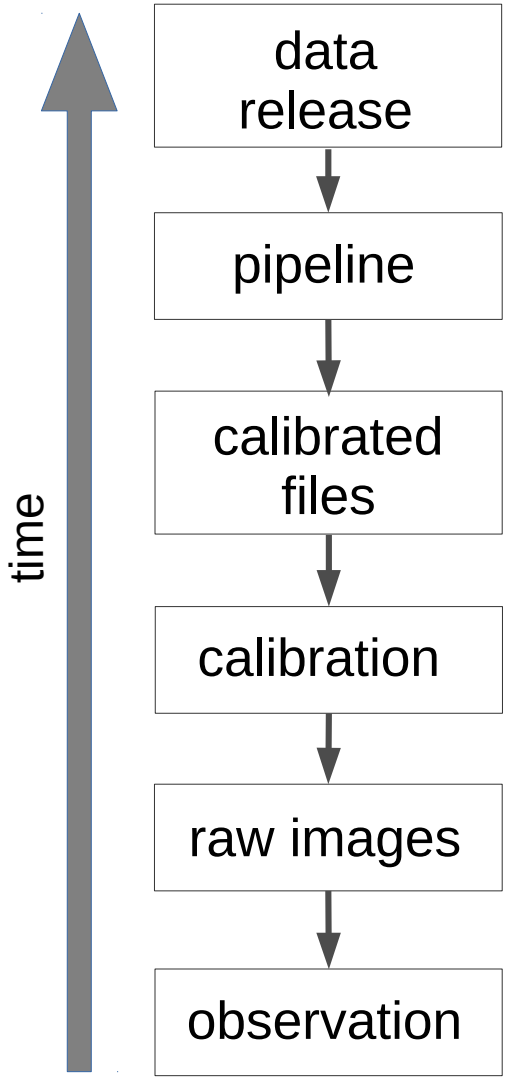
- Where is the data coming from?
- What were the input files for the pipeline?
- Have calibrated files been used for the pipeline?
- How were they calibrated?

Example in astronomy



- Where is the data coming from?
- What were the input files for the pipeline?
- Have calibrated files been used for the pipeline?
- How were they calibrated?
- Can I get the raw images?
- Were there perfect conditions during the observation?

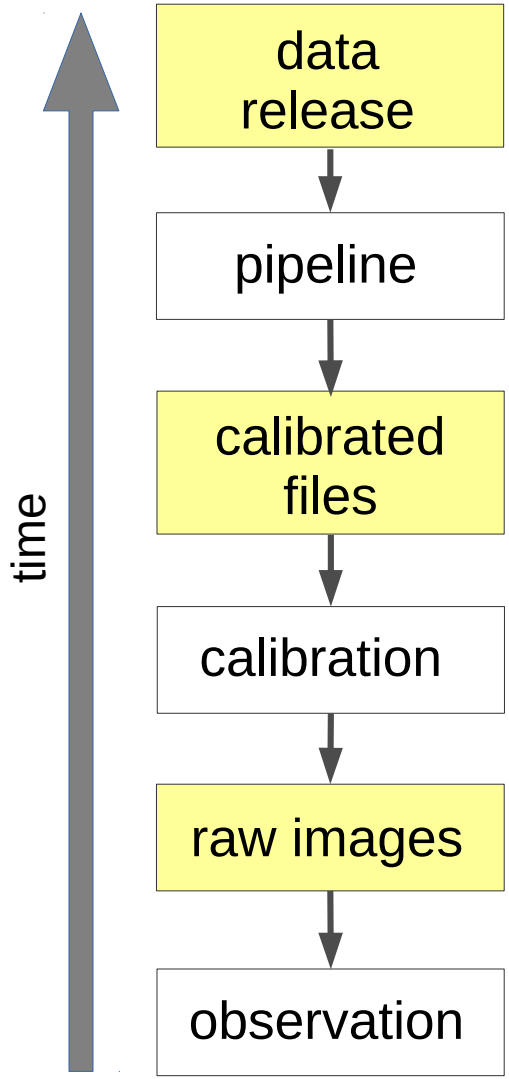
Example in astronomy



- Where is the data coming from?
- What were the input files for the pipeline?
- Have calibrated files been used for the pipeline?
- How were they calibrated?
- Can I get the raw images?
- Were there perfect seeing conditions during the observation?

