Leibniz-Institut für
Astrophysik Potsdam

Provenance Data Model

Introduction

DM session at IVOA InterOp

May 2016, Cape Town

Kristin Riebe, AIP, GAVO

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:

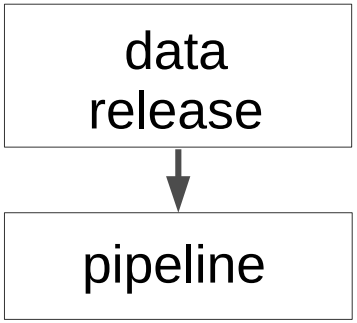
> Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness.

Example in astronomy

data release

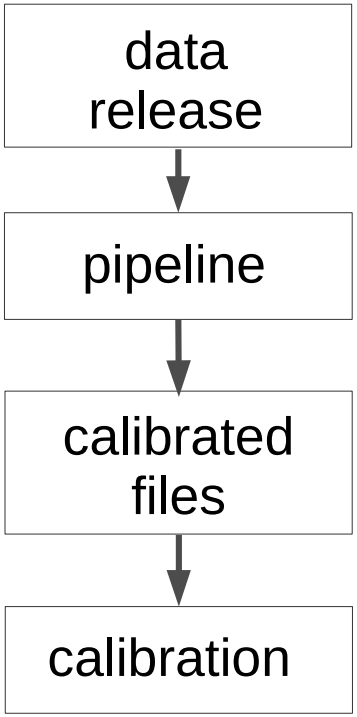
- Where is the data coming from?

Example in astronomy



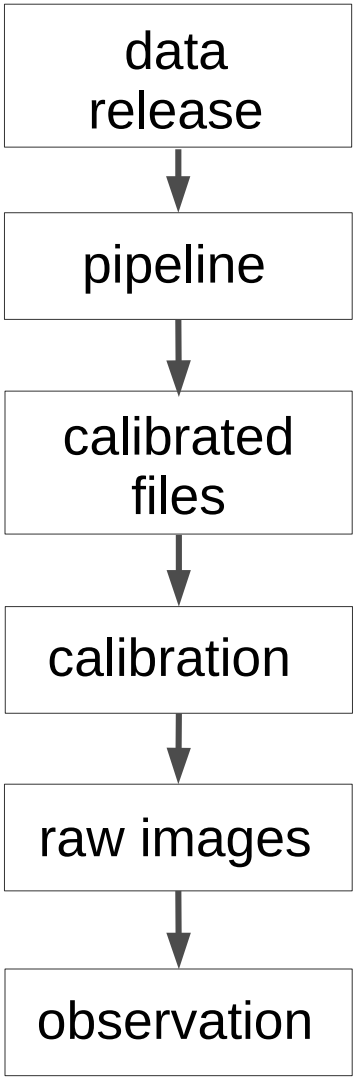
- Where is the data coming from?
- What were the input files for the pipeline?

Example in astronomy



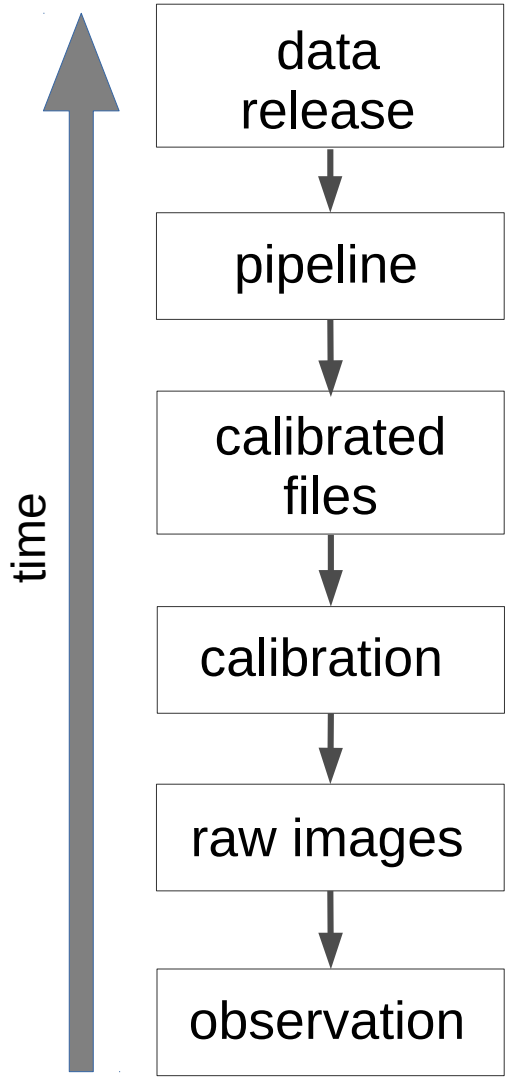
- Where is the data coming from?
- What were the input files for the pipeline?
- Have calibrated files been used for the pipeline?
- How were they calibrated?

