Discovery and Provenance Metadata for Persistent Data Objects in the Virtual Observatory

Arnold Rots
2012-05-13

When data objects are published, proper, accurate, and complete metadata should be attached for three reasons:

- Curation of the data objects for preservation
- Allowing data discovery through detailed queries
- Facilitating the tagging of provenance information when these objects are passed along in the literature, modified, enriched, or further analyzed

The VO currently has a mechanism for curation and discovery of data objects contained in repositories that have a fair degree of homogeneity (for instance, observatory and mission archives): the resource metadata recorded in VO registries. This allows data discovery tools, given a particular user query, to make a reliable selection of repositories to be queried further.

However, with the advent of data publishing tools that offer researchers the opportunity to deposit valuable data objects into a persistent repository where these objects will be available to the community in perpetuity, we have come to realize that another type of registry is needed and that a modified set of metadata is required to make that availability a reality. Since the repositories will contain very heterogeneous collections of data objects, the resource metadata for the repository will be little help in data discovery and the discovery tools need direct access to individual objects' metadata. In addition, provenance metadata become crucial, since many objects will contain highly analyzed data derived from one or more parent objects.

We suggest that the list of metadata enumerated in this note be used to tag stored data objects and that all, most, or at least relevant items be attached to citations of the data objects in the literature. The purpose is to enable proper citation, credit, and accountability of the data objects. Persistent dataset identifiers play a central role in recording the provenance. The ADS and the AAS journals have implemented a system of such identifier that is in full working order, but a more comprehensive agreement is needed; this is a project that is being pursued under the aegis of the DC&P IG.

This list of metadata keywords is directly derived from the IVOA recommended Resource Metadata for the Virtual Observatory – the registry metadata. Therefore any data cited from repositories in an IVOA registry can easily be provided with metadata from the registry. A limited number of keywords are considered essential for a basic understanding of the data objects, and are thus denoted as ***required***. Others are highly recommended if ***relevant***. All others are optional, or may be applied to certain classes of data objects only. The term "*data objects"* refers to data in any form and at any granularity. It may refer to a tar file, an image, a spectrum, a light curve, a table, the data behind a graph, or even a single number (isolated or in a table); what it does require is that there is a unique identifier that allows the user to find it.

1 Identity metadata

Title (string) [Dublin Core] [Required]
Definition: A name given to the data object.

ShortName (string)
Definition: A short abbreviation for the name given to the data object.

Identifier (URI) [Dublin Core] [Required]
Definition: An unambiguous persistent reference to the data object within a given context. The syntax for Identifiers is described in *IVOA Identifiers* in the IVOA document collection (http://www.ivoa.net/Documents/).

2 Curation metadata

Publisher (string) [Dublin Core] [Required]
Definition: An entity responsible for making the data object available

PublisherID (URI)
Definition: The identifier for the entity responsible for making the data object available. The syntax for Identifiers is described in *IVOA Identifiers* in the IVOA document collection (http://www.ivoa.net/Documents/).

Creator (string) [Dublin Core]
Definition: An entity primarily responsible for making the content of the data object.

> *Creator.Logo* (URL)
> Definition: A URL pointing to a graphical logo, which may be used to help identify the information resource.

Contributor (string) [Dublin Core]
Definition: An entity responsible for making contributions to the content of the data object.

Date (string) [Dublin Core] [Required]
Definition: A date associated with an event in the life cycle of the data object. Typically, Date will be associated with the creation or availability (i.e., most recent release or version) of the data object. ISO8601 is the preferred format (YYYY-MM-DD).

Version (string)
Definition: A label associated with the creation or availability (i.e., most recent release or version) of the data object.

Contact (string, e-mail address)
Definition: The e-mail address for contacting the persons responsible for the data object.

> *Contact.Name* (string)
> Definition: The name of the contact.
>
> *Contact.Address* (string)
> Definition: The mailing address of the contact.

Contact.Email (e-mail address)
Definition: The e-mail address of the contact.

Contact.Telephone (string)
Definition: The telephone number of the contact.

3 General content metadata

Subject (string, list) [Dublin Core]
Definition: A list of the topics, object types, or other descriptive keywords about the data object.

Description (string, free text) [Dublin Core]
Definition: An account of the content of the data object.

Source (string) [Dublin Core]
Definition: A bibliographic reference from which the present data object is derived or extracted.

ReferenceURL (URL)
Definition: A URL pointing to additional information about the data object. In general, this information should be human-readable.

Type (string, list) [Dublin Core] [Required]
Definition: The nature or genre of the content of the data object.
Comment: Type includes terms describing general categories, functions, genres, or aggregation levels for content. VO Types include:

Type	Description
Observation	Collection of data objects (files) associated with one or more observations
Object	Collection of data objects (files) associated with one or more celestial objects
Image	One or more 2-D images
Mosaic	Mosaic of multiple 2-D images
Cube	One or more 3-D data cubes
Spectrum	One or more 1-D spectra
LightCurve	One or more 1-D light curves
EventList	One or more event lists
Catalog	Collection of derived data, primarily in tabular form
Table	Table of values; at least two columns
Value	Single value
ValuePair	Keyword-value pair
Library	Collection of published materials (journals, books, etc.)
Simulation	Theoretical simulation or model
Survey	Collection of observations covering substantial and contiguous areas of the sky
Animation	Animation clips of astronomical phenomena
Artwork	Artists' renderings of astronomical phenomena or objects
Facsimile	Digitized facsimile of (historical) document
Historical	Historical information about astronomical objects.
Other	A data object not described by any of the above types.

This list is extensible. Resources providing more than one type of content should list all relevant types.

ContentLevel (string, list)
Definition: A description of the content level, or intended audience.

