Data Model for Astronomical DataSet Characterisation

Version 1.0

IVOA *Working draft*

October 16, 2006

Working Group:
http://www.ivoa.net/twiki/bin/view/IVOA/IvoaDataModel
This version: Presentation of the Characterisation Data Model and its XML representations

Editors:
Jonathan McDowell, François Bonnarel, Igor Chilingarian, Mireille Louys, Alberto Micol, Anita Richards.
Authors:
IVOA Data Model Working Group

Abstract

This document defines the high level metadata necessary to describe the physical parameter space of observed or simulated astronomical data sets, such as 2D-images, data cubes, X-ray event lists, IFU data, etc.. The Characterisation data model is an abstraction which can be used to derive a structured description of any relevant data and thus to facilitate its discovery and scientific interpretation. The model aims at facilitating the manipulation of heterogeneous data in any VO framework or portal.

A VO Characterisation instance can include descriptions of the data axes, the range of coordinates covered by the data, and details of the data sampling and resolution on each axis. These descriptions should be in terms of physical variables, independent of instrumental signatures as far as possible.

Status of this document

This is an IVOA Note expressing suggestions from and opinions of the authors. It is intended to share best practices, possible approaches, or other perspectives on interoperability with the Virtual Observatory. It should not be referenced or otherwise interpreted as a standard specification. A list of *current IVOA recommendations and other technical documents can be found at http://www.ivoa.net/Documents/.* A more preliminary version of this work announced on the DM list in April 2005 is available *at http://alinda.u-strasbg.fr/Model/Characterisation/characterisation.pdf* which includes a previous revision history. This version has been significantly reorganised so that a detailed list of changes would be unwieldy; instead, we note the relationships with other models, notably the rest of the Observation model, Quantity and STC.

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

This document defines an abstract data model called "Data Set Characterisation" (hereafter simply "Characterisation"). In this Introduction we present requirements and place the model in the broader context of VO data models. In Section 2 we introduce the concepts (illustrated with some examples) and discuss their interactions. In Section 3 we present a formal UML class model using the concepts defined earlier. XML and VOTABLE serializations are presented in Section 4 and the Appendices give further examples.

1.1 The purpose of the Characterisation model

Characterisation is intended to define and organize all the metadata necessary to describe how a dataset occupies multidimensional space, quantitatively and, where relevant, qualitatively. The model focuses on the axes used to delineate this space, including but not limited to *Spatial* (2D), *Spectral* and *Temporal* axes, as well as an axis for the *Observable* (e.g. flux, number of photons, etc.), or any other physical axes. It should contain, but is not limited to, all relevant metadata generally conveyed by FITS keywords.

Characterisation is applicable to observed or simulated data[1] but is not designed for catalogues such as lists of derived properties or sources (see Section 1.2).

The model is intended to describe:

- A single observation;
- A data collection;
- The parameter space used by a tool or package accessed via the VO.

The model describes the available data, not its history. For instance, spatial resolution expresses the level of smearing of the true sky brightness distribution in a data set without differentiating between contributions from different atmospheric, instrumental and software processing effects (see Section 1.2).

Characterisation has to satisfy two sets of requirements:

I Data Discovery requirements:

This model prescribes elements for use in requests to databases and services and thus forms a fundamental part of the standards for VO requests). The use of this model should enable a user[2] to select relevant observations from an archive efficiently. The selection will be based purely on the geometry of the observations, that is, how and how accurately the multidimensional space is covered and sampled.

[1] Unless otherwise stated, we use the terms "dataset", "observations" etc. to mean any applicable observed or simulated data.

[2] A user is either a human or a software agent

Discovery may only require a simplified overview (e.g. position, waveband, average spatial resolution). The user may opt for the inclusion of data where there is insufficient information to respond to certain parts of a query. Eventually, it should be possible for a client to generate a detailed multidimensional footprint of an observation. For example:

- What observations from a particular archive are likely to have covered a specific VO Event? (Spatial and Temporal Coverage)
- Which CCD frames in a mosaic actually cover the position of a particular galaxy? (detailed Spatial Coverage)
- What observed spectra have a resolution comparable with a given simulated spectrum e.g. matching the Shannon criterion? (Sampling Precision).

II Data Processing/Analysis requirements:

Characterisation should detail the variation of sensitivity on all relevant axes (e.g. variation of sampling or sensitivity across the field of view, detailed bandpass function), in order to provide data to an analysis tool or for reprocessing.

Errors may be provided for any or all axes.

Version 1 will fulfill all Data Discovery requirements, and allow some simple automatic processing such as cross-correlation and data set comparisons. Full implementation of Data Processing/Analysis requirements will only become available in a future version of this model.

1.2 Links to other IVOA modeling efforts

Characterisation arose out of the "Observation Data Model", a high level description of metadata associated with observed data, described in an IVOA note available at `http://www.ivoa.net/Documents/latest/DMObs.html`. The connection is summarised in Fig.1. It became obvious that there was an urgent need for a model to characterise the physical properties of data, alongside Provenance, DataCollection, Curation etc. (which provide instrumental, sociological and other information). For example, Provenance will be linked with Characterisation to provide the telescope location (needed for some coordinate transformations), calibration history, etc.

Characterisation complements and extends some of the metadata adopted by the VO Registry (`http://www.ivoa.net/Documents/latest/RM.html`), providing the finer level of detail needed to describe individual datasets.

Data models for Catalogues and Sources are also being developed.

Ideally, all these models must be mutually consistent and employ the definitions supplied by the STC and Quantity models (see Section 3.4), but some overlap and duplication is required to allow data and service providers to use the parts they need without excessive effort.

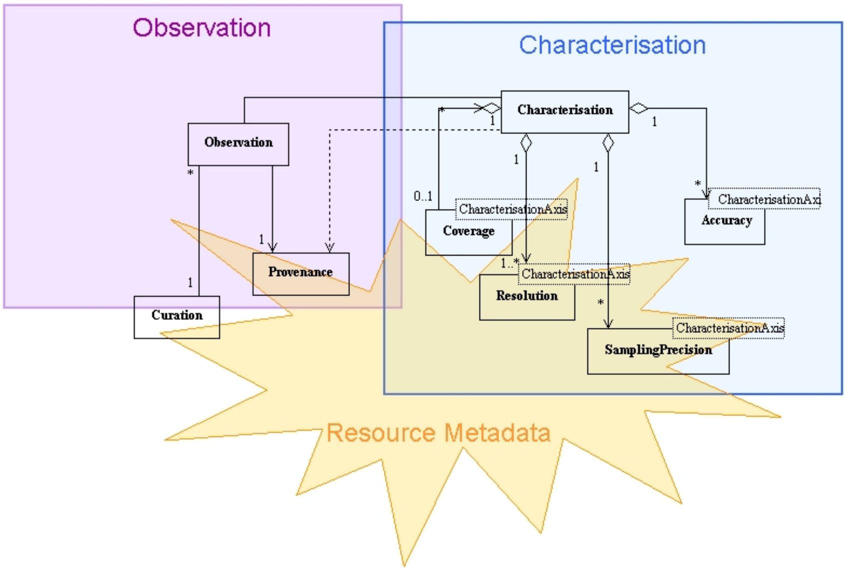

Figure 1: *Interaction between the Observation and Characterisation data models: Characterisation focuses on the physical information relative to an observation. Data management aspects such as the VO identifier, data format, etc.. are handled elsewhere in the Observation model.*

2 Exploring the Characterisation concepts

2.1 Overview: a geometric approach

We introduce the physical axes used to define the N-dimensional space occupied by any data set or required for interpretation. When considering a typical astronomical observation, we have identified various Properties:

- *Coverage:* describes what direction the telescope was pointing in, at which wavelengths and when; and/or the region covered on each axis. This is described in increasing levels of detail (see Section 2.6.1) by:
  - *Location*
  - *Bounds*
  - *Support*
  - *Sensitivity*

  If the data contain many small regions then the Bounds may be qualified by a

  - *Filling Factor*

  (especially if the Support is not precisely defined).