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 seeing conditions during the observation?

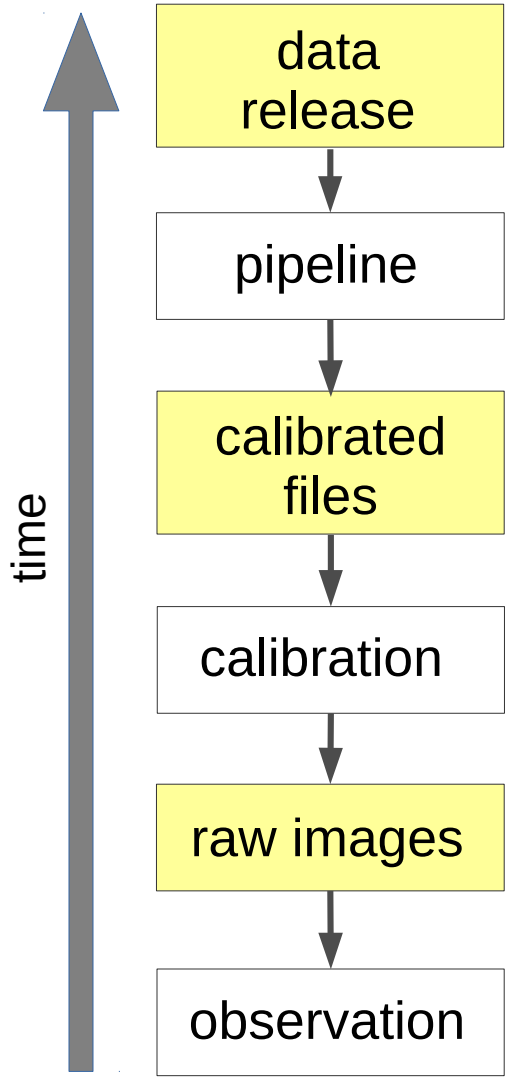
Example in astronomy



- 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?

