Data Provenance: Use Cases for the ESO archive, and interactions with the Virtual Observatory

J. Santander-Vela[1,2], A. Delgado[1], N. Delmotte[1], M. Vuong[1]

[1] European Southern Observatory [2] Instituto de Astrofísica de Andalucía (CSIC)

J. Santander-Vela

In the Virtual Observatory era, where we intend to expose scientists, or software agents on their behalf, to a stream of observations from all existing facilities, the ability to access and to further interpret the origin, relationships, and processing steps on archived astronomical assets (their Provenance) is a requirement for proper observation selection, and quality assessment. In this poster we present the different use cases Data Provenance is needed for, the challenges inherent to the ESO archive and their link with ongoing work in the IVOA.

Data Provenance and Astrophysics

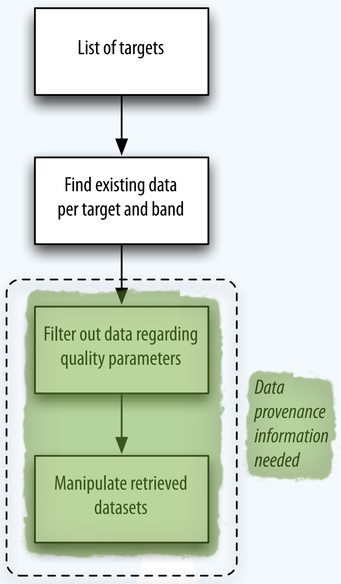

Figure 1 Astronomical, multi-wavelength analysis work-flow using e-Science tools such as the **VO**, showing where data Provenance information is a requirement for e-science.

Data Provenance in e-Science is the *tracking of the origin of any piece of information that has been recorded, with knowledge of all processing steps undergone.* In the astronomical realm, we track data Provenance in order to answer questions like *what was done to the photons captured by this archived entity, and who were responsible for it?* And we need to be able to answer that question because for typical astronomical workflows (such as that shown in Figure 1) across large datasets Provenance must be obtained, and used throughout the workflow, in order to get meaningful results.

For the typical detection and processing steps in astronomy, a certain parameter (i.e., energy flux) is measured and registered in files by a telescope's detectors, which are related to photon emission from an astronomical source (See Figure 2). Concrete questions for data Provenance are **which were the files used to create a particular data product (Associations)**, and **which are the data products created from it? (Inverse Association)**. Additionally, we need to know **are we using the same source data *(photons)* for any two given data products? (Photon Accountancy)**, as the use of the same photons in different data products which are later combined can produce unreliable science. Plus, we need to be able to **track parameters related to the models for emission, absorption and detection** used by the reduction software.

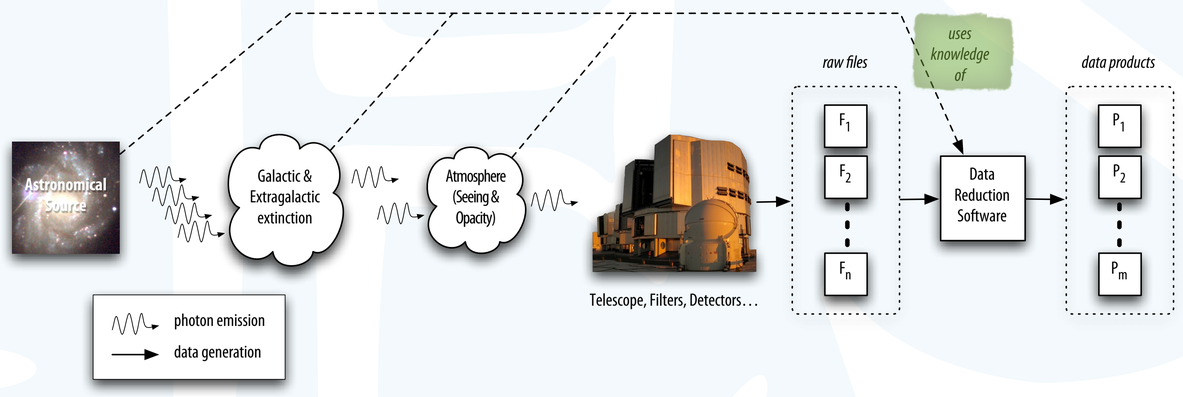

Figure 2 Elements of astronomical data Provenance: data reduction pipelines, or standalone software are used to convert some property measured with the telescope detectors into up to *n* files (F_n), and later into *m* physical parameters and data products (P_m). For that, they use knowledge of the telescope and detectors settings (for details on the light-path, for instance, see [Delgado 2009] in this poster session), plus modelling of the telescope, source, and the different absorbing media photons have traversed.