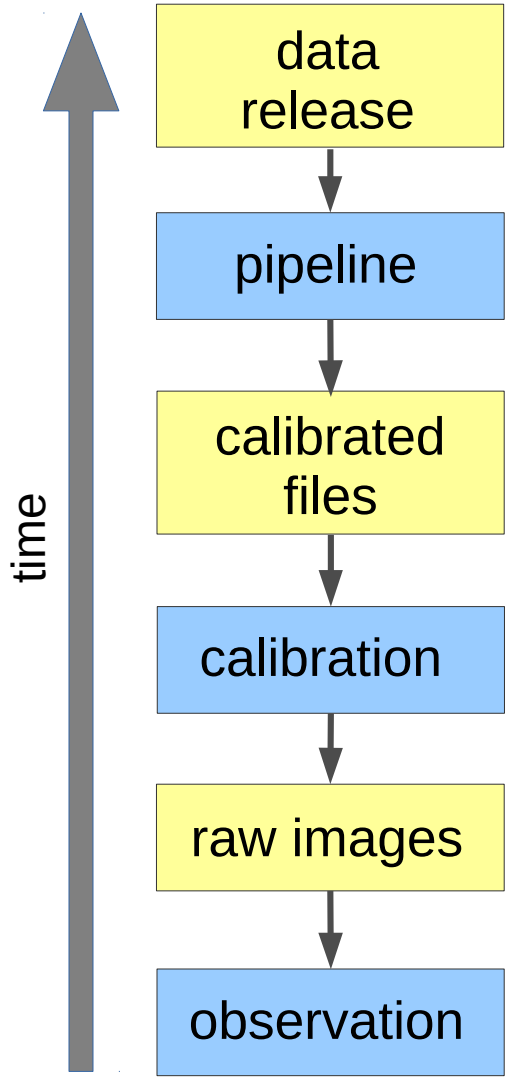
  => Track data back in time

Example in astronomy



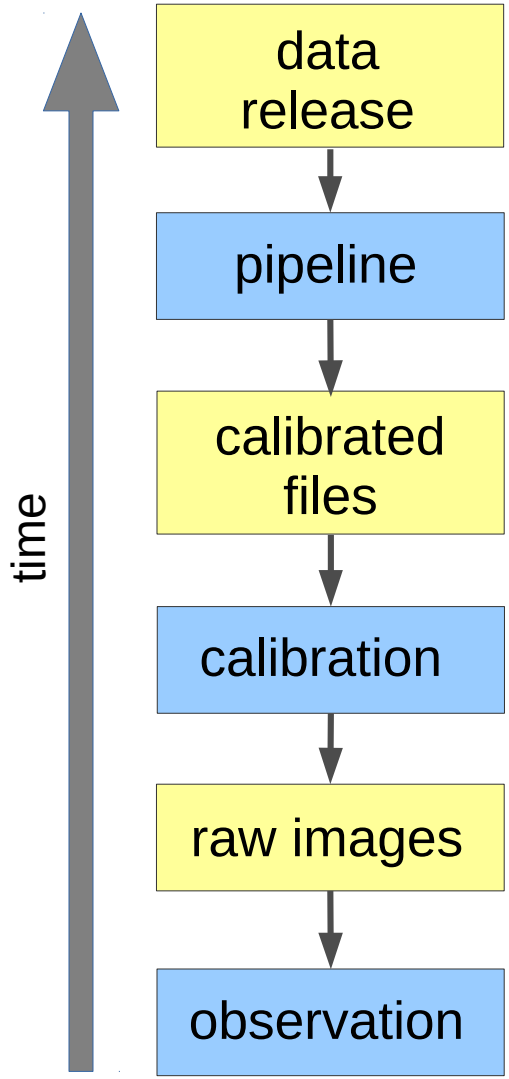
- identify data entities

Example in astronomy



- identify data entities
- identify processes (activities)