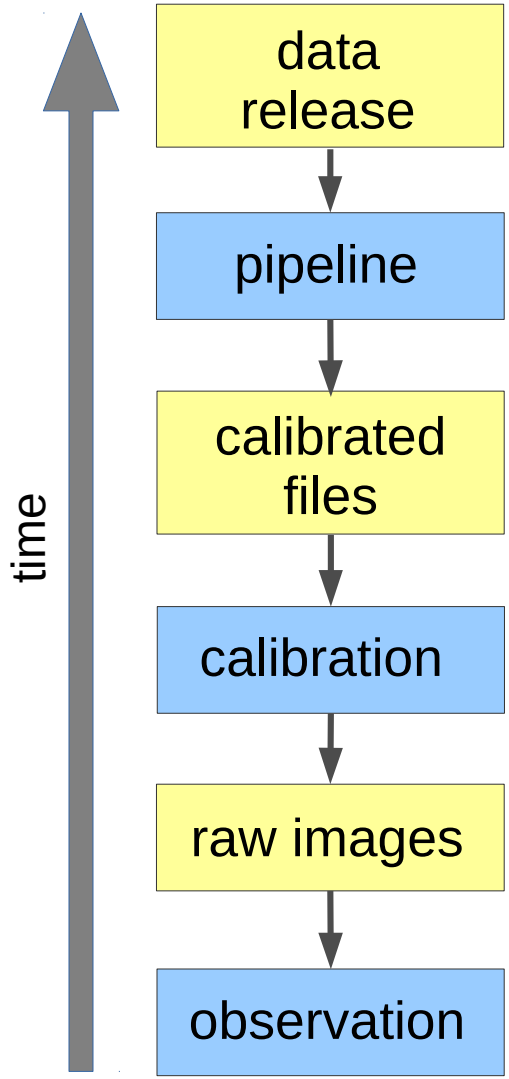
  => Track data back in time

Example in astronomy



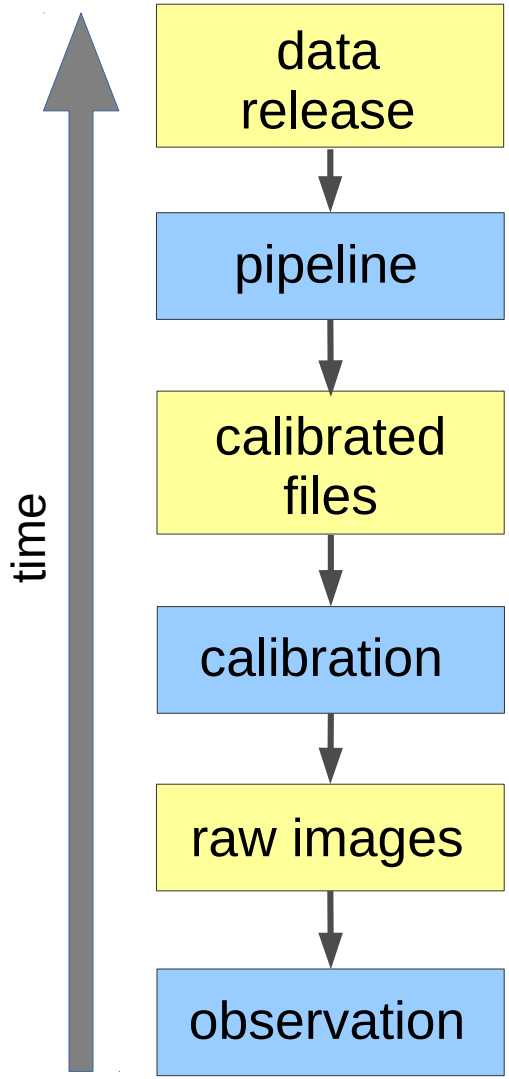
- identify data entities

Example in astronomy



- identify data entities
- identify processes (activities)

Example in astronomy



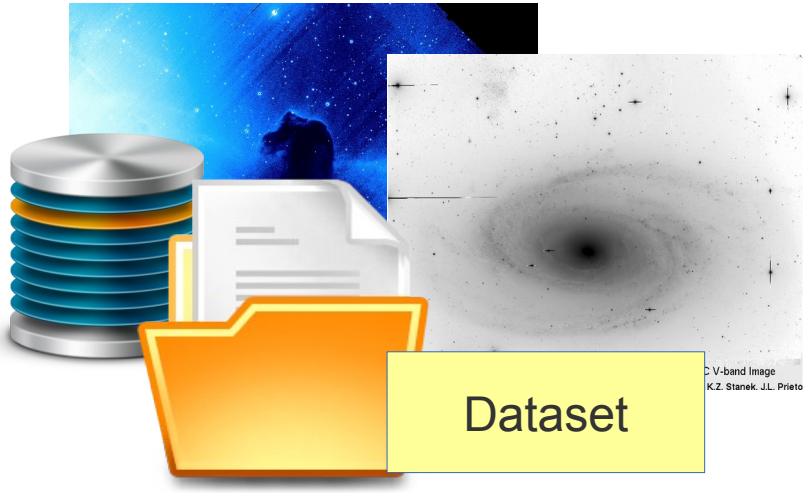
- identify data entities
- identify processes (activities)
- provenance is defined by the relations between data and activities
- provenance is about history
  => points backwards in time

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"*

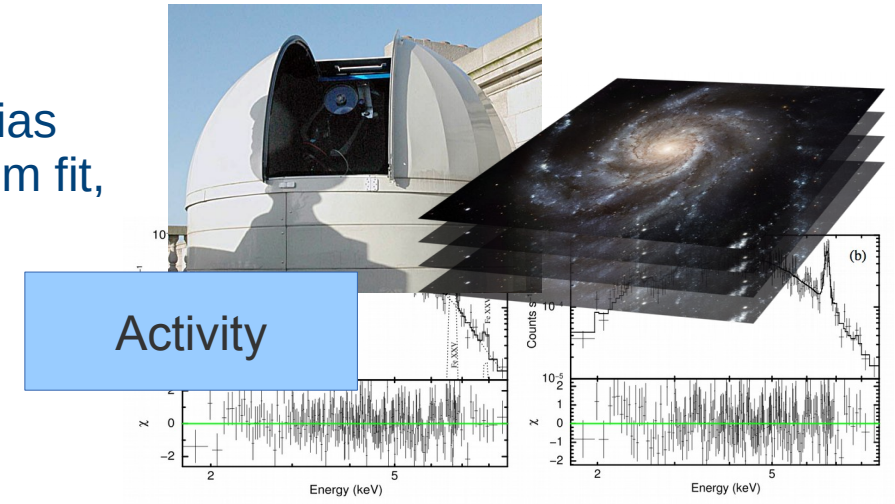
  Provenance:
  Dataset = one or more data entities with a common origin



