



Leibniz-Institut für
Astrophysik Potsdam

IVOA Provenance Data Model

Current status

InterOp, 21th October 2016, Trieste

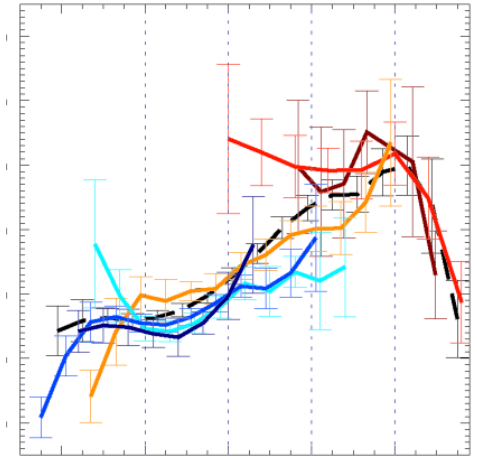
Kristin Riebe

IVOA Data Model Working Group

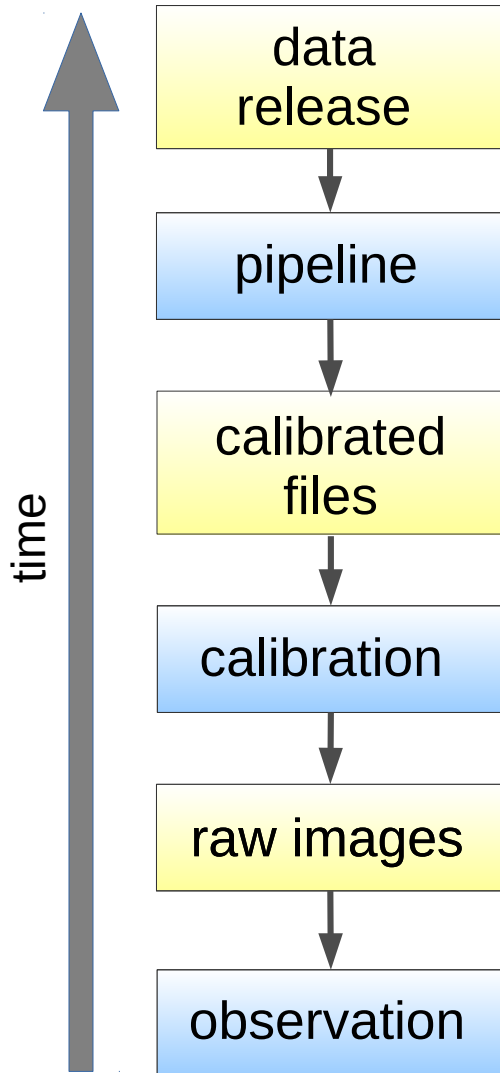


What is provenance?

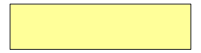
- In general: tracking the history, origin of something:
 - art
 - food industry
 - information (data vis) on news webpage
 - scientific data!
- In astronomy: explain how data sets were produced:
 - Who created the data?
 - Which algorithm was used to produce it?
 - Which steps were undertaken to process the image?
 - Can I get access to the original, uncalibrated files from the observation?