Example in astronomy



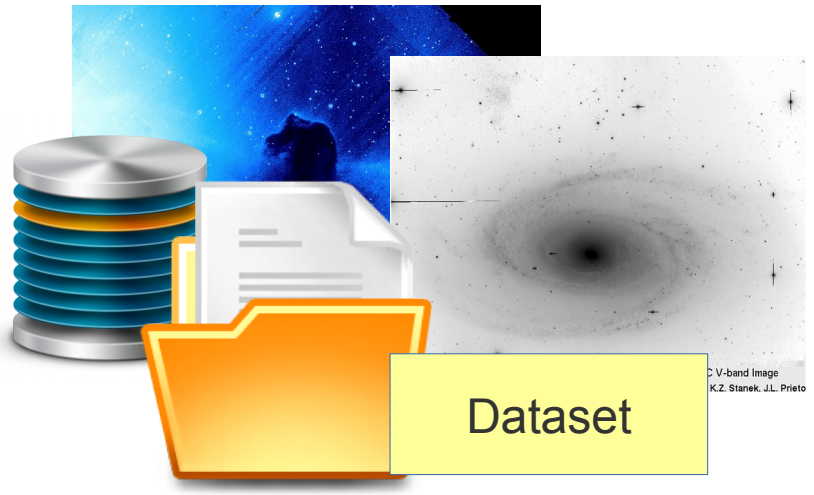
- identify data entities
- identify processes (activities)
- provenance is defined by the relations between data and activities
- provenance is about history => points backwards in time

Central provenance objects

- **Datasets:**
  fits files (images), votables, database tables, spectra, log files, parameters, ...

  DatasetDM:
  Dataset = *"a file or files which are considered to be a single deliverable"*

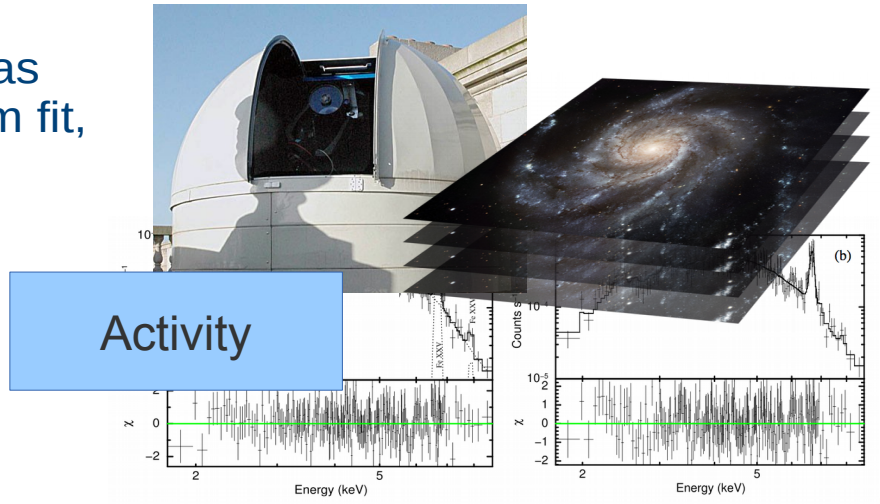
  Provenance:
  Dataset = one or more data entities with a common origin



- **Activities:**
  observations; processing steps like bias subtraction, image stacking, continuum fit, object extraction; simulations, ...

- **Persons/Organizations:**
  data creator, publisher, contact, ...

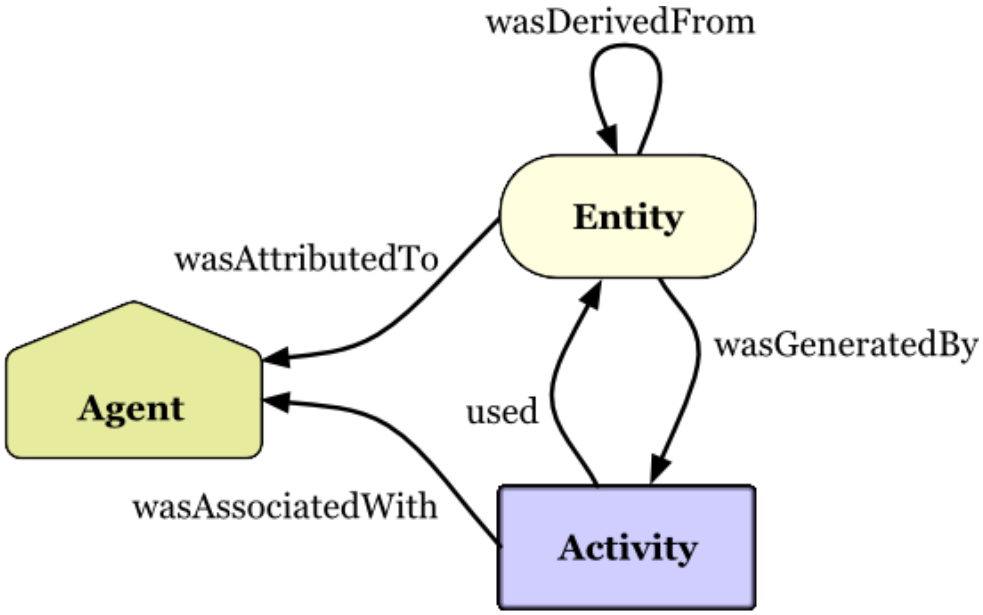
- ... also see ProvDM of W3C ...