- *Sampling Precision:* describes the sampling intervals on each axis;
- *Resolution:* describes the effective physical resolution (e.g. PSF, LSF, etc.).

Each property can be related to one or more physical axes, described in more detail in Section 2.6. For each axis:

- *Accuracy:* describes the measurement precision, see Section 3.2.2.

2.2 Examples of Characterisation

The tables below illustrate how the spatial, temporal and spectral domains and the observable quantity of some typical data sets can be described, at various levels of complexity, using the properties from Section 2.1. Table 1 shows some of the Characterisation metadata for an X-ray event list. Additional examples are presented in Appendix C : Table 2 for a 2-D image, Table 3 for a 1D spectrum, Table 4 for an IFU Dataset, Table 5 for a radio interferometry image service and Table 6 for simulated data.

In some of these examples, some concepts are interdependent, discussed further in Section 2.4.1.

All these concepts can be applied to any data set but some elements may not have defined values, or the origin may be arbitrary, for example the spatial location of a generic simulated galaxy cluster (Table 6).

2.3 Structure and development strategy

Characterisation provides a framework to present the metadata necessary to specify a dataset in a standard format and to make any interrelationships explicit. The description can be presented from the perspective of the Properties or the Axes in a succession of progressively more detailed description layers. This will allow evolution of the model in three independent directions: new properties may be added as well as new axes, and if necessary new levels of description may be considered without breaking the overall structure.

2.4 The Axis point of view

2.4.1 Axes and their attributes

The physical dimensions of the data are described by axes such as: SPATIAL, SPECTRAL, TIME, VELOCITY, VISIBILITY, POLARISATION, OBSERVABLE. We recommend that data providers use these axes names but this is not compulsory (e.g. FITS names can be used). The data provider will be

Properties/Axes	Spatial	Temporal	Spectral	Flux/Observable
Coverage				
Location	Central position	Mid- Time	Central energy	Average flux
Bounds	RA,Dec [min,max] or Bounding box [center, size]	Start/stop time	Energy [min, max]	Probability above background (min) Pileup (max)
Support	FOV as accurate array of polygons	Time intervals (array)	Energy filter intervals (array)	
Sensitivity	Quantum efficiency (x,y); vignetting		ARF (effective area) as fn of energy	Out-of-time events (saturation); wings of PSF
Filling factor	Good pixel fraction	Live time fraction	not used	
Resolution	PSF (x,y) or its FWHM	Time resolution	RMF (spectral redist. matrix)	SNR
Sampling Precision	Pixel scale (x,y)	Frame time	PI bin width	ADU quantization

Table 1: *Property versus Axis description of metadata describing an X-ray CCD Event List. This also characterises the potential images and other products which can be derived. During exposure, the instrument moves with respect to the sky, so, for example, the sensitivity is a function of the support on the first three axes.*

required to supply a UCD for each axis, as well as the units. This ensures uniqueness and recognition by standard software. There is no limit on the number of axes present and they may be dependent or overlapping (e.g. one frequency axis and two velocity axes, representing the velocities of two separate molecules with transitions at similar frequencies).

Some axes may not even be explicit in the data, but are implicit, present only as a header keyword or elsewhere. For example, a simple 2D sky image usually has celestial coordinate axes, but the time and spectral axes may not be present in the main data array although the observation was made using a finite integration time and wavelength band (a single sample on each of the temporal and spectral axes). These implicit axes may be represented in Coverage to provide their location an/or bounds, or even, for purposes such as color corrections, their sensitivity as a function of the coordinate within the bounds.

2.4.2 Axes flags

Axes flags (Section 3.2.1) are used to indicate Boolean and other qualifying properties. These include whether the axis represents a dependent variable (e.g. the Observable); the calibration status and whether the data are undersampled.

2.5 Accuracy

Accuracy characterises any uncertainties associated with each axis (Section 3.2.2) – astrometric uncertainties are attached to the Spatial axis, photometric to the Observable etc. Note that this is a level of detail distinct from the assessment of the overall accuracy of data provided by the Registry metadata.

2.6 The Property point of view

The main properties needed for data description and retrieval are categorized under Coverage, Resolution, and SamplingPrecision, introduced in Section 2.1.

The values of the properties characterising an Observation may be derived from instrumental properties given in Provenance or from other Characterisation features. For example, high energy missions move the telescope during the observation (Table 1), leading to a time-variable mapping from detector to celestial coordinates (the 'aspect solution'), giving a spatially variable effective exposure time derived from the temporal bounds multiplied by the filling factor, or the sum of all the support intervals weighted by sensitivity, or derived from the sampling precision and period within the bounds. The sensitivity across the spectral band may be a function of spectral position (ARF). Such dependencies should be restricted to areas of significance to users, such as the Sensitivity class. At present, a single value, or the extrema, can be given for each element; more complex formulae will be available in a future version of Characterisation.

2.6.1 Coverage

Coverage has several levels of depth, providing a range of detail to meet the needs of any user/developer, illustrated in Fig. 2. The simplest approximation to a spatial field of view presumes that a sharp-edged region of the celestial sphere has 100% sensitivity inside and 0% outside. In reality the transition is fuzzy and the region may be irregular and contain gaps. For example, some applications only need to know what range of coordinate axes values might contain data; others need to know the variation in (flux) sensitivity as a function of position on an axis. Coverage provides answers to these questions at different levels of precision, with the idea that software implementations will be able to convert between the levels.

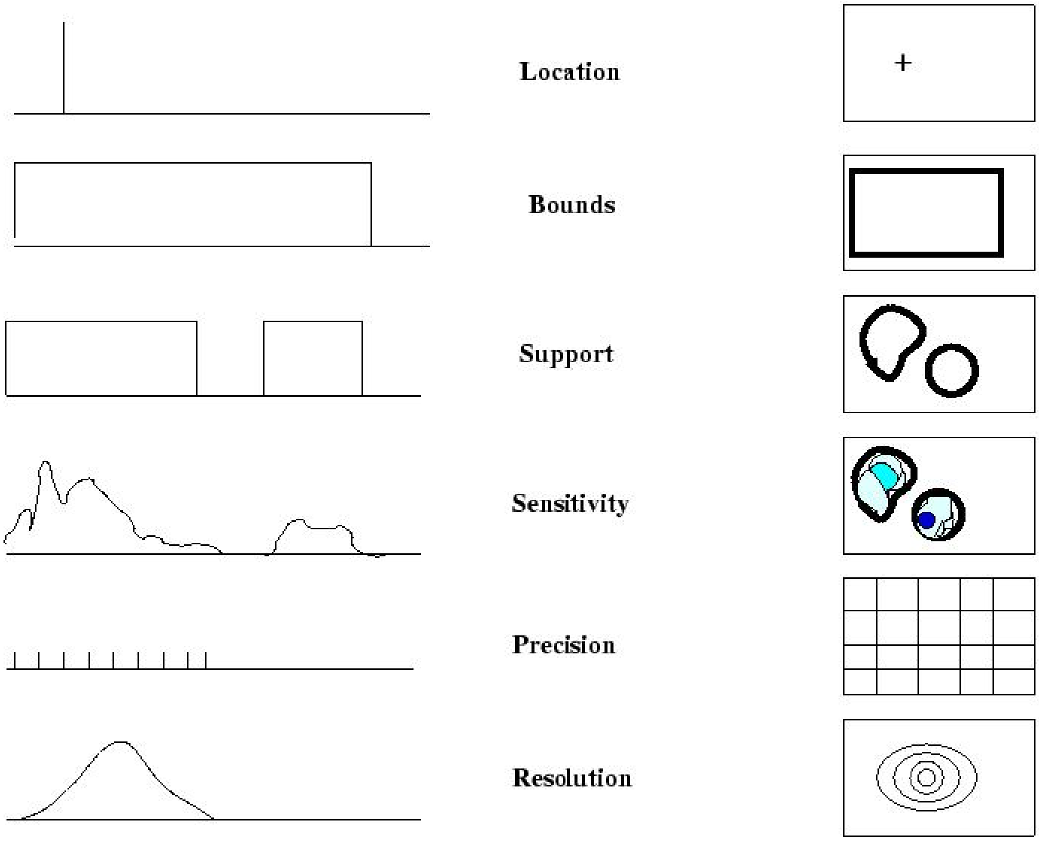

Figure 2: *Illustration of the different levels of description. left: for a 1-dimensional signal, right: for a 2D signal.*

Coverage is described by four layers which give a hierarchical view of increasing detail:

1. *Location:* The simplest Coverage element is the Location of a point in N-dimensional parameter space, such as an image described by a single value each of RA, Dec, wavelength and time. These are fiducial values representative of the data. A precise definition (mean, weighted median, etc.) is not required, but Location can serve as a reference value or origin of coordinates in frames with no absolute position (e.g. Table 6).

