The Case for Provenance

Juande Santander-Vela, Arancha Delgado
European Southern Observatory, Archive Department
Talk Outline

- What is Provenance?
- How does Provenance fit in the VO?
- The case(s) for Provenance
- Conclusions
What is Provenance?
What is Provenance?

provenance  |ˈprævənəns|
noun
the place of origin or earliest known history of something: an orange rug of Iranian provenance.
• the beginning of something's existence; something's origin: they try to understand the whole universe, its provenance and fate.
See note at origin.
• a record of ownership of a work of art or an antique, used as a guide to authenticity or quality: the manuscript has a distinguished provenance.

ORIGIN late 18th cent.: from French, from the verb provenir ‘come or stem from,’ from Latin provenire, from pro- ‘forth’ + venire ‘come.’
What is Provenance?

**provenance |ˈprævənəns|**

noun
the place of origin or earliest known history of something: an orange rug of Iranian provenance.
- the beginning of something's existence; something's origin: they try to understand the whole universe, its provenance and fate.
See note at origin.
- a record of ownership of a work of art or an antique, used as a guide to authenticity or quality: the manuscript has a distinguished provenance.

ORIGIN late 18th cent.: from French, from the verb provenir ‘come or stem from,’ from Latin provenire, from pro- ‘forth’ + venire ‘come.’
What is Provenance?

**provenance |ˈprævənsən|**

noun
the place of origin or earliest known history of something: *an orange rug of Iranian provenance.*
- the beginning of something's existence; something's origin: *they try to understand the whole universe, its provenance and fate.*
See note at *origin*.
- a record of ownership of a work of art or an antique, used as a guide to authenticity or quality: *the manuscript has a distinguished provenance.*

ORIGIN late 18th cent.: from French, from the verb *provenir* ‘come or stem from,’ from Latin *provenire*, from pro- ‘forth’ + venire ‘come.’
What is Provenance?

provenance  |ˈprɑːvənəns|
noun
the place of origin or earliest known history of something: an orange rug of Iranian provenance.
• the beginning of something's existence; something's origin: they try to understand the whole universe, its provenance and fate.
See note at origin.
• a record of ownership of a work of art or an antique, used as a guide to authenticity or quality: the manuscript has a distinguished provenance.
ORIGIN late 18th cent.: from French, from the verb provenir ‘come or stem from,’ from Latin provenire, from pro- ‘forth’ + venire ‘come.’
What is Provenance?

- Definition points out to:
  - **History** of astronomical data products
  - **Ownership/Attribution** (Observer, Proposal, Telescope/Instrument)...
  - **Quality**
What is Provenance?
What is Provenance?

TYPICAL ASTRONOMICAL WORKFLOW

1. List of targets
2. Find existing data per target and band
3. Filter out data regarding quality parameters
4. Manipulate retrieved datasets

Data provenance information needed
What is Provenance?

TYPICAL ASTRONOMICAL WORKFLOW

- List of targets
- Find existing data per target and band
- Filter out data regarding quality parameters
- Manipulate retrieved datasets

CHARACTERISATION

Data provenance information needed
What is Provenance?

TYPICAL ASTRONOMICAL WORKFLOW

- List of targets
- Find existing data per target and band
- Filter out data regarding quality parameters
- Manipulate retrieved datasets

CHARACTERISATION

Data provenance information needed
What is Provenance?

Astronomical Source

Galactic & Extragalactic extinction

Atmosphere (Seeing & Opacity)

Telescope, Filters, Detectors…

Data Reduction Software

raw files

uses knowledge of

F1
F2
Fn

P1
P2
Pm

photon emission
data generation

Astronomical Source
What is Provenance?

- Astronomical Source
- Galactic & Extragalactic extinction
- Atmosphere (Seeing & Opacity)
- Data Reduction Software
- Photon emission
- Data generation
- Telescope, Filters, Detectors...

raw files: $F_1, F_2, ..., F_n$

data products: $P_1, P_2, ..., P_m$

Uses knowledge of
What is Provenance?

Astronomical Source

Galactic & Extragalactic extinction

Atmosphere (Seeing & Opacity)

Telescope, Filters, Detectors...

Data Reduction Software

raw files

F1

F2

F_n

uses knowledge of

NOT ONLY FITS FILES

data products

P1

P2

P_m

photon emission

data generation
What is Provenance?

- Astronomical Source
- Galactic & Extragalactic extinction
- Atmosphere (Seeing & Opacity)
- Telescope, Filters, Detectors…

Data Reduction Software

Raw files

F_1, F_2, ..., F_n

Data products

P_1, P_2, ..., P_m

NOT ONLY FITS FILES

ALSO CATALOGUES

Uses knowledge of

- photon emission
- data generation

FITS FILES

CATALOGUES
What is Provenance?

