Prototyping Provenance metadata for the Virtual Observatory

Mireille Louys, CDS & ICube Laboratory
University of Strasbourg

M.Louys, Provenance in astronomy, RDA 6th plenary, sept 2015
International Virtual Observatory Alliance

M. Louys, Provenance in astronomy, RDA 6th plenary, sept 2015
Science driven

Information technology developments for astronomical science

Fosters interoperability

Standardisation process a la W3C

Working groups

Technical coordination

Science priority committee

Executive Board representing all national projects
Working Groups / Interactions

Data Model

Applications

Visualise, compare, compute

DataAccess Layer

Exchange, circulate

VOQL

Query Language for astronomical databases

Grid/Webservices

WS FrameWork
Grid Computing facilities

VOTable

Transport: XML

Semantics

Unified Content Descriptor, Vocabulary, Units

Registry

Localise resources and services
Assisting astronomers for data search

- What astronomers may look for?
  - Select data sets according to their science topic
    - Project, instrument, facility (telescope name and type)
    - Location (position in the sky or class of object)
    - Physical properties in space, time, spectral domain, flux
    - Types, Formats, Size

- Existing VO data models for metadata
  - Coarse description only on data Provenance
Provenance metadata in the IVOA

- Explains how data sets were produced
  - Observing process and conditions
  - Data reduction, selection and extraction methods applied to raw measures to build up science-ready data products
    (source lists, spectra, light curves, images, …)

- Helps VO users to:
  - Derive selection criteria to filter out suitable data for his/her scientific needs
  - Estimate better which data release fits the best for their needs
  - Run his/her own reduction method on intermediate data products in order to refine data analysis.

→ Expose progenitors of science data products
Provenance in the W3C

W3C Provenance definition

“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. PROV-DM is the conceptual data model that forms a basis for the W3C provenance (PROV) family of specifications.”

PROV-OVERVIEW (Note), an overview of the PROV family of documents
PROV-PRIMER (Note), a primer for the PROV data model
PROV-O (Recommendation), the PROV ontology, an OWL2 ontology allowing the mapping of the PROV data model to RDF
PROV-DM (Recommendation), the PROV data model for provenance (this document)
PROV-N (Recommendation), a notation for provenance aimed at human consumption
PROV-XML (Note), an XML schema for the PROV data model
PROV-AQ (Note), mechanisms for accessing and querying provenance
W3C Provenance pattern

- Makes explicit:
  - Processing steps
  - Chain of dependencies
  - Responsibilities

- Useful for all execution sequence of tasks, workflow, reduction pipeline, analysis workflow, etc.

- Applies for both acquisition and reduction steps

M.Louys, Provenance in astronomy, RDA 6th plenary, sept 2015
In our context

- **Entity**
  - data products (files), ancillary data (calibration, instrumental response, etc.), processing parameter files

- **Activity**
  - data acquisition, mosaicing, regridding, fusion, calibration, …, transformation

- **Agent**
  - Telescope astronomer, pipeline operator, principal investigator, etc.
Customized re-use
Binding to existing data models

Dataset Metadata Model
Work Package 4

- More use-cases to work out with this pattern
- Explore the ActivityDescription class for various use-cases
  - M. Servillat, C. Boisson, M. Sanguillon, J. Bregeon
    → CTA data products (4 levels of progenitors)
  - High energy physics
    → Fitting parametric models profiles for XMM spectra
  - Theoretical spectra
    → Provenance for the Pollux data base at LUPM
Need for a serialisation format

- Currently most of the provenance information in astronomy collections is available as:
  - log files
  - list of launched command lines in FITS headers in `COMMENT` keywords

- W3C offers various forms of syntax, translators
  - see Kristin Riebe astronomical use-case for the RAVE pipeline

- PROV-N (W3C)
  - Traces the execution scenario in simple text
  - Defines a grammar
Conclusion

- Emergence of massive projects (LSST, …)
  - The « code to data » strategy requires a precise and interoperable description of processing

- Appropriate time to consider Provenance metadata for astronomical data products

- Which levels of details according to use-cases
  - Science user → highlights for data quality
  - Pipeline /workflow management → reproducibility