Requirements for Data Provenance Systems: High-Level Use Cases

From a general classification of data Provenance systems' properties, such as that shown in Figure 3, we can see that for some systems Provenance is considered simply informational. Ideally, however, Provenance can be used to *attribute* dataset originators, *replicate* derivations of parameters, and ultimately assess *dataset quality* (both being retrieved, and being archived). If the Provenance collection starts with observatory systems, *audit trails* can be created to see which people and systems have been responsible for each particular processing step from the origin.

Data Provenance can be data or process oriented (but data oriented Provenance is called, in the IVOA context, **Characterisation**), or process oriented, so the metadata for Provenance outside of Characterisation must be process oriented.

Figure 4, on the other hand, shows a classification of the different use cases that have been compiled as desireable for the ESO archive. Any implementation of astronomical data Provenance within ESO must take into account those use cases.

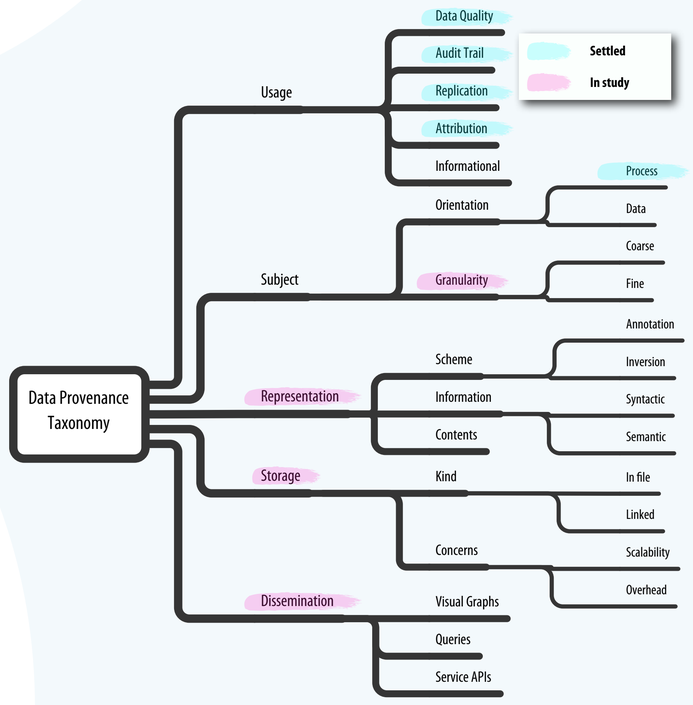

Figure 3: Taxonomy (based on [Simmhan 2005]) of the different aspects to be taken into account for any kind of e-science Provenance system.

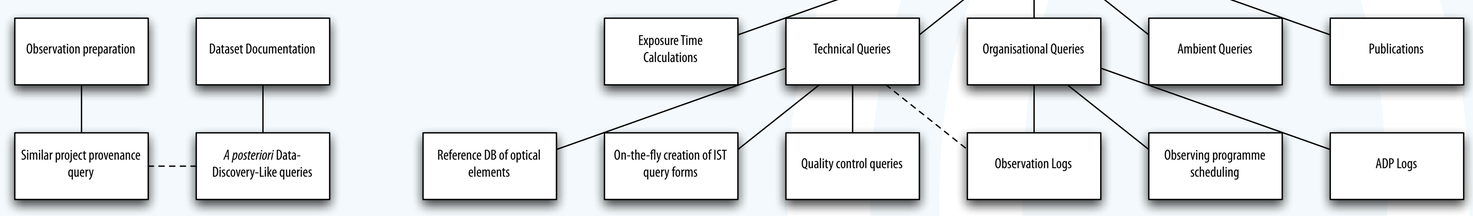

Figure 4 Classification of use cases compiled for the ESO archive. The main group is Data Discovery using Provenance information (mostly through inverse association queries), while similar queries are desireable for dataset documentation (using direct association queries).

Glossary

Characterisation Observation metadata within a multi-axis space (usually space, time, frequency and observed property).

Curation Project-related metadata, such as what kind of project does an observation belong to, but also encompassing data collections which have been but together *a posteriori*.

e-Science Enhanced science performed through the use of networked, distributed resources. See **VO**.

IVOA International Virtual Observatory Alliance, organisation in charge of creating the standards for the description and interoperation of astronomical observations in archives. http://www.ivoa.net

Provenance Observation metadata related to the origin of every realisation of an observation (i.e., files or data products), from instrumental configuration, to ambient conditions, and processing steps.

VO Virtual Observatory, a federation of data archives sharing the same description, including Curation, Characterisation and Provenance metadata, plus common data access protocols.

Implementing Data Provenance at ESO: Challenges and Jump-Starts

The remaining elements to be set up for the establishment of a proper data Provenance, as shown on Figure 3, are **Granularity**, **Storage**, **Representation**, and **Dissemination**.