2. *Bounds:* The next level of description is the SensitivityBounds, i.e. a single range in each parameter providing the lower and higher limits of an N-dimensional "box". The scalar intervals between the limits (the sizes and centres of each box-side) should also be available if required. The Bounds are guaranteed to enclose all valid data but there may be excluded edge regions for which there is no valid data, such as (on the wavelength axis) the 'red leak' end of a spectral filter. These provisions satisfy the intent of typical data discovery queries.

3. *Support:* Mathematically, the support of a function is the subset of its domain where the function is non-zero. Here, Support describes quantitatively the subsets of space, time, frequency and other domains, onto which the observable is mapped, where there are valid data (according to some specified quality criterion). Support may include one or many ranges on each axis (e.g. Table 4).

4. *Sensitivity:* Sensitivity, (unlike the previous 'on/off' properties), provides numerical values indicating the variation of the response function on each of the axes, such as the relative cell-to-cell sensitivity in the data. This includes filter transmission curves, flat fields, sensitivity maps, etc. The final limits on Sensitivity are determined by the bounds of the Observable; for example, the minimum and maximum given by a single count and by the saturation level for some types of detector.

The *Bounds* may also contain the

- *Filling Factor* sub-level, which gives the useful fraction of Bounds on any axis. It may not be appropriate to detail multiple small interruptions to data (for example detectors requiring dead time between each sample) if it is conventional for analysis systems to solve the problem using a statistical correction based on the Filling Factor. Very regular filling may also be described by Sampling (see below). Even if Support provides a complete description, the Filling Factor may be used to rank the suitability of data during discovery.

A method should be provided to derive the Filling Factor from the Sampling Extent and Sample Precision (Section 2.6.2) if these are are given, but if all three values are entered separately there needs to be a means of checking consistency.

2.6.2 Resolution and Sampling Precision

Resolution is often a smoothly decaying (e.g. Gaussian) function but the data product is subject to further discrete Sampling, e.g. CCD pixels, Table 2. Resolution may, however, be a top hat function determined by the Sampling interval – e.g. the temporal resolution of an image made from a single integration. We maintain a distinction between the concepts to facilitate different requirements in data processing, whether during data discovery services which allow resampling or flexible resolution (Table 5), or during post-discovery processing (Table 4).

- *Resolution* Resolution is usually the minimum independent interval of measurement on any axis. Mathematically, if the physical attributes (e.g. position, time, energy) of the incident photons, or other observable, are $\mathbf{x}$ (e.g. x_0 = energy, x_1 = RA, x_2 = Dec, x_3 = time, etc.), and the measured attributes are $\mathbf{y}$ (e.g. y_1 = spectral channel, y_2, y_3 = pixel position, y_4 = time bin) then given a flux of photons $S(\mathbf{x})$ the detected number of photons is

  $$N(y_1, y_2, ...) = N(\mathbf{y}) = \int \mathbf{S}(\mathbf{x})\mathbf{A}(\mathbf{x})\mathbf{R}(\mathbf{x}, \mathbf{y})\mathbf{dx}$$

  where A is the probability that a photon is detected at all (the quantum efficiency) and $R(x_1, x_2, ..., y_1, y_2, ...)$ is the smearing of measured values (PSF, line spread function, etc.).

  In the most detailed case, **R(x, y)** may be a complicated function, such as a PSF which varies as a function of detector position and energy. The first level of simplification is to specify a single function which applies to the whole observation - e.g. a single PSF. This function may either be provided as a parameterized predefined function (e.g Gaussian) or as an array. The concept of Resolution Bounds provides the extreme values of resolution (see Table 5)

  The final level of simplification is to give a single number characterising the resolution, such as the the standard deviation of a Gaussian PSF.

- *Sampling*

  Sampling (or pixelization or precision or quantization) describes the truncation of data values as part of the data acquisition or data processing. If sampling is non-linear, simplification may be necessary, by giving limiting values or a single ‘characteristic sampling precision’. The Sampling Period gives the sample separation and the Sample Extent shows the deviation from the pure “Dirac comb” case. The Nyquist parameter – the ratio between the resolution FWHM and the Sampling period – will also be provided by a method. The Sampling flags (Section 3.2.1) provide a simple guide as to whether these properties are significant.

2.7 Presentation of layered information

The layered structure allows tasks to retrieve only the metadata which is actually required (see Section 1.1). The lower levels can be very detailed, for example the variation in Sensitivity to the Observable(s) along the spatial, spectral and other axes, or as described for Resolution, Section 2.6.2. This could take various forms:

- A simple value or range
- An analytic function of other property values
- A variance map for 2D data
- A look-up table for the bandpass correction to 1D spectral data

The more complex properties may be provided using pointers to ancillary data with the same types of axes and dimensions as the observation itself, e.g. a weight map packaged with a 2D image; this capability exists in the first version of this model. The provision of "attribute formulae" or attributes pointing to functional descriptions, such as the aspect solution for an X-ray observation (Section 2.6), is left for the future development of Characterisation; a first step may be to decompose a complex coupled description into non-coupled expressions. Where it is possible to provide separate values for interdependent elements (see also the end of Section 2.6.1), there must be a validation method to avoid contradictions.

A later version of the model will also allow links to other aspects of the Observation model (Section 1.2), external calibration and documentation. Advanced VO tools could use such metadata to recalibrate data on demand. Characterization is used to describe potential as well as static data products (e.g. Tables 1 and 5) and could therefore also provide pointers to Registry entries for tools and services capable of extracting images etc. from event or visibility data or atlas cut-outs.

3 The Model

3.1 The role and structure of the Model

We use UML diagrams to describe the organisation of Characterisation metadata following the Properties/Axis/Levels perspective. The model offers different views of the characterisation concepts. Figure 3 shows the relationships between the main concepts. The CharacterisationAxis box attached to each property class represents the axes along which the property (e.g. Resolution) is assessed; for example, there can be one Resolution class for each relevant axis. Fig 4 illustrates how the properties of the data are gathered under the Characterisation container class. The Coverage class is shown with the four increasingly detailed properties introduced in Section 2.6; such a Characterisation tree is available for each CharacterisationAxis value.

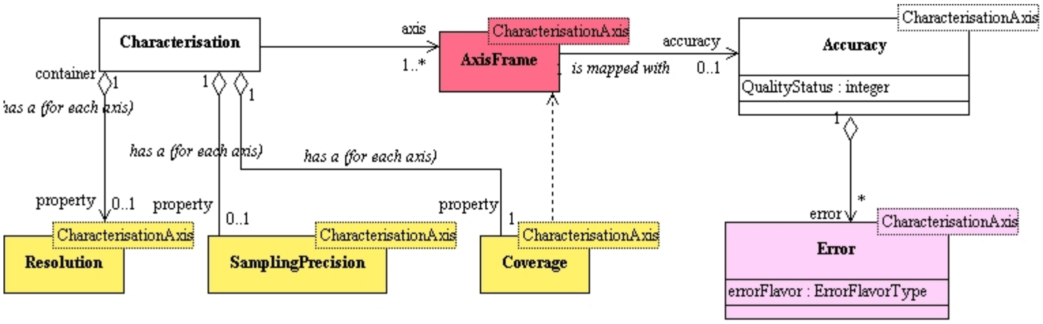

Figure 3: *This class diagram emphasises the Property/Axis perspective. The Characterisation class is a container that gathers the properties for each axis. The axis is described by the AxisFrame class. The link 'shows immersion in' represents the list of all relevant axes for one observation/dataset. The Characterisation axis is a template parameter for each Property, valid for all levels. The Accuracy class is linked to the Axis class and gathers different types of Error descriptions (systematic, statistical).*

3.2 AxisFrame

The information related to this CharacterisationAxis parameter is described by a specific class: AxisFrame. This can have common "factorised" attributes applicable to the property layers on that axis (Section 2.1). It is related to the Frame concept in Quantity, containing the UCD, units, name, and a holder for the STC coordinate frame (see Section 3.4) which also provides the base class for the observatory location (Observation – Provenance model).