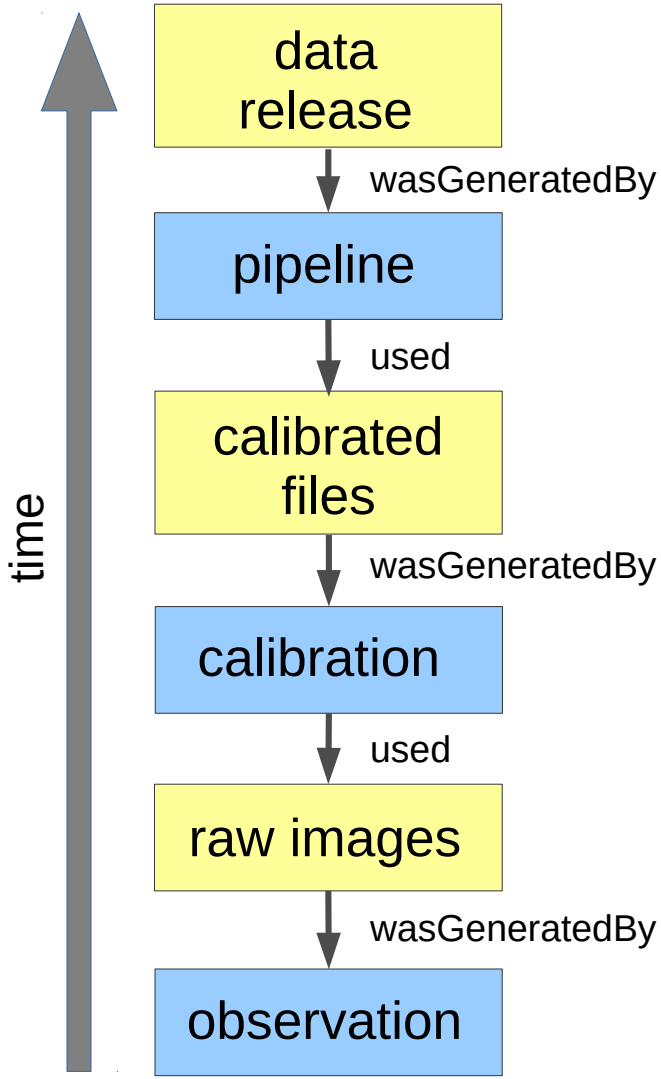
Provenance DM from W3C

http://www.w3.org/TR/prov-dm/, published 2013

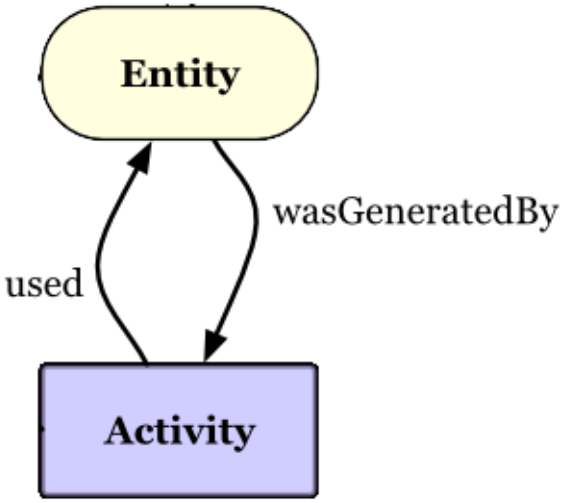
- 3 core classes:
  - Activity
  - Entity
  - Agent
- core relations:
  - used
  - wasGeneratedBy
  - wasDerivedFrom
  - wasAttributedTo
  - wasAssociatedWith
- + many more classes and relations