Project Metadata

Observation

Provenance

Processing

Software

Calibration

Ambient

Instrumental

Observation

Provenance

INTERNAL
PROVENANCE
WHITEPAPER
BY N. DELMOTTE
What is Provenance?

ORGANISATIONAL

- Project Metadata
- Observation
- Provenance
  - Processing
    - Software
    - Calibration
  - Ambient
  - Instrumental

INTERNAL PROVENANCE WHITEPAPER BY N. DELMOTTE
What is Provenance?

ORGANISATIONAL

- Project Metadata

“ACQUISITIONAL”

- Observation
- Provenance
- Processing
- Ambient
- Instrumental
- Software
- Calibration

- Observation
- Provenance

INTERNAL PROVENANCE WHITEPAPER BY N. DELMOTTE
How does Provenance fit in the VO?
How does Provenance fit in the VO?

Astronomical Source

Galactic & Extragalactic extinction

Atmosphere (Seeing & Opacity)

Telescope, Filters, Detectors...

Data Reduction Software

raw files

F_1

F_2

F_n

Uses knowledge of

P_1

P_2

P_m

Photon emission

Data generation
How does Provenance fit in the VO?

Astronomical Source

Galactic & Extragalactic extinction

Atmosphere (Seeing & Opacity)

Telescope, Filters, Detectors...

raw files

Data Reduction Software

data products

photons emission data generation

uses knowledge of

F₁ F₂ Fₙ

P₁ P₂ Pₘ
How does Provenance fit in the VO?
How does Provenance fit in the VO?

RADAMS

Data Model for Observation

Observation

Characterization

Provenance

Target

Packaging

ObsData

Coverage

Resolution

Precision

Sensitivity

Accuracy

Data Model for Dataset Characterisation

RADAMS

Data Model for Observation
How does Provenance fit in the VO?

RADAMS
Ph.D. Thesis

Data Model for Observation

Observation

Provenance

Curation

Policy

Target

Packaging

Characterization

Coverage

Resolution

Precision

Sensitivity

Accuracy

Data Model for Dataset Characterisation

Data Model for Observation
How does Provenance fit in the VO?

RADAMS Ph.D. Thesis

Data Model for Observation

Observation

Curation

Policy

Packaging

Extended Provenance

Provenance

Target

Data Model for Dataset Characterisation

Coverage

Resolution

Precision

Sensitivity

Accuracy

Data Model for Observation

Characterization

ObsData
How does Provenance fit in the VO?

DATA MODEL FOR OBSERVATION

- Observation
- Curation
- Policy
- Target
- Packaging

DATA MODEL FOR DATASET CHARACTERISATION

- Coverage
- Resolution
- Precision
- Sensitivity
- Accuracy

DATA MODEL FOR OBSERVATION

- ObsData
- Characterization

SEE TALK BY FRANÇOIS BONNAREL

RADAMS Ph.D. Thesis
The case(s) for Provenance
The case(s) for Provenance

- Use cases for Provenance, centred around the core concepts:
  - **History** of a data set (**documentation**)
  - Establishing **ownership**/providing **attribution**
  - **Quality** assessment
History of a data set (ESO)
History of a data set (ESO)
History of a data set (ESO)

Compilation by A. Delgado
History of a data set (ESO)

COMPILATION BY
A. DELGADO
History of a data set (ESO)

COMPILATION BY
A. DELGADO
## History of a data set (ESO)

### Telescope Information
- **Site**
- **Dates: First Light, Commissioning…**
- **Instruments & Optical Elements**

### Instruments
- **Technologies**
- **Configurations**

### Optical Elements
- **Installation dates**
- **Kind (Filter, Grisms, Gratings…)**
- **Names ↔ IDs ↔ Transmission Curves**

**Compilation by A. Delgado**
History of a data set

This set is defined for each instrument band AND receiver configuration.
History of a data set

INSPIRED BY LAMB & POWER’S RAW RADIO DM NOTE

This set is defined for each instrument band AND receiver configuration
History of a data set

- **Observation**
- **Provenance**
  - `Processing.timeStamp`
- **AmbientConditions**
- **InstrumentConf**
  - `Processing.timeStamp`
- **ProcessingStep**
  - `Processing.timeStamp`
  - `kind`
  - `softwarePackage`
  - `parameter.name`
  - `parameter.kind`
  - `parameter.value`
- **Calibration**
  - `Processing.timeStamp`
  - `parameter.name`
  - `parameter.kind`
  - `parameter.value`
  - `parameter.sigma`
  - `parameter.calCoeff[n]`
History of a data set