If a deep level (higher number, Section 2.6) object, e.g. Sensitivity, needs to have its own AxisFrame object, this can be defined locally, overriding the factorised top level AxisFrame object. The redefinition can be partial, e.g. a change of unit or a change of spatial orientation requiring a new CoordFrame.

3.2.1 Flags and other qualifying information

Other elements in the AxisFrame class include the number of bins present on any axis, and flags to indicate the calibration status, independency and sampling properties of the axis, as described in Section 2.4.2

Independent or dependent status Axes may include both 'independent' variables (which may have associated errors) and the *"Observable"* axis or axes which represent phenomemena measured along some other axes. For

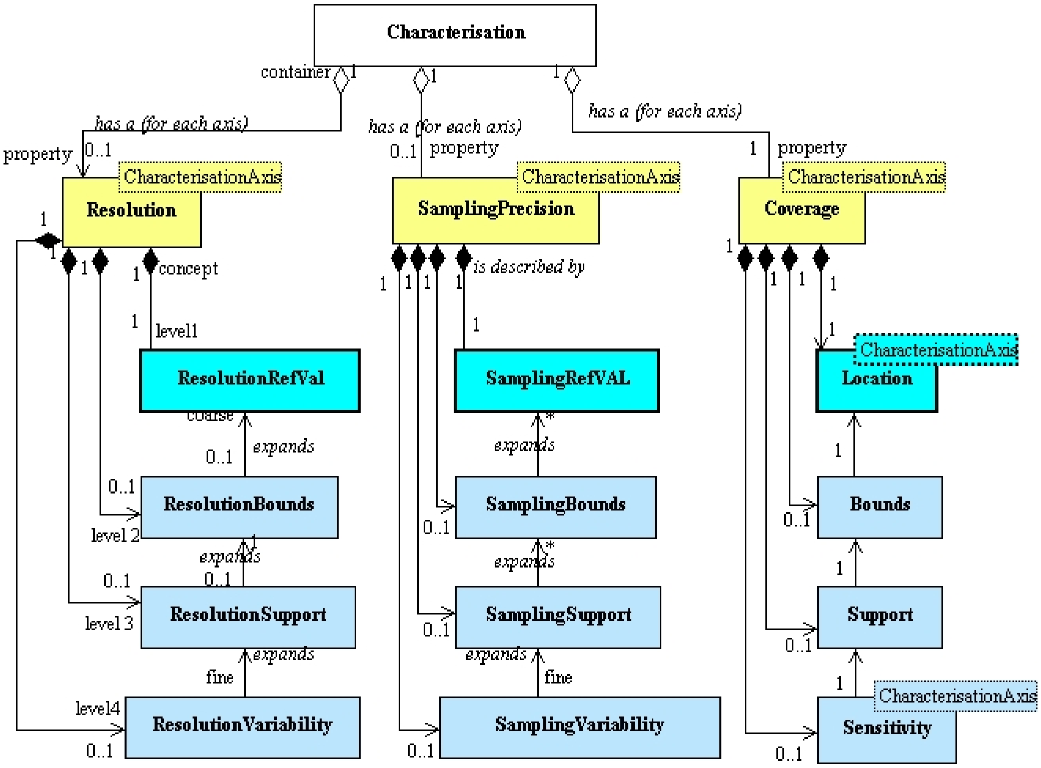

Figure 4: *UML diagram: The layered structure of characterisation. This diagram synthesises the Property/Axis/Layer approach. The concepts are represented in yellow. The coarse description is designed by the first level (blue boxes), while the pale blue ones represent the complementary metadata. The Bounds, Support and Sensitivity classes are nested levels of detail to add knowledge about the Coverage of an Observation. Symmetrically, Resolution and Sampling may also have the 4-level structure of description. The complete Characterisation for one observation is obtained by filling the tree for each relevant axis: spatial, spectral, temporal, etc.*

instance, in a 3D datacube of the sky, the Spatial axis is an independent axis (flag TRUE), as is the (implicit) Spectral axis, but the Flux axis is dependent (flag FALSE), and the velocity axis is dependent on the frequency axis.

Calibration status The AxisFrame object in the Characterisation model provides a calibration status flag for each axis, so that a user can insist on calibrated data only where necessary. The CalibrationStatus is given separately for each axis and can be

- UNCALIBRATED: not in units which can be directly compared with other data (but often still useful, for example the presence of spectral lines at known wavelengths can give a redshift regardless of absolute flux densities).
- CALIBRATED: in reliable physical units or other accepted units such as magnitudes.[3]
- RELATIVE: calibrated to within a constant (additive or multiplicative) factor which is not precisely known, such as arising from uncertainty in the flux density of a reference source.
- NORMALIZED: dimensionless data, divided by another data set (or a local extremum).

The calibration process itself is described elsewhere in the Observation Data Model (Section 1.2).

Sampling status

- Undersampling: TRUE if the sampling precision period is coarse compared to the resolution and the precision of a single data value is limited by the sampling; FALSE if the sampling precision period is small compared to the resolution and precision is limited by the resolution
- Regular sampling: TRUE if the pixellation or binning is close to linear with respect to the axis world coordinate (so that an accurate position can be obtained by counting samples from a Bound); FALSE if this would introduce an error significant with respect to other uncertainties.
- The total number of samples along each axis may be given, normally used for multiple regular sampling.

[3] In such cases the coarser levels of description should also be given in physical units and the need for a tool such as a look-up table of zeropoints etc. and conversion algorithms has been identified.

3.2.2 Errors in Characterisation: the Accuracy class

The values along Coordinate axes and measurements of Observables may all suffer from systematic and statistical uncertainties. Errors may be in the units of the axis or may be represented by quality flags. These Error classes are gathered in an Accuracy object (linked to the AxisFrame object, see Fig. 5, and related Quantity and STC data model elements, see Section 3.4)) which supports multiple levels of description, analogous to Coverage. The uncertainty in the position or measurement on any axis can be described by a typical value, by the bounds on a range of errors, and/or by very detailed error values for each sampling element (e.g. pixel).[4] A pointer may be provided to error maps packaged with the data, as described for the more detailed levels of Coverage (Section 2.7).

3.3 Navigation in the model: by axis or by properties?

The structure of Characterisation is clearly hierarchical with the characterisation class as the root element. The model can be serialised using two alternative sets of primary elements:

- *Properties*, with the corresponding classes for each axis attatched; used, for example, to represent data where the axes values are interdependent (e.g. Table 1);
- *Axes*, factorising each description into the multi-layer property levels; this provides more compact XML.

Either structure could be applied to the examples tabulated in Section 2.2. This UML model could be used to build two different XML schemas, giving access primarily by property or by axis. Here, we present the “Axis First” serialisation only; the “Property First” serialisation will be presented in the next version of this model.

3.4 Implementing the model using elements of Quantity and STC

The Quantity model (`http://ivoa.net/internal/IVOA/IvoaDataModel/qty.v0.2.pdf`) provides the means to supply values for dimensionality, coding, errors, units, UCDs and so on. Characterisation could make a fundamental use of the Quantity Frame class, as a template for its AxisFrame container. The Q:Quantity data type also provides for uncertainties. Any basic class such as Location , Support or Bounds, could also be implemented

[4] Measurement errors are distinct from any ‘fuzziness’ in the values provided by the coarsest levels of Characterisation, e.g. Location may be an arbitrary approximation (Section 2.6.1), but that kind of uncertainty is catered for by going to deeper levels of Characterisation, and by the concept of Region of Regard in the Registry Resource model.