Example in astronomy

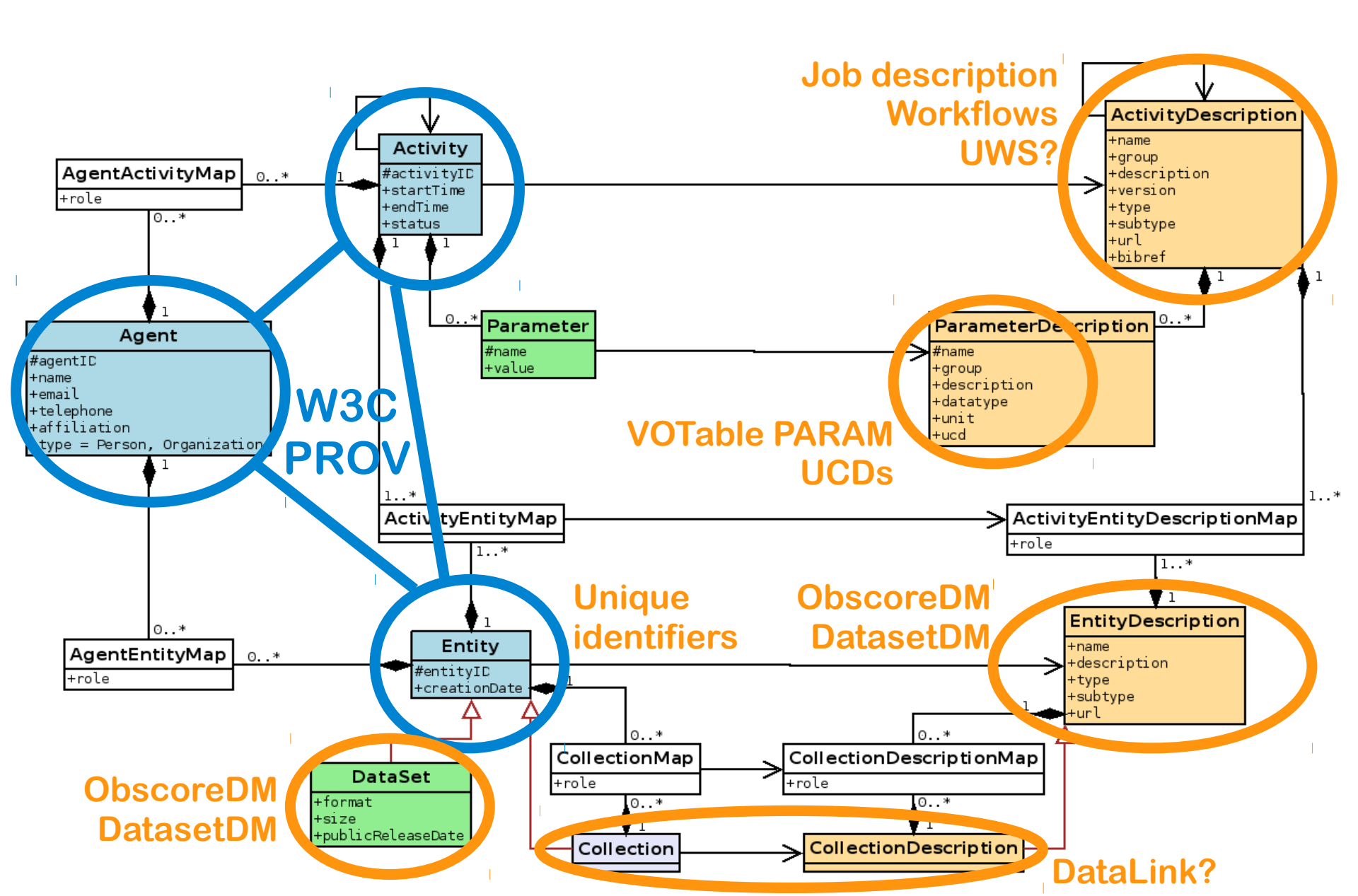


- input:
  data that is “used” by an activity
- output:
  data that “wasGeneratedBy” an activity

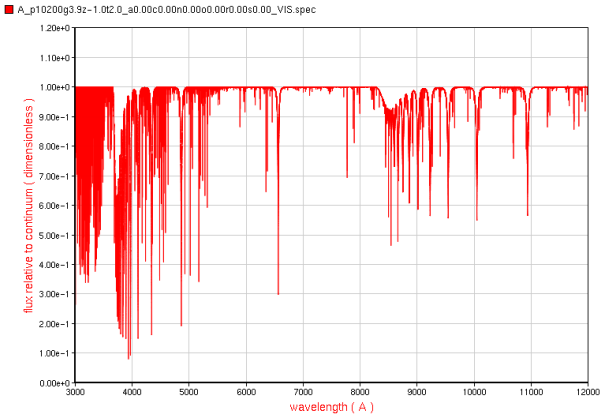


W3C or more?