Relationship (string) [Required]
Definition: A data object may be related to another data object in a way that is important to document, so that associated services or duplicate copies may easily be located. The highlighted values are recommended.

primary	The data object is the original copy, published by the creator
mirror-of	The data object is a mirror of another data object. Information gathered from the data objects is indistinguishable.
service-for	The data object is a service associated with a data collection.
derived-from	The data object is a derivative of another data object, e.g., a subset selected for a particular scientific interest, or a reprocessed data collection.
copy-of	The data object is a copy of an object that was obtained from another repository.
served-by	The data object can be accessed via another service resource.

RelationshipID (URI)
Definition: The identifier of an associated data object. The relationship is described in the Relationship metadata element. The syntax for Identifiers is described in *IVOA Identifiers* in the IVOA document collection (http://www.ivoa.net/Documents/).

4 Collection and service content metadata

Facility (string, list)
Definition: The observatory or facility where the data was obtained.

Instrument (string, list)
Definition: The instrument used to collect the data.

Coverage (string) [Dublin Core, with modifications]
Definition: The extent of scope of the content of the data object.

Coverage.Spatial (string)
Definition: The sky coverage of the data object.

Coverage.Spectral (string, list)
Definition: The spectral coverage of the data object.

Coverage.Spectral.Bandpass (string, list)
Definition: A specific bandpass specification.

Coverage.Spectral.CentralWavelength (float)
Definition: The central wavelength of the spectral bandpass, in meters.

Coverage.Spectral.MinimumWavelength (float)
Definition: The minimum wavelength of the spectral bandpass, in meters.

Coverage.Spectral.MaximumWavelength (float)
Definition: The maximum wavelength of the spectral bandpass, in meters.

Coverage.Temporal.StartTime (string)
Definition: The earliest temporal coverage of the data object.

Coverage.Temporal.StopTime (string)
Definition: The latest temporal coverage of the data object.

Coverage.Depth (float)
Definition: The (typical) depth coverage, or sensitivity, of the data object. Coverage.Depth is specified in flux density (Jy).

Coverage.ObjectDensity (float)
Definition: The (typical) density of objects, catalog entries, telescope pointings, etc., on the sky, in number per square degree.

Coverage.ObjectCount (int)
Definition: The total number of objects, catalog entries, etc., in the data object.

Coverage.SkyFraction (float)
Definition: The fraction of the sky represented in the observations, ranging from 0 to 1.

Resolution (float)
Definition: The resolution of the data object contents.

Resolution.Spatial (float)
Definition: The spatial (angular) resolution that is typical of the observations, in decimal degrees.

Resolution.Spectral (float)
Definition: The spectral resolution that is typical of the observations, given as the ratio $\lambda/\Delta\lambda$ (so that higher spectral resolution has a larger number).

Resolution.Temporal (float)
Definition: The temporal resolution that is typical of the observations, given in seconds.

UCD (string, list)
Definition: A list of the UCDs (Unified Content Descriptors, http://cdsweb.u-strasbg.fr/doc/UCD.htx) represented in the data object.

Format (string, list) [Dublin Core] [Required?]
Definition: The physical or digital manifestation of the information provided by the data object: fits, jpg, ascii, csv, tsv, tar, etc.

Rights (string) [Dublin Core]
Definition: Information about rights held in and over the data object.

5 Data and metadata quality assessment

Users of virtual observatory data objects need some way to assess the quality of the data and of the associated descriptive information in the registry. Data quality is both subjective and quantitative. While the completeness and consistency of the provenance metadata itself may be a reasonable indicator of the quality of the associated data object, this is at best a qualitative measure. The following metadata elements are intended to capture the most basic measures of data quality, and may well require extensions as VO usage practices evolve and become more sophisticated.

DataQuality (char)
Definition: An overall assessment of the integrity, consistency, and level of documentation concerning uncertainty estimates and calibration procedures, of the data object provided. We suggest 3 general grade levels, plus codes for unknown or undocumented cases:

- A Data are fully calibrated, fully documented, and suitable for professional research.
- B Data are calibrated and documented, but calibration quality is inconsistent. Users are advised to check data carefully and recalibrate.
- C Data are uncalibrated.
- U Data quality is unknown. If a data object does not provide a data quality assessment, class U should be assumed.

ResourceValidationLevel (int)
Definition: A numeric grade describing the quality of the data object description and interface, when applicable, to be used to indicate the confidence an end-user can put in the data object as part of a VO application or research study. The allowed values are:

0 The data object has a description that is stored in a registry. This level does not imply a compliant description.

1 In addition to meeting the level 0 definition, the data object description conforms syntactically to this standard and to the encoding scheme used.

2 In addition to meeting the level 1 definition, the data object description refers to an existing data object that has been demonstrated to be functionally compliant.

3 In addition to meeting the level 2 definition, the data object description has been inspected by a human and judged to comply semantically to this standard as well as meeting any additional minimum quality criteria (e.g., providing values for important but non-required metadata) set by the human inspector (see comment below).

4 In addition to meeting the level 3 definition, the data object description meets additional quality criteria set by the human inspector and is therefore considered an excellent description of the data object.

If no value is provided, level 0 should be assumed.

ResourceValidatedBy (URI)

Definition: The IVOA identifier for the registry or organisation that set the value of ResourceValidationLevel.

Uncertainty.Photometric (float)
Definition: The uncertainty of the photometric measurements provided by the data object, given in Jy.

Uncertainty.Spatial (float)
Definition: The uncertainty of the astrometric, or positional measurements, provided by the data object, given in degrees.

Uncertainty.Spectral (float)
Definition: The uncertainty of the wavelengths provided by the data object, given in meters.

Uncertainty.Temporal (float)
Definition: The uncertainty of the temporal measurements provided by the data object, given in seconds.