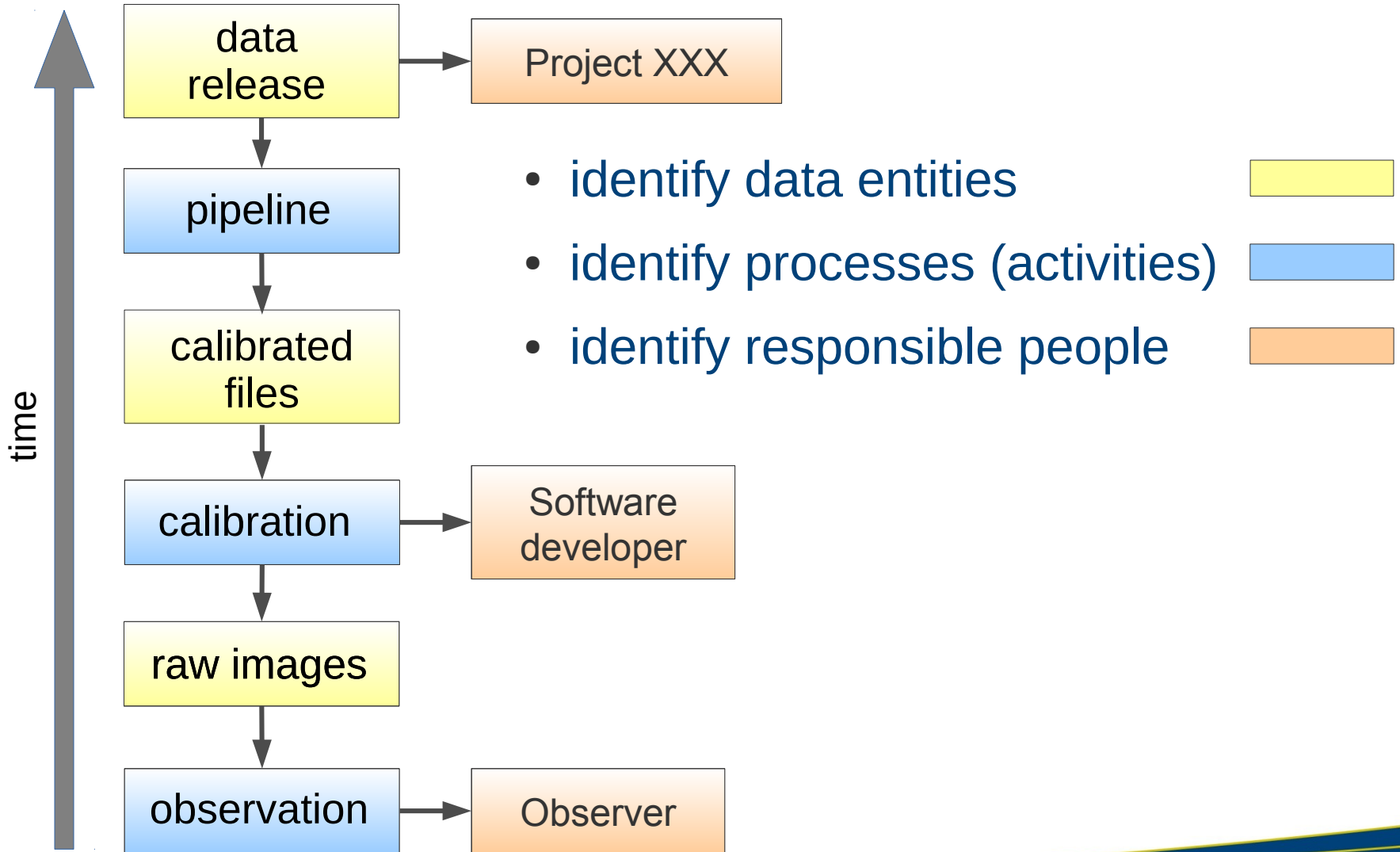
Example in astronomy



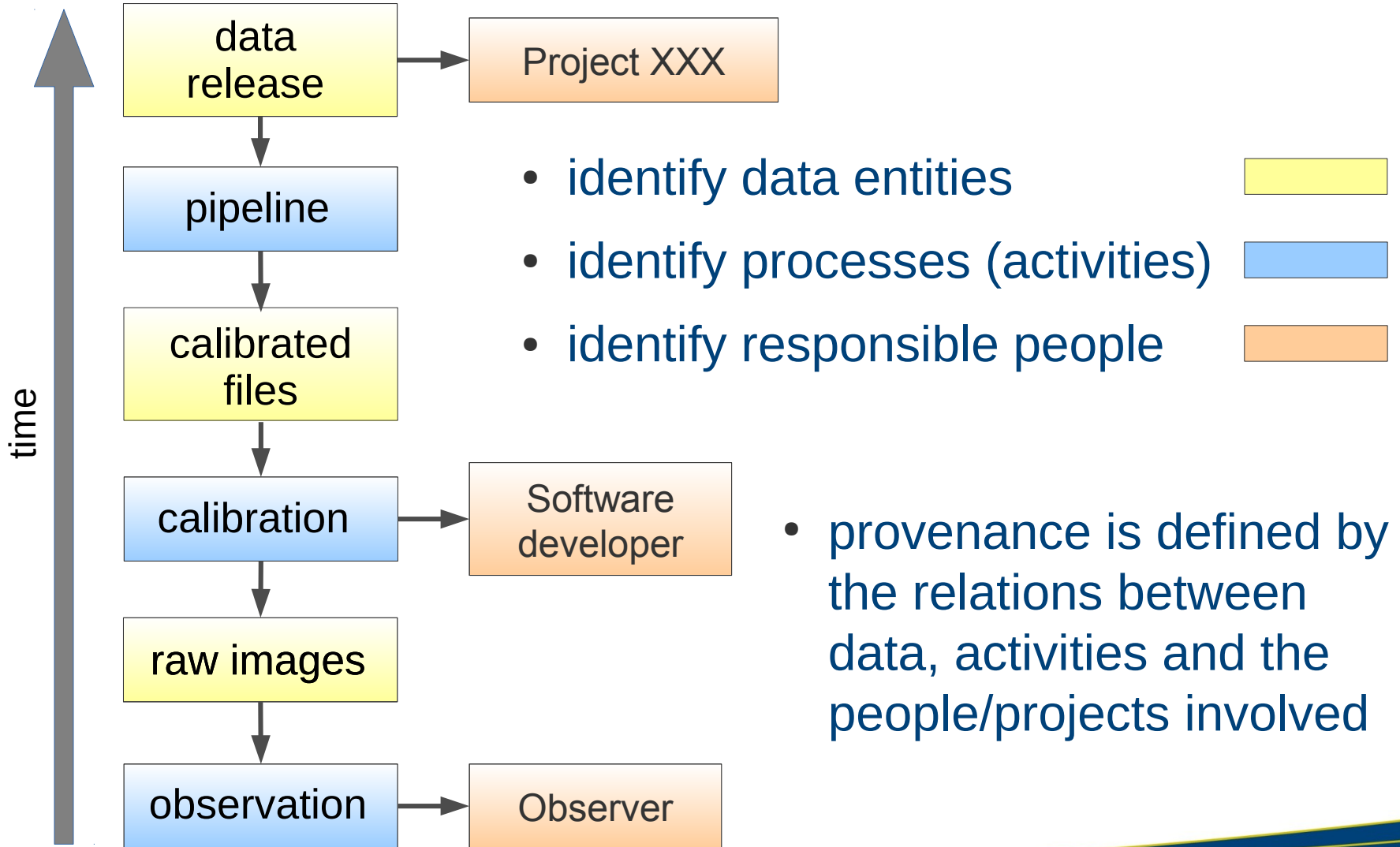
- identify data entities
- identify processes (activities)
- identify responsible people