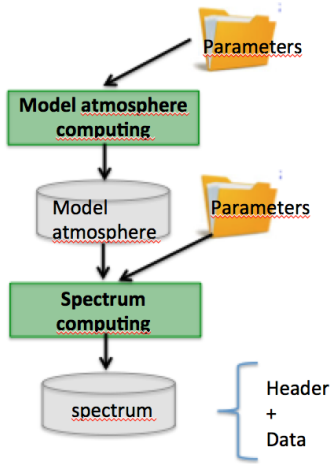
- Is W3C enough?
  - Many implementations already exist, also see:
    - Southampton Provenance Suite, https://provenance.ecs.soton.ac.uk/ includes validator, converter, visualisation tools
    - Prov Implementation report: http://www.w3.org/TR/prov-implementations/
- In astronomy:
  - know most common processes => predefine activities
  - => 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, …)

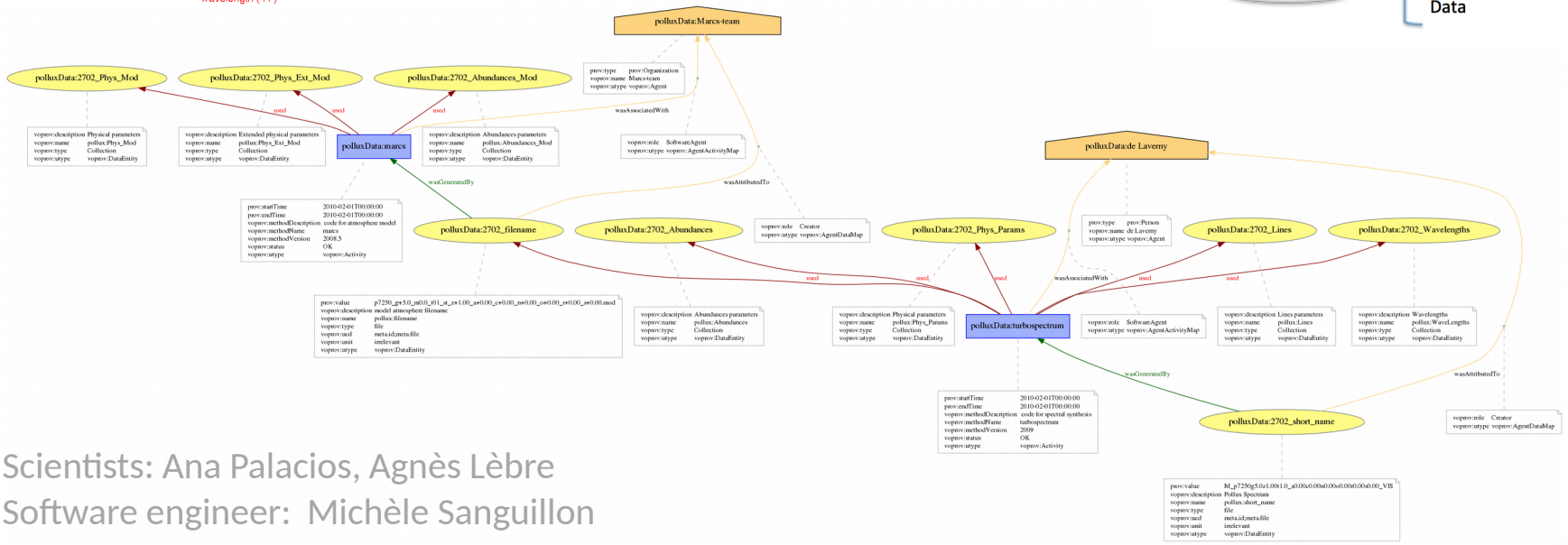

Job description
Workflows
UWS?
ActivityDescription
+name
+group
+description
+version
+type
+subtype
+url
+bibref
Activity
#activityID
+startTime
+endTime
+status
AgentActivityMap
+role
Agent
#agentID
+name
+email
+telephone
+affiliation
type = Person, Organization
W3C
PROV
Parameter
#name
+value
ParameterDescription
#name
+group
+description
+datatype
+unit
+ucd
VOTable PARAM
UCDs
ActivityEntityMap
ActivityEntityDescriptionMap
+role