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

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

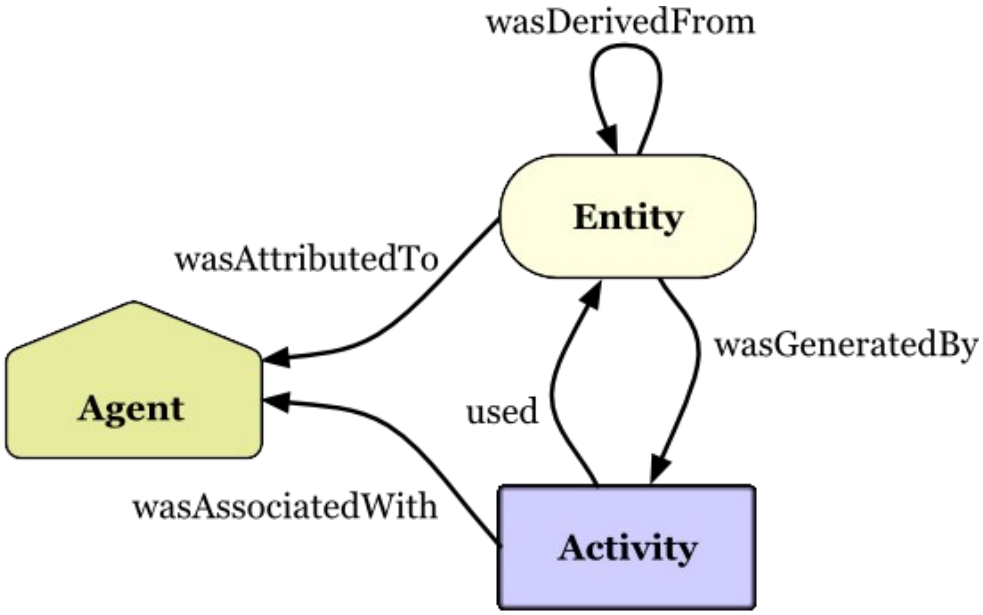
- ... also see ProvDM of W3C ...



Provenance DM from W3C

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

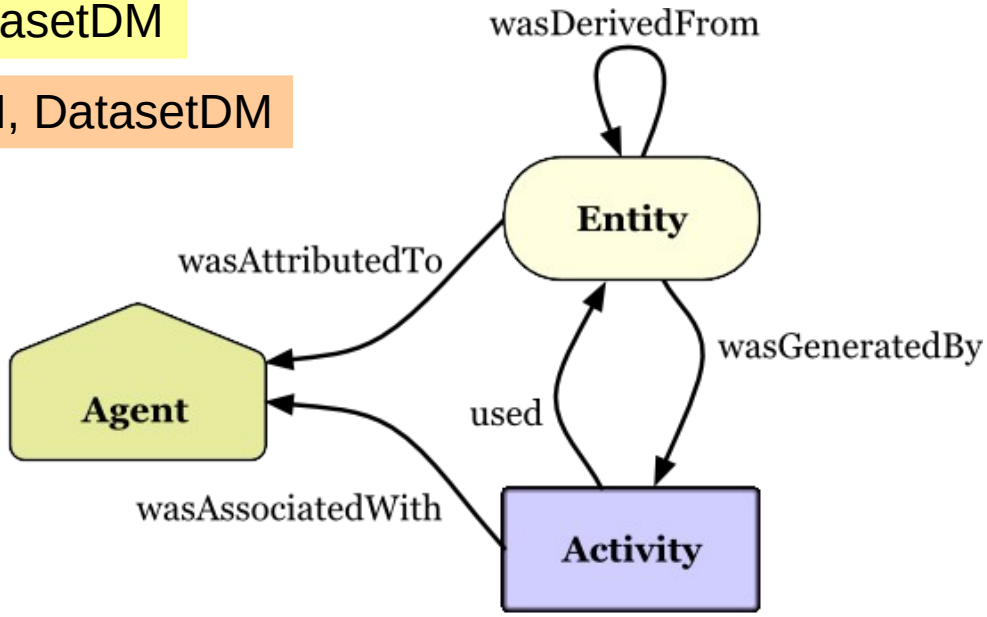
- 3 core classes:
  - Activity
  - Entity
  - Agent
- core relations:
  - used
  - wasGeneratedBy
  - wasDerivedFrom
  - wasAttributedTo
  - wasAssociatedWith
- + many more classes and relations