Example in astronomy

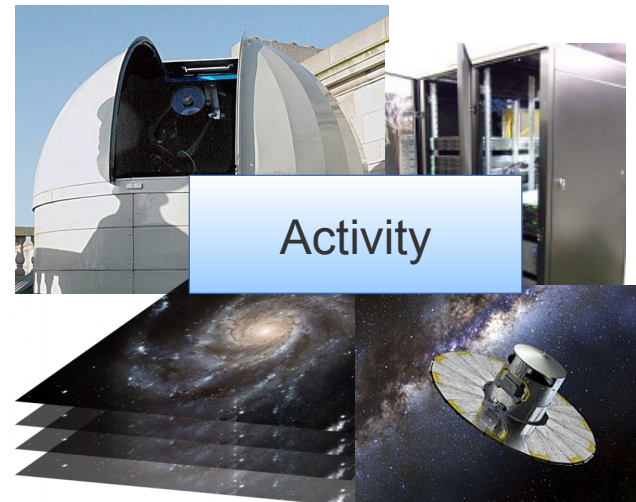
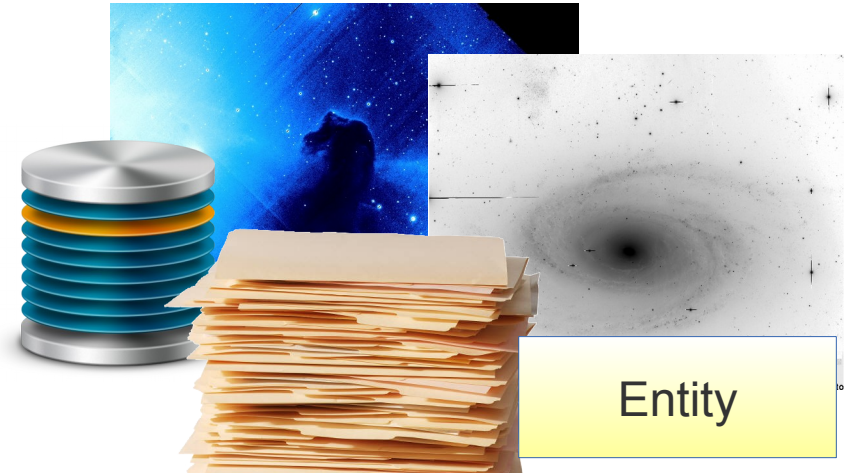


Example in astronomy



Examples for core objects

- **Entities (datasets):**
images, catalogs,
database tables, spectra,
log files, parameters, ...
- **Activities:**
observations;
processing steps like bias subtraction,
image stacking, continuum fit, object
extraction; simulations, ...
- **Persons/Organizations:**
creator, publisher, developer, ...



Provenance DM core classes

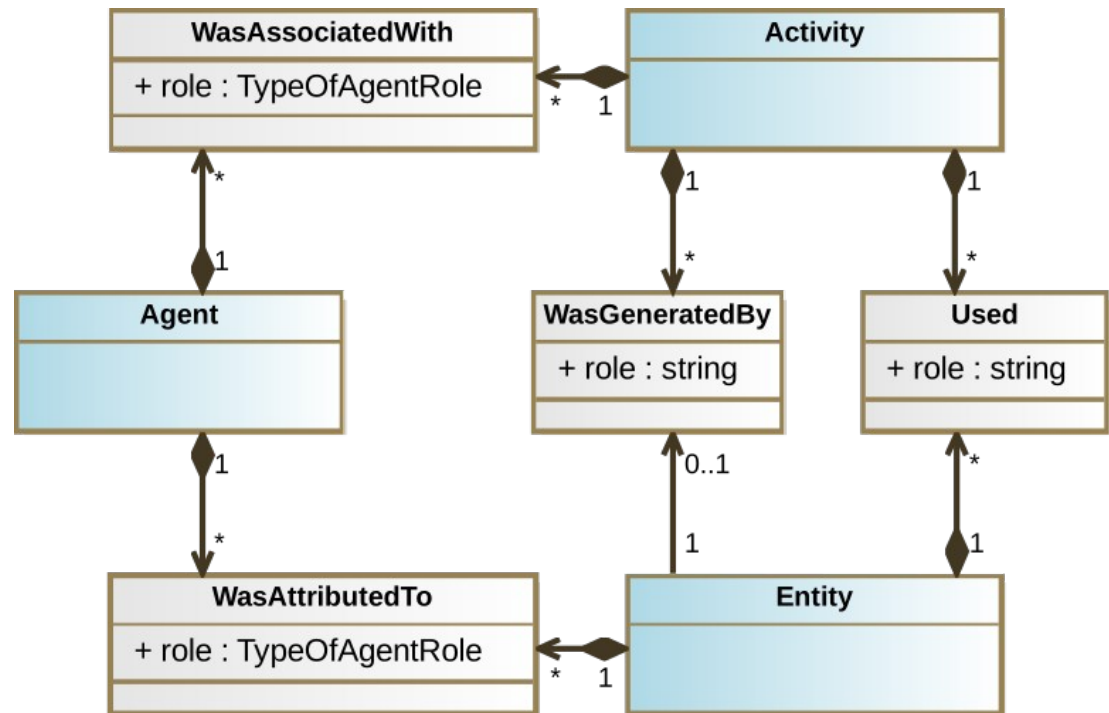
- same core classes as in W3C ProvDM model:
 - <http://www.w3.org/TR/prov-dm/>, published 2013