Unique
identifiers
ObscoreDM
DatasetDM
AgentEntityMap
+role
Entity
#entityID
+creationDate
EntityDescription
+name
+description
+type
+subtype
+url
DataSet
+format
+size
+publicReleaseDate
ObscoreDM
DatasetDM
CollectionMap
+role
CollectionDescriptionMap
+role
Collection
CollectionDescription
DataLink?

Pollux use case



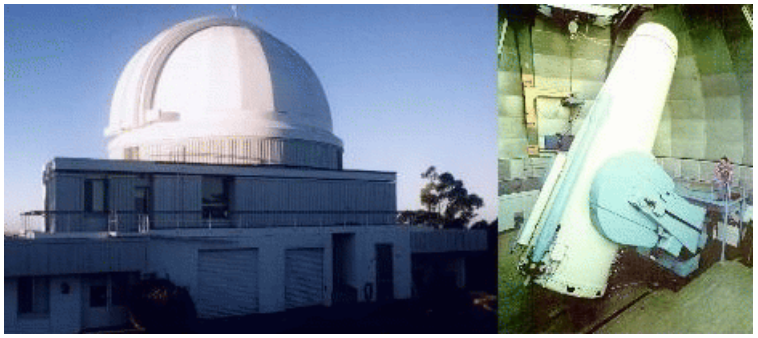
Database of more than 8000 very high resolution synthetic spectra in the optical domain (3000 Å to 12000 Å).



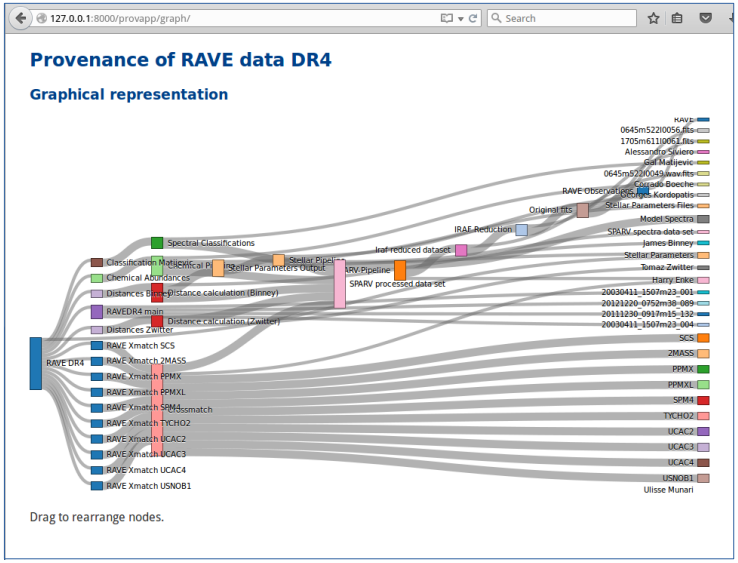


Scientists: Ana Palacios, Agnès Lèbre
Software engineer: Michèle Sanguillon

RAVE survey use case



- Radial velocity experiment
- multi-fibre spectroscopic survey of the southern hemisphere, 2003 - 2013
- different calibration, reduction and analysis steps
- radial velocities + other stellar properties for ~ half million stars
- use provenance to track history of datasets, where data is coming from

