IVOA Provenance Data Model
Current status

InterOp, 21\textsuperscript{th} 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

- identify data entities
- identify processes (activities)
- identify responsible people
Example in astronomy

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

- provenance is defined by the relations between data, activities and the people/projects involved
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/](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:

```xml
<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**

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