- Observation
- Provenance
- InstrumentConf

AmbientConditions
- timeStamp
- opacity
- temperature
- humidity
- wind

OpacityCurve
+ AmbientConditions.timeStamp
- opacity
- skydipStart
- azimuth
- elevation.[n]
- tsky.[n]
- atmosphericModel
History of a data set

- **Observation**
  - AmbientConditions
    - timestamp
    - opacity
    - temperature
    - humidity
    - wind

- **Provenance**
  - OpacityCurve
    + AmbientConditions.timestamp
    - opacity
    - skydipStart
    - azimuth
    - elevation
    - tsky
    - atmosphericModel

- **InstrumentConf**
History of a data set

This set is defined for each instrument band AND receiver configuration.
History of a data set

UNIFICATION OF PROVENANCE AT THE AMBIENT, TELESCOPE_CONF LEVELS:
THE REST IS VERY HETEROGENEOUS
History of a data set

• Can only be done with:
  • Historical archive of:
    • optical elements
    • software (possibly with virtualization)
    • configuration values / fudge factors
    • ambient information
  • DIMM seeing, opacity, conductivity…
History of a data set

- Can only be done with:
  - Historical archive of:
    - optical elements
    - software (possibly with virtualization)
    - configuration values / fudge factors
    - ambient information
  - DIMM seeing, opacity, conductivity…
  + SYSTEMATIC OBSERVING LOGS!
Ownership Tracking

Level 0
- Period
  - Program Type
    - Program ID
      - Run ID

Level 1
- Run ID
  - Night
    - Night Log
  - OB ID
  - Raw File (DP_ID)
    - DPR CATG

Level 2
- Run ID
  - Night
  - Reduced Science Data
  - Master Calib. Data
  - Ancillary
  - File

Level 3
- Data Release ID
  - Structured ReadMe
  - Association IDs
  - Main File(s)
  - Associated File(s)
    - DP_ID
  - DP_ID

Level 4
- RunID
  - bibcode
    - ProgID
    - DataID
Ownership Tracking

Level 0
- Period
- Program Type
- Program ID

Level 1
- Run ID
- Night
- OB ID
- Raw File (DP_ID)
- DPR CATG

Level 2
- Run ID
- Night Log
- Reduced Science Data
- Master Calib. Data
- Ancillary
- File

Level 3
- Data Release ID
- Structured ReadMe
- Association IDs
- Main File(s)
- Associated File(s)
- DP_ID

Level 4
- bibcode
- ProgID
- DataID
- RunID

ProjID
Ownership Tracking
Ownership Tracking

- Can only rely on unique identifiers being maintained
- Other VO manipulations (SAMP messages, cross-matching) can lose associations
- Need to provide services for identification of key IDs
  - Increase of the role of the IVOA Registry?
- This is stopping some small publishers!
Ownership Tracking

- Can only rely on unique identifiers being maintained
- Other VO manipulations (SAMP messages, cross-matching) can lose associations
- Need to provide services for identification of key IDs
  - Increase of the role of the IVOA Registry?
- This is stopping some small publishers!
Quality Assessment
Quality Assessment

- Linked to:
  - Acquisition **configuration**
  - **Actual problems** with instruments/telescope
  - **Weather**
  - **Intended usage** of the dataset(s)!
Quality Assessment

- Linked to:
  - Acquisition configuration
  - Actual problems with instruments/telescope
  - Weather
  - Intended usage of the dataset(s)!

NO OBJECTIVE, ONE-SIZE-FITS-ALL, QUALITY ASSESSMENT METRIC
Quality Assessment

- Linked to:
  - Acquisition \textit{configuration}
  - \textbf{Actual problems} with instruments/telescope
  - \textbf{Weather}
  - \textbf{Intended usage} of the dataset(s)!
Quality Assessment

- Linked to:
  - Acquisition configuration
  - Actual problems with instruments/telescope
  - Weather
  - Intended usage of the dataset(s)!

TIME CHANGING
Queriable Provenance?

- Most queries, on **Characterisation/Target/Curation**
- **Instrument-specific forms** for instrument configurations
  - Allow UType/UCD/IVOAT **key-value pairs**?
- Instrument-specific *a priori* quality assessment
- *A posteriori*, usage-specific quality assessment, offline > usage-specific query to data Provenance
Conclusions
Conclusions

- Provenance is an integral part of the Observation DM
- Provenance comes with discipline, but allows for quality science
- Different approach for different kinds of instruments, but all under the same general framework
- Provenance should be accessible for any item; specialised TAP version?
Conclusions

- Data centres: **consistent naming/coding**, plus mappings for existing data

- **History, history, history!**

- Should we **forget** about **past data**, and focus in the **future**?
Vielen Dank!

STARTING WITH MY GERMAN CLASSES ;-)}