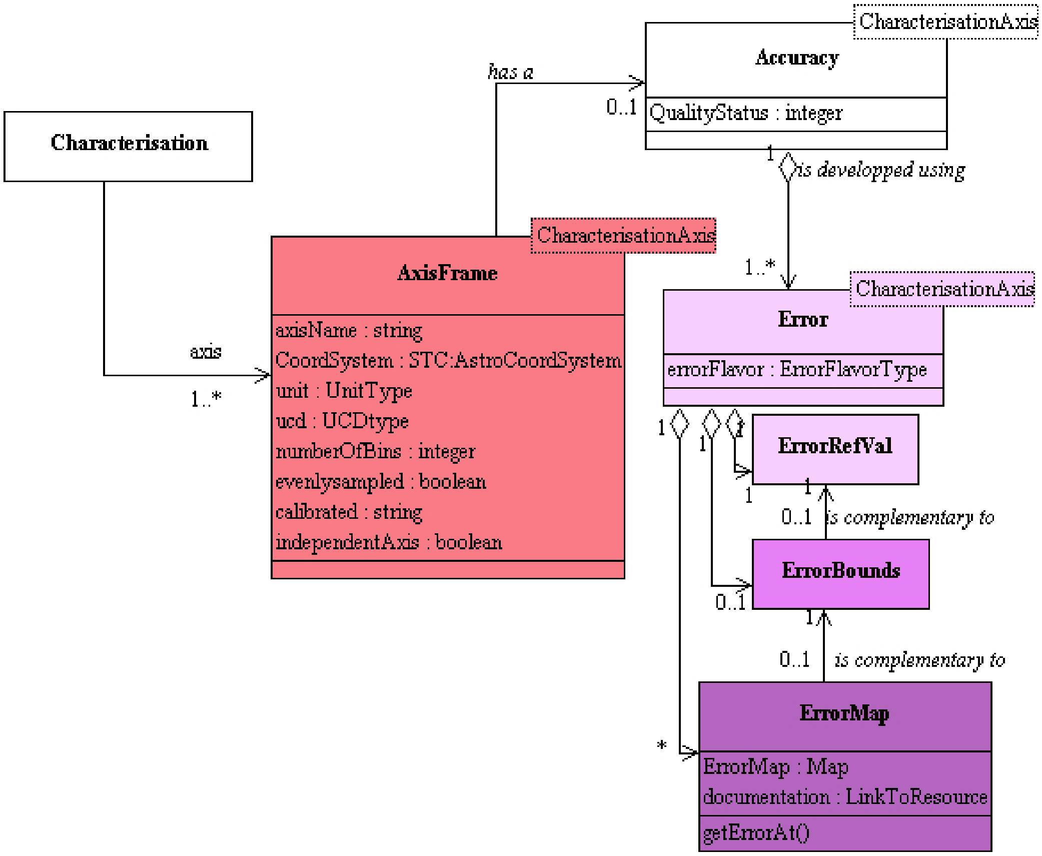

Figure 5: *This class diagram illustrates the AxisFrame class and its relationship with the Accuracy class, which encompasses various classes of errors such as systematic or statistical.*

as a Quantity, but this would require another relationship between the Quantity datamodel and STC.

STC, the metadata scheme for Space-Time Coordinates (see `http://www.ivoa.net/Documents/latest/STC.html`) encompasses the description of most of the Characterisation Axes examples in Section 2.4.1 with the exception of Observables. Sensitivity is the only Property not present in STC. However, the full STC structure cannot simply be reused, as it does not have the flexibility needed to deliver the alterntive schemata for both multilayered views presented in Sections 3.1 and 3.3. We do use STC intermediate level objects as building blocks for the Charactersation model.

The STC:AstroCoordSystem object is needed as a reference to define the Coverage axes. STC substructures may be reused in the following way:

- *Location* implements STC:AstroCoords
- *Bounds* encapsulates STC basic types, some STC:Interval elements and STC:Coords in a structure similar to STC:AstroCoordArea.
- *Support* uses STC:AstroCoordArea
- *Resolution* ResolutionRefval can be implemented via STC:CResolution
- *SamplingPeriod and SampleExtent* can be implemented via STC:PixSize elements .
- *Accuracy Error* is provided by reusing STC:Error elements in the same way.

This is represented for the spatial axis using implementation links in the UML diagram in Fig.6.

In simple cases Data handlers will probably reuse predefined elements included from an external STC library. For example, AxisFrame includes the STC elements for CoordSys and the (possibly variable) space-time coordinates of the ObservatoryLocation[5] or of the origin of coordinates (e.g. for barycenter-corrected data).

Many parameters (i.e. most numerical-valued elements at a finer level than Location) are customarily expressed either as maximum and minimum values or as a centre and scalar range (or both). In some cases an array of such values is needed, e.g. 2 dimensions on the spatial axis in most but not all cases; upper and lower bounds to (separately) the major and minor axes of Resolution in Table 5; higher dimensionality is possible such as the inclusion of beam position angle in this Resolution example.

The Resolution and Pixel-Size concepts are represented in STC at a deep level inside the Coordinates class (together with the Name/Value/Error in

[5] This should, where necessary, be consistent with the Provenance section of the Observation model (Section 1.2).

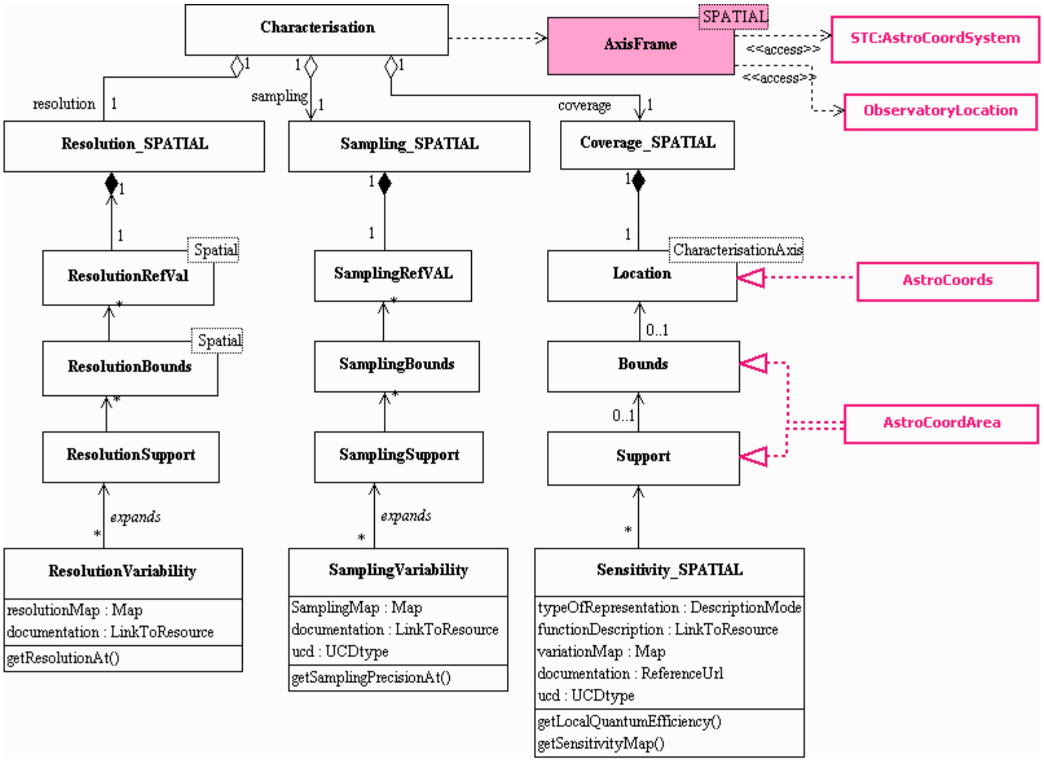

Figure 6: *UML diagram: Expressing the spatial properties as a subtree of Characterisation . Here is an example of how STC components (in pink italics) may be used to implement the different levels of the Coverage description. The Location element uses a STC:AstroCoords. Bounds encapsulates STC basic types some STc:Interval elements and STC:Coords in a structure similar to STC:AstroCoordArea.*

the Coordinate object). This allows any coordinate to be expressed to the appropriate degree of numerical precision. Characterisation needs to allow selection of metadata by resolution, which therefore must be accessible at the upper level of description. This can be achieved by linking Coordinates to the Resolution Class, which meets the requirements of Section 3.3.