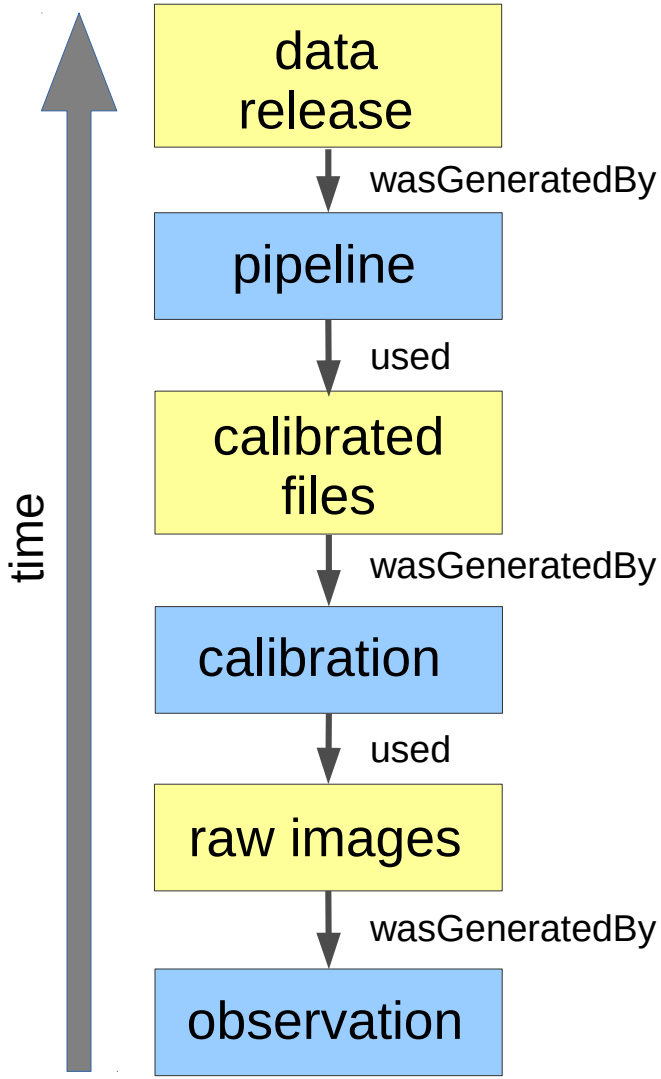
Provenance DM from W3C

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

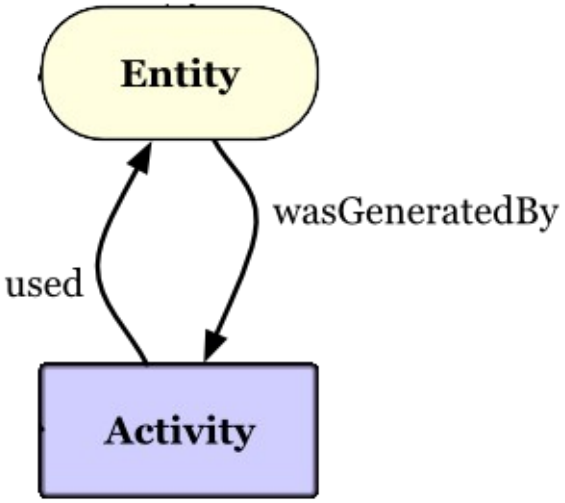
- 3 core classes:
  - Activity "Experiment" in SimDM
  - Entity "DataSet" in DatasetDM
  - Agent "Party" in SimDM, DatasetDM
- core relations:
  - used
  - wasGeneratedBy
  - wasDerivedFrom
  - wasAttributedTo
  - wasAssociatedWith
- + many more classes and relations



