Data Model for Astronomical DataSet Characterisation - compliance

Addenda to Version 0.93

Working Group:
`http://www.ivoa.net/twiki/bin/view/IVOA/IvoaDataModel`
This version: Summary and suggestions for mandatory keywords
Editors:
Jonathan McDowell, François Bonnarel, Igor Chilingarian, Mireille Louys, Alberto Micol, Anita Richards.
Authors:
IVOA Data Model Working Group

Abstract

We present a summary of the Characterisation Data model and suggestions for mandatory keywords. The keywords MUST, SHOULD and MAY are used as in SIAP (**insert reference). An implimentation is compliant if it satisfies all the MUST and SHOULD requirements. An implimentation which satisfies all the MUST but not all the SHOULD requirements is partially compliant.

The prime goal is to get this model applied by data providers in useful ways. We should make it as easy as possible to describe any kind of observed or simulated data by minimising the number of compulsory fields. At the same time we must encourage data providers to give enough information to expand the ways in which data can be selected or manipulated by VO tools currently or imminently available.

Status of this document

This is a Draft. It should not be referenced or otherwise interpreted as a standard specification. ...

Acknowledgements

Members of the IVOA Data Model Working Group, including representatives of the US NVO, Astrogrid, and the AVO have contributed to the present draft. The editors particularly acknowledge the input from P. Didelon, G. Lemson, A. Rots, L. Shaw, B. Thomas and D. Tody. F. Bonnarel and M. Louys acknowledge support from the French ACI-GRID project, *IDHA* and the EU-funded *VOTech* project.

Contents

1 Introduction

The prime goal is to get the Characterisation data model applied by data providers in useful ways. We should make it as easy as possible to describe any kind of observed or simulated data by minimising the number of compulsory fields. At the same time we must encourage data providers to give enough information to expand the ways in which data can be selected or manipulated by VO tools currently or imminently available.

Characterization should be consistent with the Registry and with DAL-oriented models e.g. SIAP, Spectral DM, STAP (time) in areas of overlap. These models allow a query to return (lists of) data which does not explicitly exclude a region of interest. One of the goals of Characterization is to allow more refined selections without disabling very simple queries e.g. related to VOEvent.

Characterisation was not designed to describe catalogues but has been used successfully to describe e.g. an object list extracted from an image

or an observing log. We should not extend Characterisation to provide for issues exclusive to catalogues e.g. definitions of sources ('star', 'galaxy'...) which will be covered by the Source model.

The Observation model as a whole will indicate associated data sets and may also provide properties common to all the data from a particular instrument but at present Characterisation can only draw on other mature models such as STC and Registry. We now describe some considerations in using Characterisation which may help in deciding how to apply it to a given data set.

1.1 Limitations in this version

Characterization can contain algorithms (e.g. the variation of noise with position) or URLs (e.g. the location of a weight map). However, we are not aware (*??*) of any tools which can make use of these and the best that is usually useful is to provide the typical value and the minimum and maximum of a range.

It is not at present possible to impliment rules linking coverage on different axes. For example, if a survey consists of spatially distinct fields, observed in several wavebands, but not all fields contain all wavebands, then each field and/or each waveband must be described separately. Similarly, separate descriptions are required if resolution, noise or etc. behave differently in different areas of Support.

1.2 Implimenting Characterisation

1.2.1 Data Providers

It is envisaged that initially we will provide XML templates for manual editing. We will investigate what would be more convenient for large data collections depending on how they store their existing metadata. For example, manual completion of forms like
`http://nvo.stsci.edu/VORegistry/UpdateRegistry.aspx` would be suitable for coarse-grained regstry entries where one person only has a few collections to register, but not for providing Characterisation descriptions for thousands of separate observations in a telescope archive. Metadata required by Characterisation might be extracted from a number of sources such as

- An archive database;
- An observing log or other description which might be stored in a database or as ascii, xml or other documents;
- FITS headers

by more or less direct routes:

- Unambiguous identification between e.g. a database column or FITS keyword and a Characterisation element;
- Correspondance with formulaic modification, e.g. adding explicit units or calculating the field of view of an interferometer;
- A separate information source e.g. resolution of the telescope using different frequencies/configurations;
- Offline/human memory/judgement

We should consider some common cases and encourage the development of tools for automatic metadata extraction and codifying heuristics, as global or more domain-specific utilities (e.g. MEx ?).