Since the space, time and spectral axes are particularly important for astronomy, we recommend that implementations include a method to return an STC::AstroCoordSys object, which will only succeed if a complete and consistent space-time-spectral description is present. This may be nominal or arbitrary for some axes e.g. for simulated data.

4 XML Serialization

4.1 XML schema (Axis First)

4.1.1 Design of the schema

Due to the Hierarchical nature of the Model, the XML serialization of Characterisation is based here on a single tree. The appropriate elements are taken from STC and Quantity as described in Section 3.4. The root element called "Characterisation" is the aggregation of a set of CharacterisationAxis elements[6] for each of the axes. The AxisFrame element is defined at the top level of the CharacterisationAxis so that each axis can be given an obvious label ("spatial", "temporal"), etc. Coverage implements different elements according to the four levels of description detailed in Section 2.6.1. Lower levels of AxisFrame may be modified with respect to the top-level object for that axis as described in Section 3.2.

A full XML serialisation is provided, as an XML schema, for simple observations, at the following site:
`http://alinda.u-strasbg.fr/Model/Characterisation/schema/characterisation.1.0.xsd`.
An XML instance document describing an IFU dataset characterisation is available at
`http://alinda.u-strasbg.fr/Model/Characterisation/examples/MPFS.1.0.xml`.

Full implementation of Characterisation software classes will probably benefit from a version of this schema based on Quantity and STC. Nevertheless, more compatibility between these two schemata is obviously needed before doing that. A future schema could, for example, define a full high level STC structure together with the Characterisation types, with each STC

[6] These elements are containers gathering the result of the dynamic grouping of properties for a given characterization axis

element referring to the appropriate Characterisation element – a variant referring in the other direction is also possible.

4.1.2 Building blocks of the schemata

In order to illustrate how the XML schemata is derived from the UML Model, building blocks of the Schemata, corresponding to some main classes of the UML diagram are shown here.

The principle is to map the main classes in XML elements, building up a hierarchy from the most englobing concept down to more specific ones. Aggregated classes are easily translated as aggregated subelements. The attributes of an UML class are also coded as sublevel elements.

The translation from UML to XML used in this serialisation applies rules and elaborates specific techniques very similar to the work of Carlson (*Modeling XML applications with UML*, Addison-Wesley, 2001). The examples shown here are 'handmade' translations of the UML model. Automated translation will be discussed in the next version of Characterisation. The derivation of the XML from the UML model is expressed in the graphical views of the XML schema in Figs. 7, 8, 10, 11 and 12.

4.1.3 Utypes generation: select one ordering strategy

One application of such a model is to provide a naming convention for every metadata considered within the model, in order to be able to identify one concept in various models or serialisations. The idea is that by navigating in the model following the logical links provided, it is possible to construct identifiers called Utypes that could be understood by any VO tool aware of the model. To avoid multiplicity, the Utypes are built from the XML schema representation of the model which already enforces a hierarchical structure. For instance, the size of the sampling element along the spatial axis in a 2D image corresponds to:
Characterisation.SpatialAxis.SamplingPrecision.SamplingRefval.

4.1.4 VOTABLE serialisation

A VOTABLE serialisation of the characterisation of the IFU MPFS data set is shown in Appendix C. Each CharacterisationAxis is seen as a table, where each property itself is seen as a Group of Fields. UML class attributes are serialised as FIELDS. In this example, Utypes are set for each Table, Group, and Field according to the following rule:

> A Utype is elaborated for each VOTable item in the serialisation as a string based on instance variable paths in object-oriented programming mode.

Other ways of deriving utypes from a valid Xpath to the equivalent XML element in the XML Characterisation schema have been studied. The main

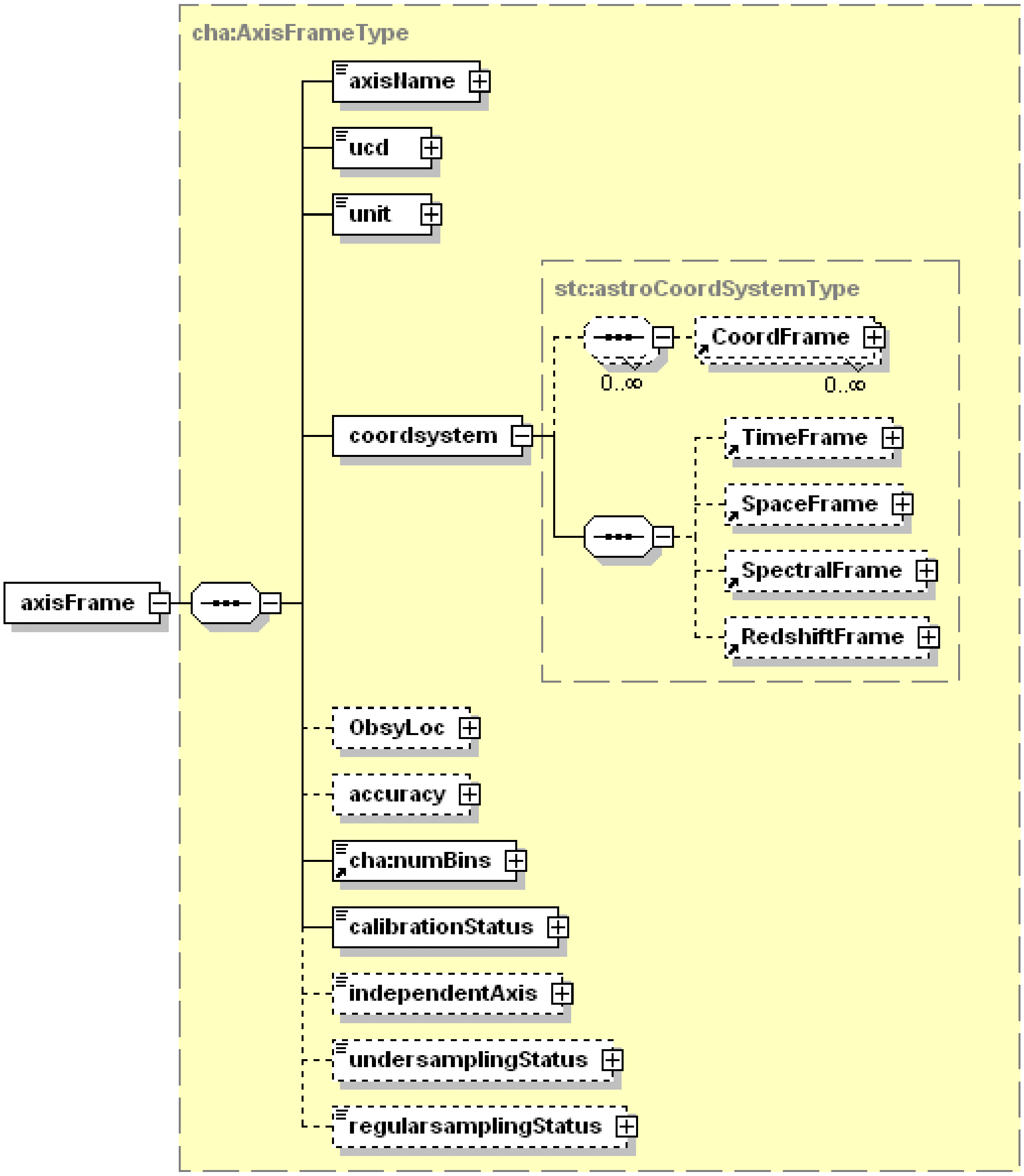

Figure 7: *The axisFrame element is built up following the corresponding UML class with* coordsystem *and* ObsyLoc *items reusing STC elements. The small arrow on cha:numbins represents a substitution group head element in XML. This allows to plug various constructs of this element (e.g. for 1D, 2D, 3D) that play the same role in the XML tree.*