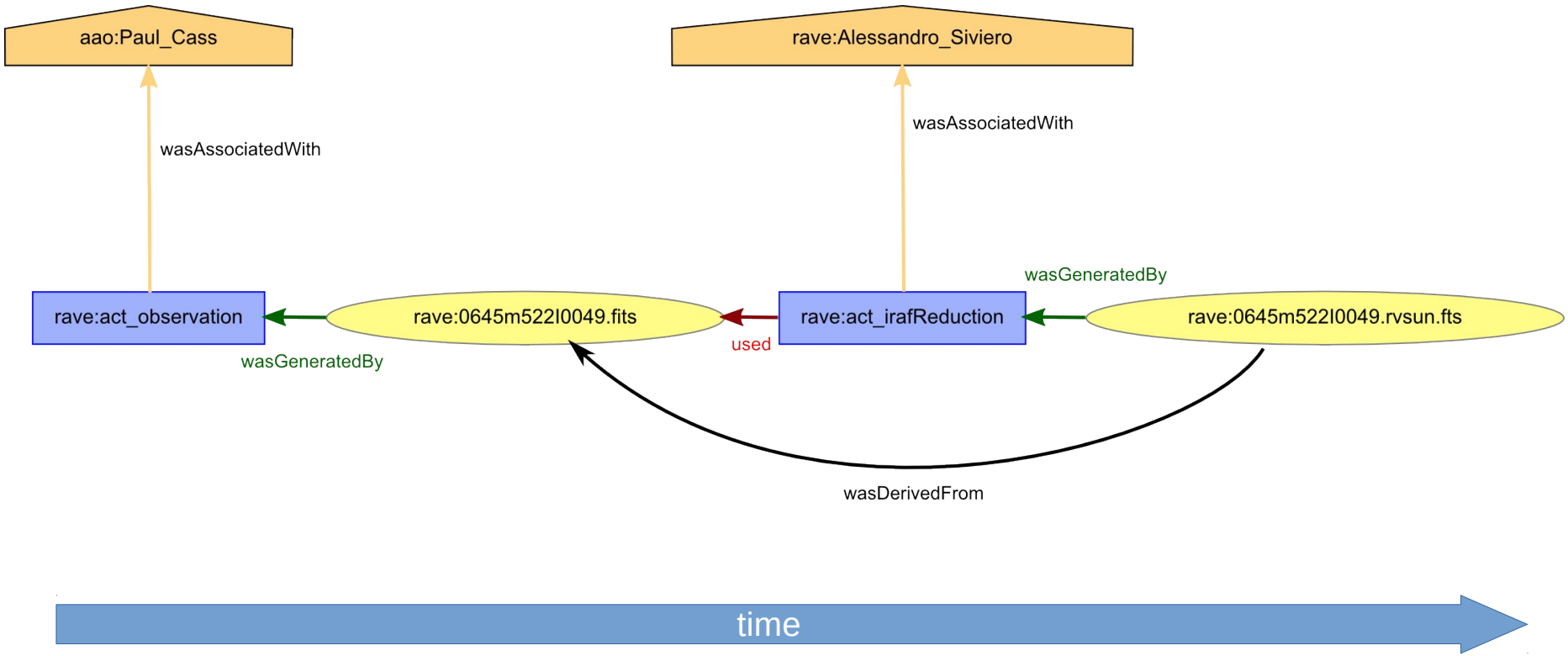
Example in astronomy



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



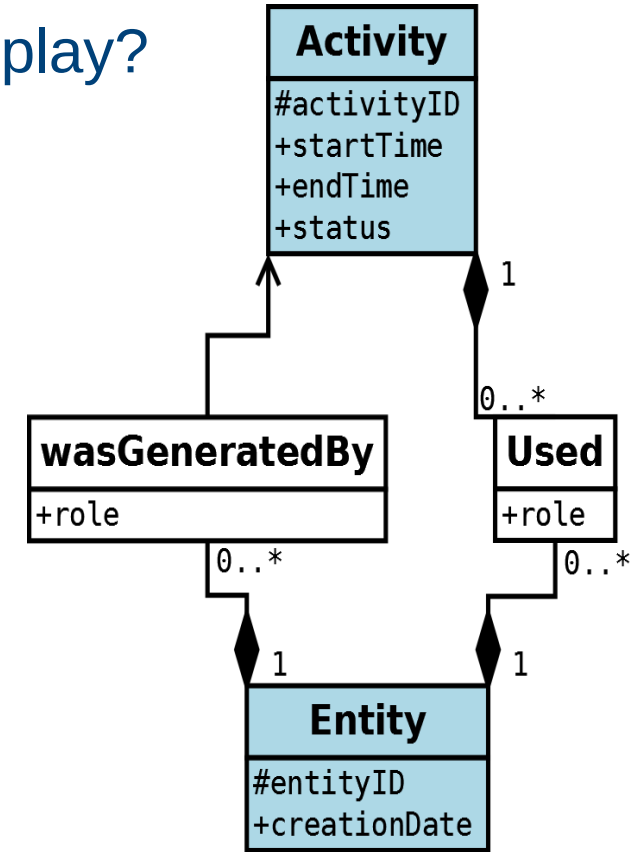
Example: Reduction of a fits file (RAVE)



Qualified relations/mapping classes

- multiple input data: which role do they play?
- example:
  - activity: sky subtraction;
    subtract one image from the other
  - input: image_A, image_B
  - output: image_C
- roles need to be given for each input (and output)