- 3 core classes:

- Activity
- Entity
- Agent

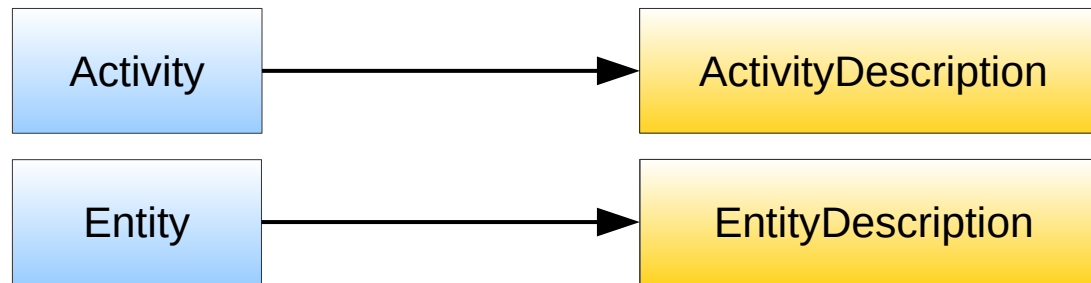
- core relations:

- used
- wasGeneratedBy
- wasDerivedFrom
- wasAttributedTo
- wasAssociatedWith



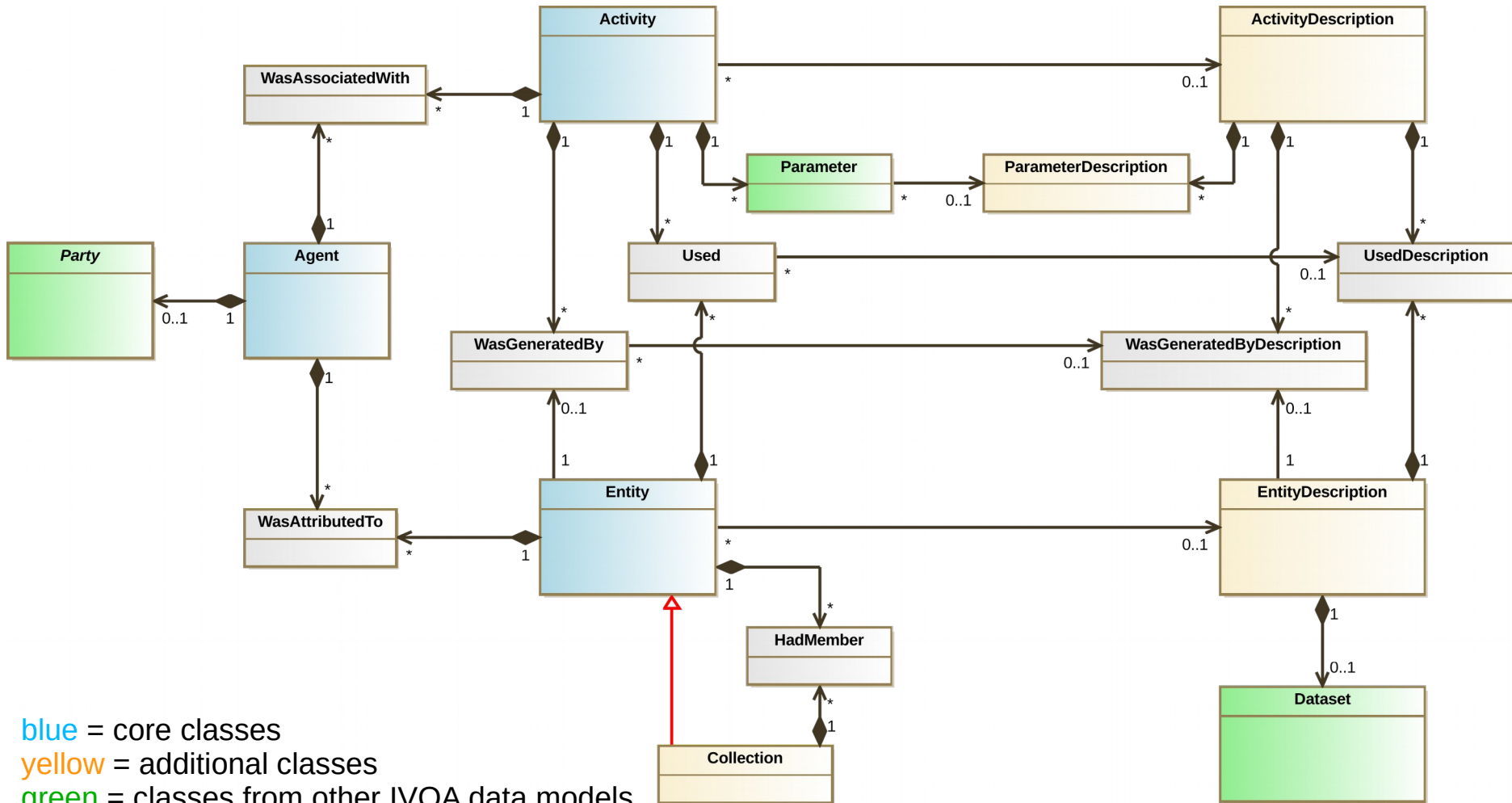
Extending the core

- in astronomy: know most common processes
- introduce new “description” classes for common processes and datatypes:



- connection to similar structures in other data models:
 - Activity => **Experiment** in Simulation Data Model
 - ActivityDescription => **Protocol** in Simulation Data Model
 - EntityDescription => **Dataset** in Dataset Data Model

Overview class diagram from working draft



blue = core classes
yellow = additional classes
green = classes from other IVOA data models
grey = relation classes

Status

- **Minimum requirements:**
 - collected and included in working draft at volute
 - “Provenance data model classes and attributes should be linked to other IVOA concepts when relevant”
 - “Provenance information should be serializable into the W3C PROV standard formats with minimum information loss.”
- **Reordered use cases (goals):**
 - new: “forward search”,
e.g. Give me all images derived from this raw image/produced using this version of the pipeline.
 - desired by pipeline developers