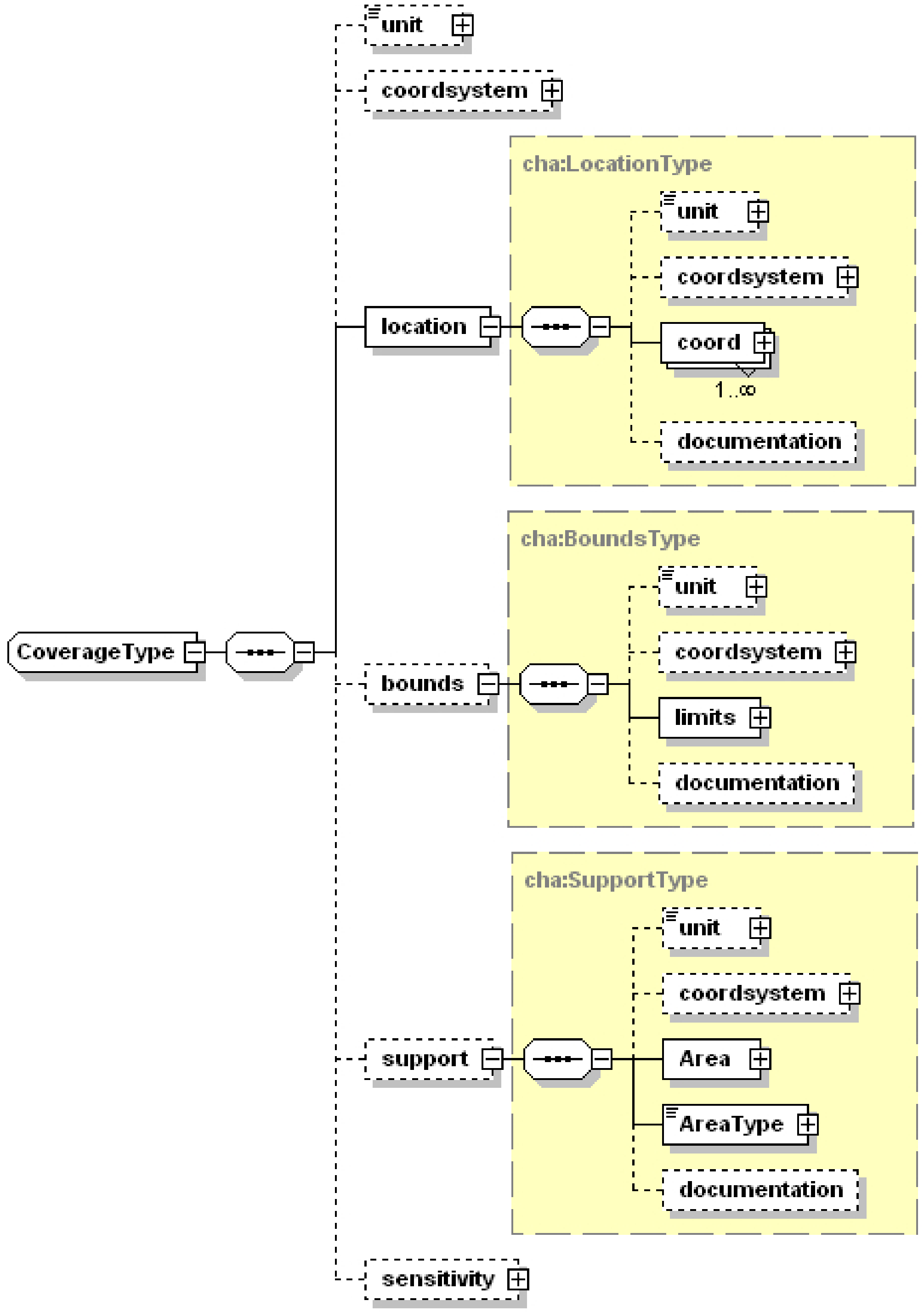

Figure 8: *The* coordsystem *and* unit *items can be factorised at the top of the Coverage structure, but may be redefined at each level when necessary.* Bounds *are expressed using a* limits *element which is developped on a general bounding box type:* CharCoordArea. AreaType *is a string describing the kind of region used:* Circle, Polygon *etc.*

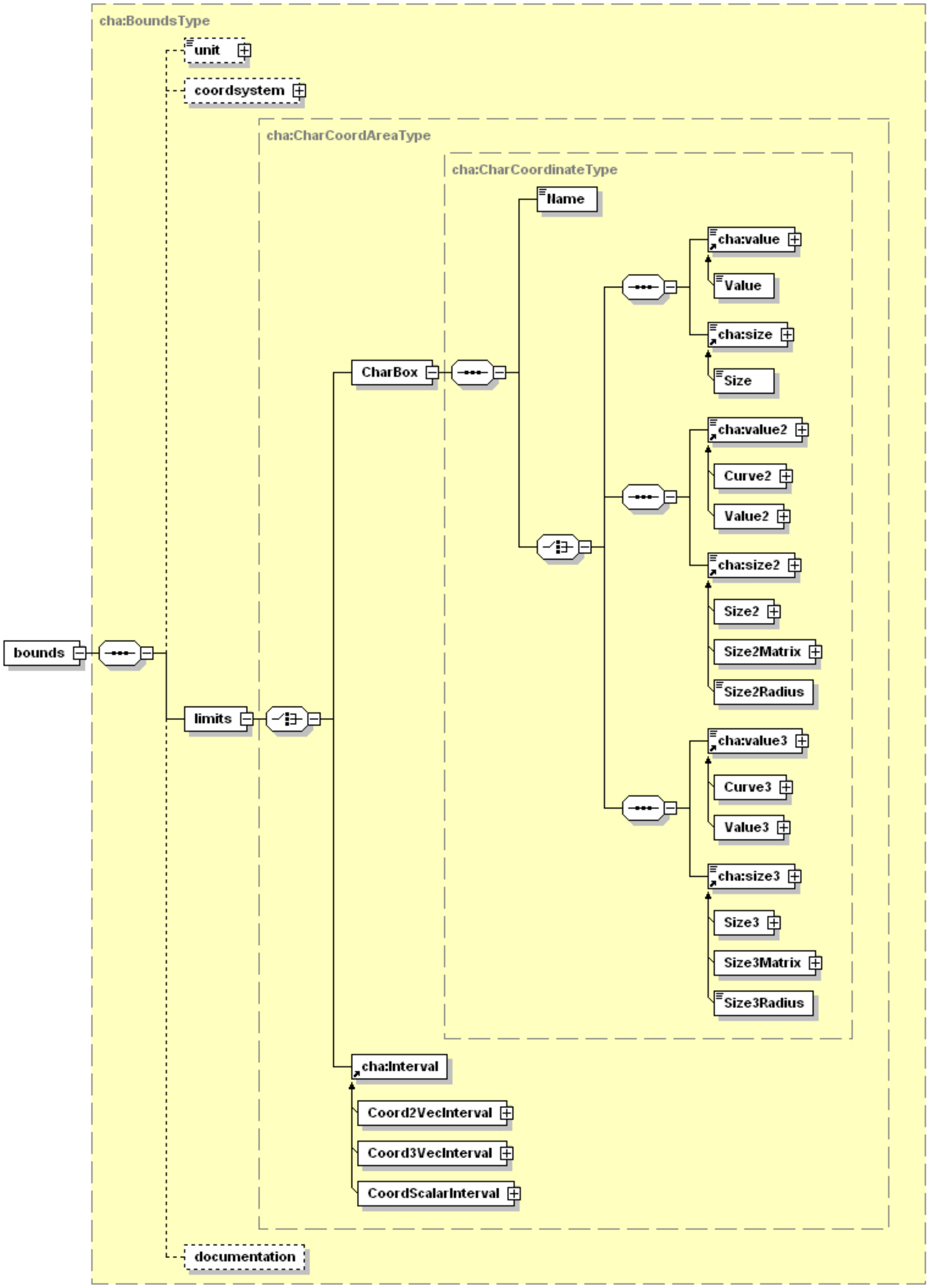

Figure 9: *Representing limits: The two expressions allowed for a bounding box are expressed using either a STC:CoordInterval embedded in a locally defined type cha:Interval or built on another type: CharBox representing a generic centered box in N-dimensions.*