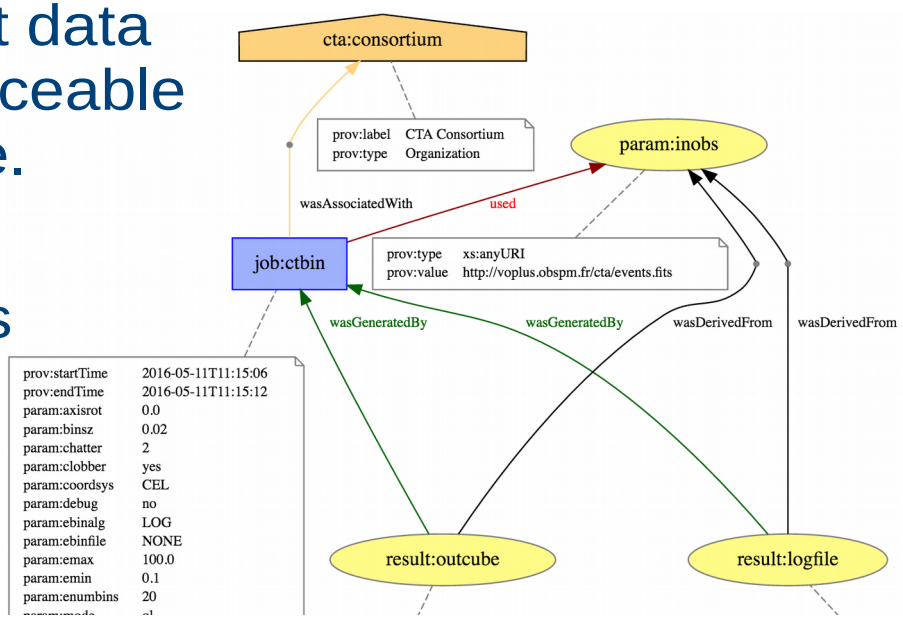


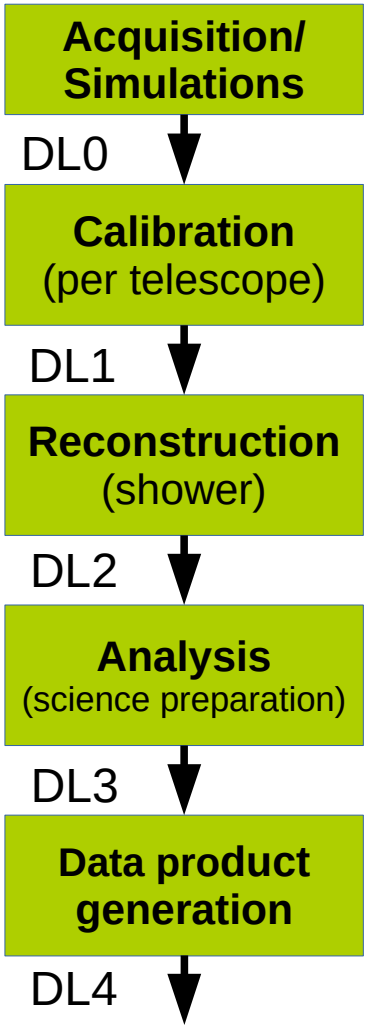
Leibniz-Institut für Astrophysik Potsdam (AIP)

@ Kristin Riebe

CTA use case

- Next Cherenkov Very High energy observatory
- **Open** observatory
- must ensure that data processing is traceable and reproducible.
- inform user on processing steps performed
- link to progenitor





cherenkov telescope array

l'Observatoire de Paris — LUTH

cnrs

PSL RESEARCH UNIVERSITY PARIS

@ Mathieu Servillat

Working group activities

http://wiki.ivoa.net/twiki/bin/view/IVOA/ObservationProvenanceDataModel

- IVOA Sesto splinter meeting, June 2015
- Provenance Day in Paris, April 2016
- IVOA Cape Town splinter meeting and DM session, May 2016
- Provenance Day in Heidelberg, June 2016
- **Next** in Paris, July or August 2016

Program of the last discussions:

- Data Model updates
- Structuring a **database** from the data model
- Storing/**serializing** the Activity/Entity Descriptions (VOTable, json, FITS frame...)
- **Access** to the Provenance database (TAP, specific access layer)
- Structure and content of the IVOA **working draft**
- **Roadmap** for Trieste (IVOA Interop in October) and beyond

What's your use case?

- Would you benefit from a standardized solution to expose your Provenance metadata?
  => contact us!
- What Provenance metadata do you need to expose?
- Does it fit in the Provenance Data Model?
- How would you store Provenance metadata?
  - Files (FITS header? FIT frame? VOTable? XML? JSON?)
  - Database
- How would you query the Provenance metadata?
  - Search for progenitors
  - Detailed search on execution context (nodes, resources), dates
  - Detailed search on activity/entity types