  => add qualified relation or mapping class in between

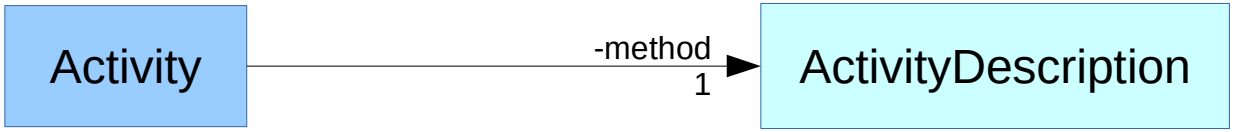


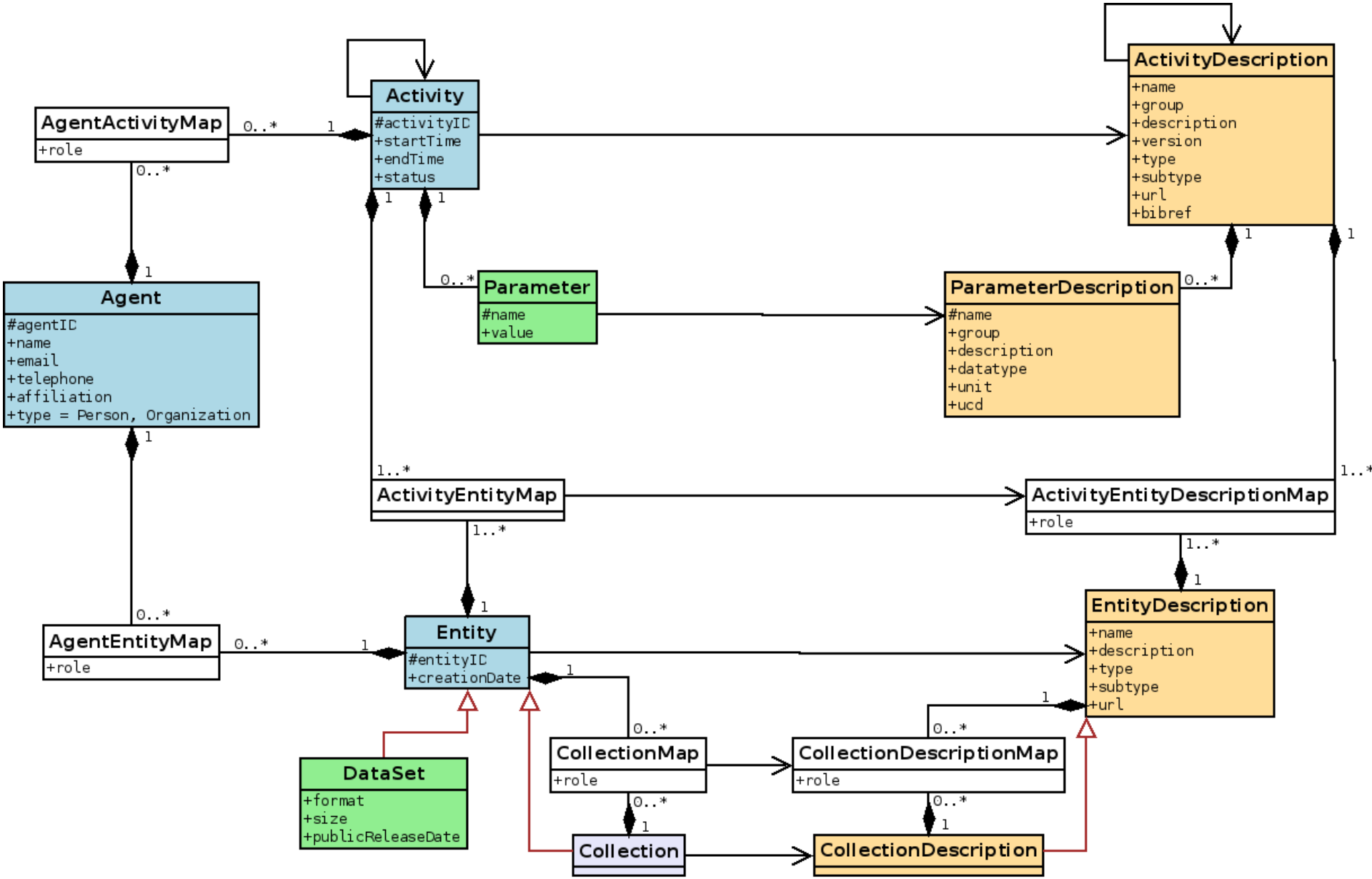
W3C or more?