1.2.2 Tools using Characterisation

TBC

2 Requirements for compliance

2.1 General considerations

The first three levels on each axis (Location, Bounds, Support), where used, MUST be given explicit numerical values (or arrays of values) in order to be accessible to any tool. Other elements MAY be given numerical values, or functions, or indirect references (URIs) but these are in general not used at present. Users are strongly encouraged to evaluate coarser levels of description explicitly even if they also provide finer levels.

We need to decide what users do if they are not giving a value for an element e.g. leave blank, consistent with other models.

Location, Bounds and other higher Characterisation levels are intentionally approximations to provide a simple inclusive description of the data, they do not themselves have errors; for example, Bounds should be the outer limits to anywhere data might be found. Accuracy properties describe uncertainties in data values, see Section 2.5

The values for some elements must be given as arrays, defined as in STC, and the required number of arguments MUST be present if any are. Bounds, for instance, describes a unique region on an axis as e.g. ($\alpha 1$ $\delta 1$, $\alpha 2$ $\delta 2$), whilst the resolutionSupport is given relatively e.g. telescope beam major and minor axes in arcsec and position angle.

2.1.1 Defaults

In some places we mention the use of defaults for values not provided; we do not think that these will be coded into the description, but software which

looks for a value of a missing element might be able to make an intelligent assumption. It is up to the writers of a software tool specification to decide whether it is more dangerous to use defaults and risk a lower level of accuracy, or to ignor data which is not completely specified as required and thus lose potentially important information. For example:

If Location is not given then, for some Axes, software may take the default Location as the mid-point of Bounds[1].

If Bounds are not given then e.g. if a spatial axis has Coordsys ICRF some software might assume all-sky coverage[2]. In other cases e.g. an error, the Location or reference value could be taken if minima and maxima are not provided.

If Support is not given Bounds, if present, could be used.

If the unit or Coordsys is not given for any level, the values for the Axis Frame are used; be careful, as this may be unsuitable (e.g. if the Axis Frame units are sexagesimal, then a single number for an error could be in degrees or arcsec or ...).

2.2 Axes

A description MUST provide at least one axis ("Axis Frame", "Axis").

All three of the Space, Time and Spectral Coverage Axes SHOULD be given[3].

The unit and coordsystem MUST be given for each Axis present (these may be relative to an internal reference only, e.g. (x,y) spatial coordinates. In such a case both the Location and Bounds MUST be given on that axis).

For Space-, Spectral- and Time-related axes and any others defined in STC, the unit and coordsystem MUST be taken from STC definitions. *We need to supply at least one example of each type of STC usage*

The Observable Axis SHOULD be given[4].

Other Axes e.g. Velocity MAY be given.

[1] this might be complicated (e.g. some spatial coordinates) or impossible

[2] a more restricted coverage might be derived once there is a link to Observation and the telescope location

[3] some might be considered irrelevant for simulated data, or not conventionally provided e.g. for old spectra with no time stamp

[4] its omission may seem reasonable for e.g. the coverage intended for a future survey

2.3 Coverage

The value MUST be given for either the Location or the Bounds on each Axis.

Values SHOULD be given for both Location and Bounds.

Support SHOULD be given.

If Support is given then Bounds MUST be given[5].

Sensitivity MAY be given (e.g. the URI of a weight map, or a function)[6].

A different Unit and/or CoordSystem MAY be used for each of these coverage layers, if not they will default to the units and CoordSystem used for the Axis Frame (i.e. when the axis was first defined).