- **Granularity** must be chosen to serve all use cases above, and also imposes restrictions on **Storage** (to be able to cope with the chosen Granularity; Storage is an implementation detail not relevant to consumers.)
- **Representation** is the way Provenance is expressed when shared with requesting entities (as opposed to Storage). Hierarchical trees, series of keyword/value pairs, or even plain-text files for informational use cases. However, the rich nature of the objects described by Provenance (for instance, see [Delgado 2009] for a detailed description of the light-path), Representation could be XML-based, for easier conversion to other **IVOA** formats.
- **Dissemination**, finally, is the way Provenance information is to be available for shared datasets. Decisions in this respect include creating Provenance services to satisfy the use cases in Figure 4.

Provenance of astrophysical datasets is being discussed within the IVOA Working Group for Data Modelling (DM WG). A **Representation, in the form of a complete IVOA Observation data model**, has already been proposed in [Santander 2009] (see Figures 5 and 6), and is currently under study.

There is also a **complete FITS keyword/metadata database** already in place at ESO [Vuong 2008], which will be one of the basis for establishing Storage and Granularity of the future ESO Provenance services. Currently, that facility stores more than **five thousand million keyword-value pairs**, for more than **ten million raw file entries**, and already includes metadata versioning support (a requirement for Provenance audit trails and historical dataset documentation).

Conclusion

By identifying the elements needed for a successful data Provenance management system, and the selection of already in place systems, the ESO archive will be able to provide even more advanced scientific-oriented data products and services, while at the same time benefiting from existing IVOA standards for the Virtual Observatory, and helping in their development, leading by example.

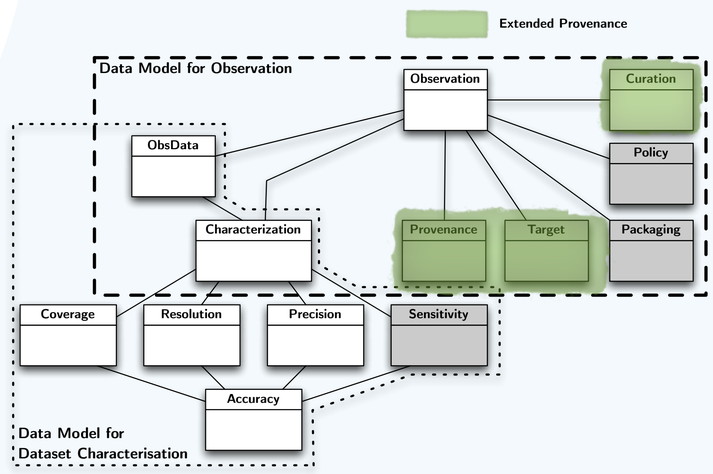

Figure 5 Proposal [Santander 2009] for an IVOA Observation Data Model. The green highlight shows the concept of *extended* Provenance, which adds spatial (Target) and project-oriented (Curation) Provenance to the Instrumental, Ambient and Processing Provenance, shown in Figure 6 .

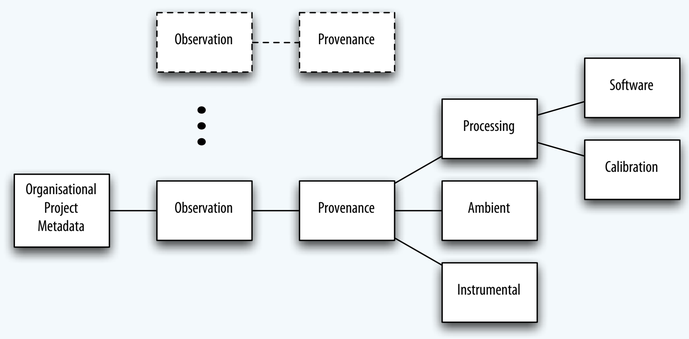

Figure 6 Detailed hierarchical Representation of the elements of data Provenance, with emphasis in non-organisational Provenance.

References

[**Delgado 2009**] Delgado, Delmotte, & Vuong, "Instrumental Provenance of ESO Archival Data", *Proceedings of the ADASS XIX*, ASP Series (in preparation)

[**Santander 2009**] Santander-Vela, "Integration of radio astronomical archives and tools in the Virtual Observatory framework," *Ph.D. Thesis*, University of Granada, May 2009

[**Simmhan 2005**] Simmhan, Plale, & Gannon, "A Survey of Data Provenance in e-Science," *SIGMOD Record* 34, 3, Sept. 2005

[**Vuong 2008**] Vuong et al., "Applications of the ESO metadata database", *Proceedings of the SPIE*, vol. 7016, pp. 70161M-70161M-8, July 2008