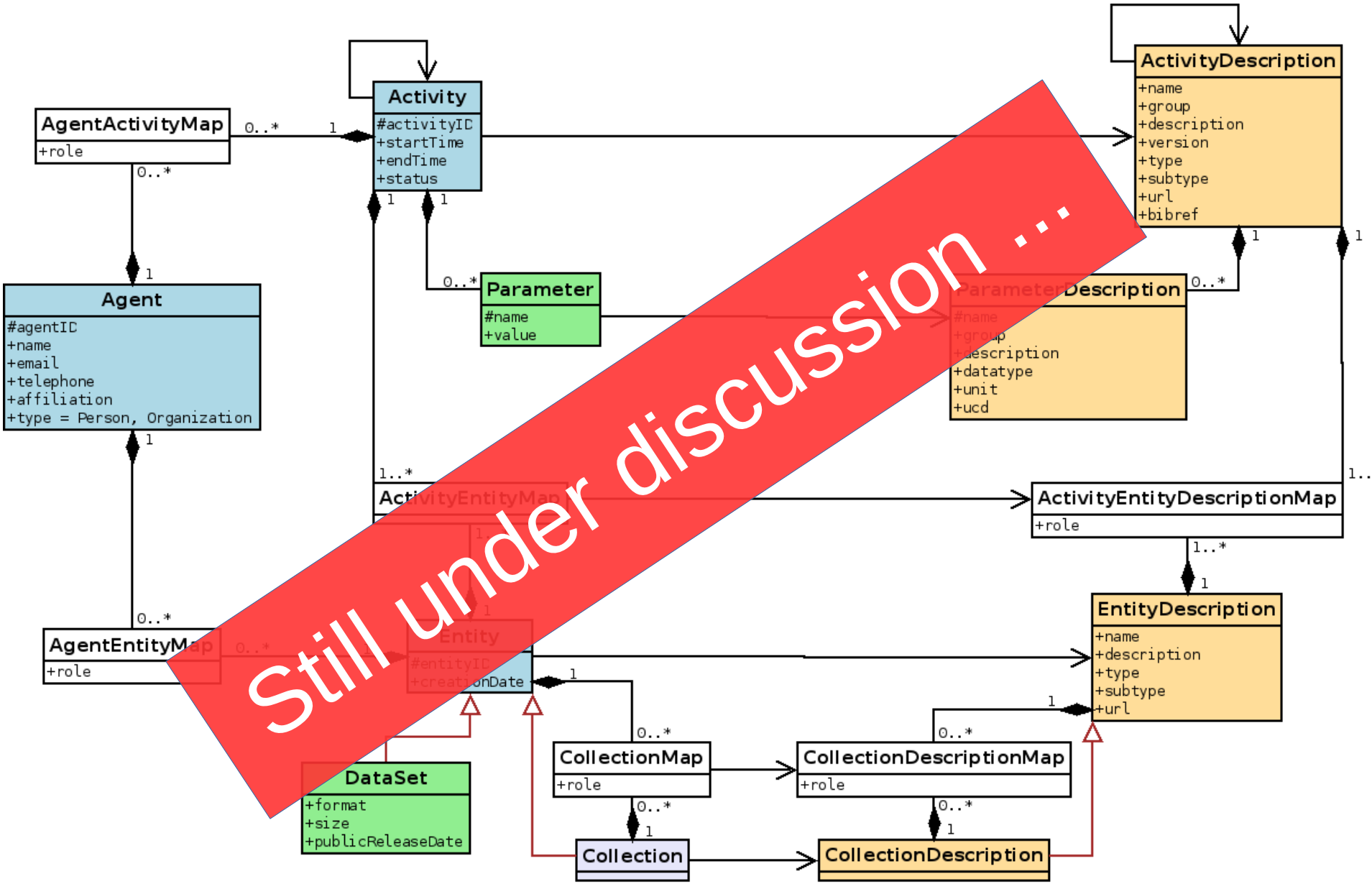
- Is W3C enough?
  - Many implementations already exist, also see:
    - Southampton Provenance Suite, https://provenance.ecs.soton.ac.uk/ includes validator, converter, visualisation tools
    - Prov Implementation report: http://www.w3.org/TR/prov-implementations/
- In astronomy:
  - know most common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

A model with prototypes

- looking at **SimDM**:
  - includes provenance for simulation data
  - two part concept for main part:
    - **Experiment:** processing, simulation etc., execution of an experiment
    - **Protocol:** design of experiment, description, reusable prototype
- adopt same structure here, but replace terms
    - Experiment => "**Activity**" (W3C)
    - Protocol => "**ActivityDescription**"
- each Activity has exactly one ActivityDescription; same for Dataset




AgentActivityMap
+role
0..*
1
Activity
#activityID
+startTime
+endTime
+status
ActivityDescription
+name
+group
+description
+version
+type
+subtype
+url
+bibref
Agent
#agentID
+name
+email
+telephone
+affiliation
+type = Person, Organization
Parameter
#name
+value
ParameterDescription
#name
+group
+description
+datatype
+unit
+ucd
ActivityEntityMap
ActivityEntityDescriptionMap
+role
AgentEntityMap
+role
Entity
#entityID
+creationDate
EntityDescription
+name
+description
+type
+subtype
+url
DataSet
+format
+size
+publicReleaseDate
CollectionMap
+role
CollectionDescriptionMap
+role
Collection
CollectionDescription


AgentActivityMap
+role
Activity
#activityID
+startTime
+endTime
+status
ActivityDescription
+name
+group
+description
+version
+type
+subtype
+url
+bibref
Agent
#agentID
+name
+email
+telephone
+affiliation
+type = Person, Organization
Parameter
#name
+value
ParameterDescription
#name
+group
+description
+datatype
+unit
+ucd
ActivityEntityMap
ActivityEntityDescriptionMap
+role
AgentEntityMap
+role
Entity
#entityID
+creationDate
EntityDescription
+name
+description
+type
+subtype
+url
DataSet
+format
+size
+publicReleaseDate
CollectionMap
+role
CollectionDescriptionMap
+role
Collection
CollectionDescription
Still under discussion ...