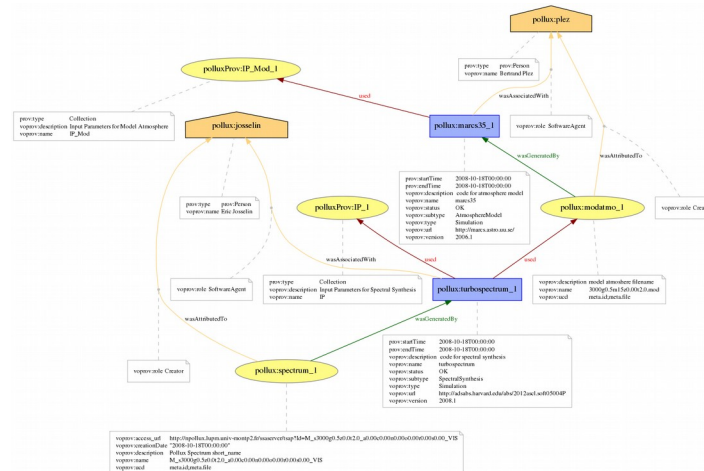
Status

- Use cases

- specific use cases:

- Pollux, RAVE, CTA (see Mathieu's talk)
- included in working draft
- short description, diagrams + link to prototype implementations

- new use case: HiPS (see François' talk)



Status

- **Provenance access**
 - ProvDAL: query provenance for entities based on their ID
 - ProvTAP: TAP service for provenance metadata
- **Serialization formats**
 - new section in working draft
 - 3 methods:
 - map to W3C classes, use W3C serialisations (PROV-N, PROV-JSON, PROV-XML)
 - TAP export into database table
 - VOTABLE: one table element for each class

Status

- VOTABLE serialization, example for activities:

```
<TABLE name="activities" utype="prov:activity">
  <FIELD name="name" utype="prov:activity.name" datatype="char" arraysize="*" />
  <FIELD name="start" utype="prov:startTime" datatype="char" arraysize="*" xtype="ISO8601" />
  <FIELD name="stop" utype="prov:endTime" datatype="char" arraysize="*" xtype="ISO8601" />
  <FIELD name="methodname" utype="voprov:method_name" datatype="char" arraysize="*" />
  <FIELD name="version" utype="voprov:method_version" datatype="char" arraysize="*" />
  <DATA>
  <TABLEDATA>
  <TR><TD>cta:telescope_stage_520</TD>
  <TD>2015-07-30T09:45:00</TD><TD>2015-07-30T10:00:00</TD>
  <TD>Telescope_stage</TD><TD>1.0</TD></TR>
  </TABLEDATA>
  </DATA>
</TABLE>
```

Status

- Mapping Dataset DM <--> ProvenanceDM

Dataset DM	Provenance DM	Comment
DataID.title	Entity.label	title of the dataset
DataID.collection	HadMember.collectionID	link to the collection to which the dataset belongs
DataID.creator	Agent.name	name of agent
DataID.creatorDID	AlternateOf.entityID	id for the dataset given by the creator
DataID.ObservationID	wasGeneratedBy.activityID	identifier to everything describing the observation; maybe it belongs to entity?
DataID.PublisherDID	Entity.ID	unique identifier for the dataset
Curation.PublisherID	Agent.ID	link to the publisher; role: publisher, type: organization/astronomer private collection)
Curation.Publisher	Agent.name	name of the publisher
Curation.Date	Entity.releaseDate	release date of the dataset
Curation.Version	Entity.version	version of the dataset
Curation.Rights	Entity.access	access rights to the dataset; one of [...]
Curation.Reference	Entity.link	link to publication
Curation.Contact	Agent.ID or name?	link to Agent with role contact
DataProductType	EntityDescription	subclass to EntityDescription
DataProductSubType	EntityDescription	subclass to EntityDescription
CalibLevel	EntityDescription	subclass to EntityDescription, calibration level

Next steps

- Finish UML description and working draft (until mid November)
- Produce VODML version of the data model
- Publish working draft on ivoa.net/documents page

- SimDM - ProvenanceDM mapping
- More prototyping:
 - VOTABLE serialization
 - ProvTAP and ProvDAL access
 - continue implementing Provenance for use cases

Weblinks

- Check the current status at:

IVOA Wiki page:

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

Volute SVN repository:

<http://volute.g-vo.org/svn/trunk/projects/dm/provenance/description/>

- Thanks!