FillFactor SHOULD be given on any axis where the actual coverage in each Support region is significantly less than 1 but the filling is too complex to be described practically using Sampling[7]. FillFactor MAY be given in addition to SamplingPrecison in which case, if SamplingPeriod and SampleExtent are given:

FillFactor = SampleExtent/SamplingPeriod

and the data provider should take care that the values and units given are consistent with this relationship.

2.4 Other Axes: Accuracy, Resolution and Sampling Precision

Accuracy, Resolution and Sampling Precision relate to a specific Coverage Axis Frame but themselves contain the relevant layers, e.g., for some axis:

Location ≡ typical or average value error
Bounds ≡ lowest and highest errors present
Support ≡ sets of discrete ranges of sampling intervals
Sensitivity ≡ the variability of resolution with location on the axis

If there are many areas of Support within the coverage, the Accuracy, Resolution and Sampling Precision refer to the inside of each Support area. It is assumed that, on any one axis, one description of Accuracy applies to all Support areas, otherwise each area must be described in a separate document (see Section 1.1). The same applies to Resolution and to Sampling Precision.

[5]If different areas of Support apply on different axes, a separate description should be used at the level where each subset of data can be described unambiguously, see Section 1.1

[6]Here, Sensitivity is the dependence of a detector response or equivalent with position on the given axis. This is not the limiting sensitivity in the sense of the faintest detectable flux, which is given by the lower Bound of the Observable axis

[7]FillFactor applies to the usable fraction of data within each Support area, as presently defined. If we find that the majority of users want it to be the useful fraction of the whole Bounds, the name and definition will be changed in a future version.

Resolution and/or SamplingPrecision MAY be present and if so, for each,:

You MUST give the unit and Coordsys on axes where the units of the Axis Frame would not make sense or are ambiguous otherwise the Axis Frame values are used.

You MAY give the the unit and Coordsys for any Accuracy, Resolution or Sampling level, otherwise the value defined at the start of the Accuracy, Resolution or Sampling definitions is used.

2.4.1 Resolution

If Resolution is present:

You MUST give resolutionRefVal (i.e. Location)

You SHOULD give resolutionBounds

You MAY give resolutionSupport.

You MAY give resolutionVariability (as a function of position on that axis).

2.4.2 Sampling Precision

If SamplingPrecision is present:

You MUST give a samplingPrecisionRefVal (i.e. Location) for the samplingPeriod.

You SHOULD give a samplingPrecisionRefVal for the sampleExtent.

You SHOULD give Bounds for the samplingPeriod and the sampleExtent

You MAY give Support for the samplingPeriod and the sampleExtent

You MAY give samplingPrecisionVariability (i.e. Sensitivity) in the form of a samplingPrecisionMap to describe variations along an axis.

2.5 Accuracy

Each Axis Frame SHOULD have associated Accuracy values for the precision of measurements along that axis, divided into statistical and systematic uncertainties. For each axis frame where Accuracy is provided:

You MUST give the unit and Coordsys on axes where the units of the Axis Frame would not make sense or are ambiguous the Axis Frame values are

used.

You MAY give the unit and Coordsys for any accuracy axis[8]

You MUST give the ErrorRefVal (typical value i.e. Location)

You MAY give the ErrorBounds for uncertainties which vary along the domain of the axis

You MAY give an ErrorMap (as a URI) to describe the variation of errors with location.

2.6 Flags

For each Axis Frame (spatial, spectral, observable (e.g. flux) etc.):

There SHOULD be a flag to indicate if it is an independent or a dependent variable ('true' or 'false').

There SHOULD be a flag to indicate its calibration status (CALIBRATED, UNCALIBRATED, RELATIVE, NORMALIZED; default UNCALIBRATED).

There MAY be flags to indicate Sampling Status (which SHOULD be provided where it is customarily relevant):
undersamplingStatus ('true' or 'false').
regularsamplingStatus ('true' or 'false').

[8]for example normalised units such as a flux accuracy of 0.03 (3given flux measurement.