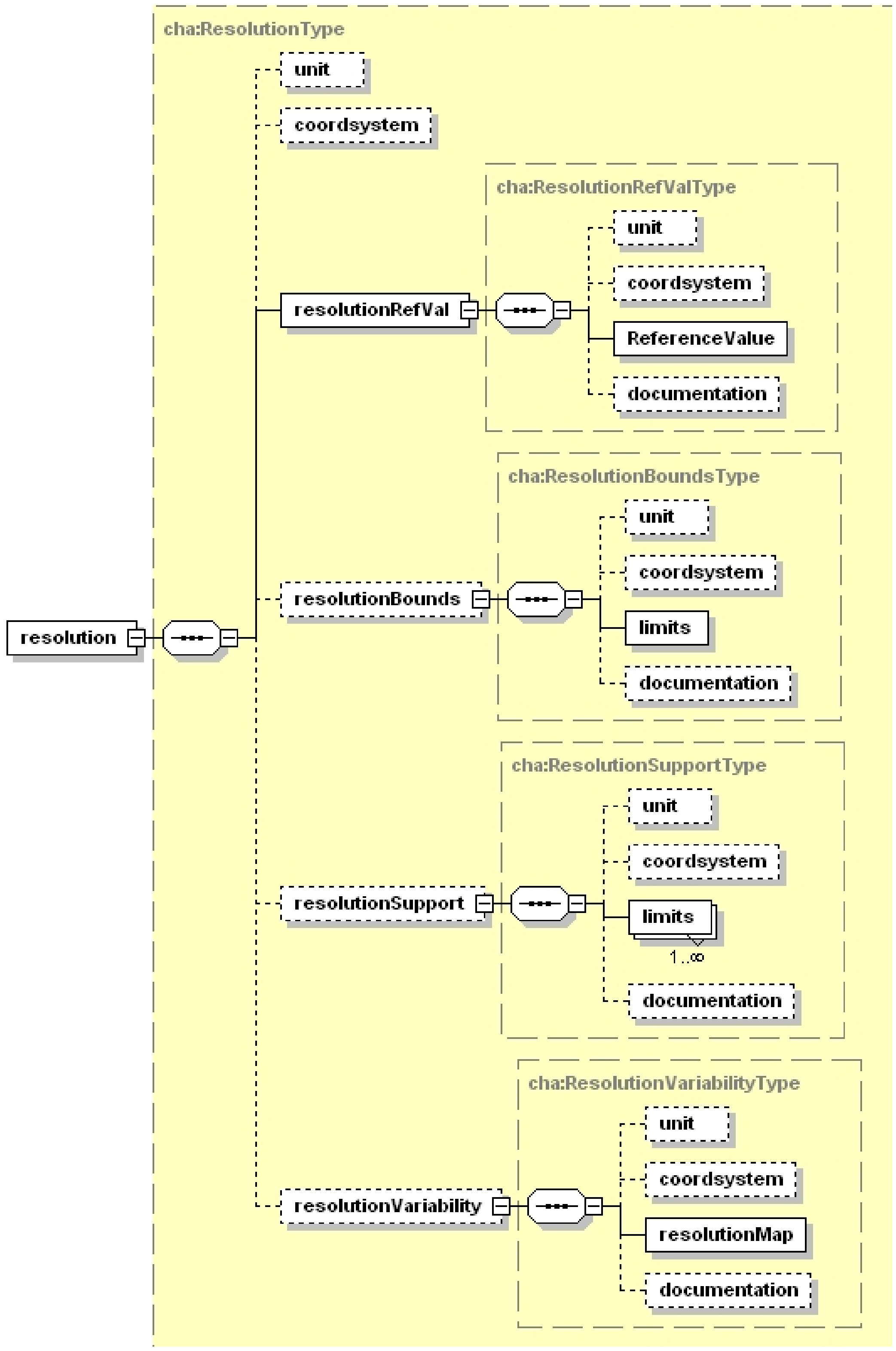

Figure 10: *This graphical view was generated with XMLSPY from the resolution element of the schema. As designed in the UML class, the resolution item contains 4 possible subelements. The RefVal element should be present but is not mandatory: some observations may have unknown resolution.*

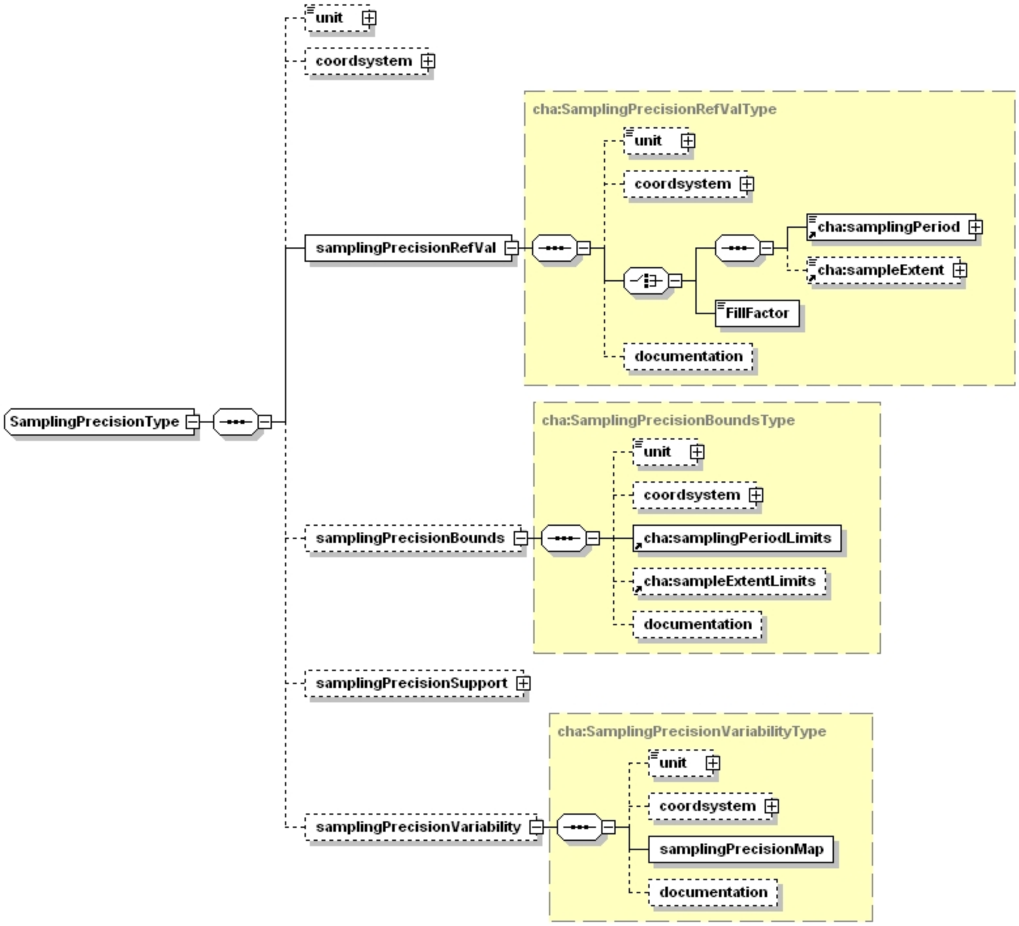

Figure 11: *The samplingPrecision item contains 4 possible subelements. One among SamplingPrecisionRefVal and SamplingPrecisionBounds should be present when possible but this is not explicitly described by the XML syntax.*

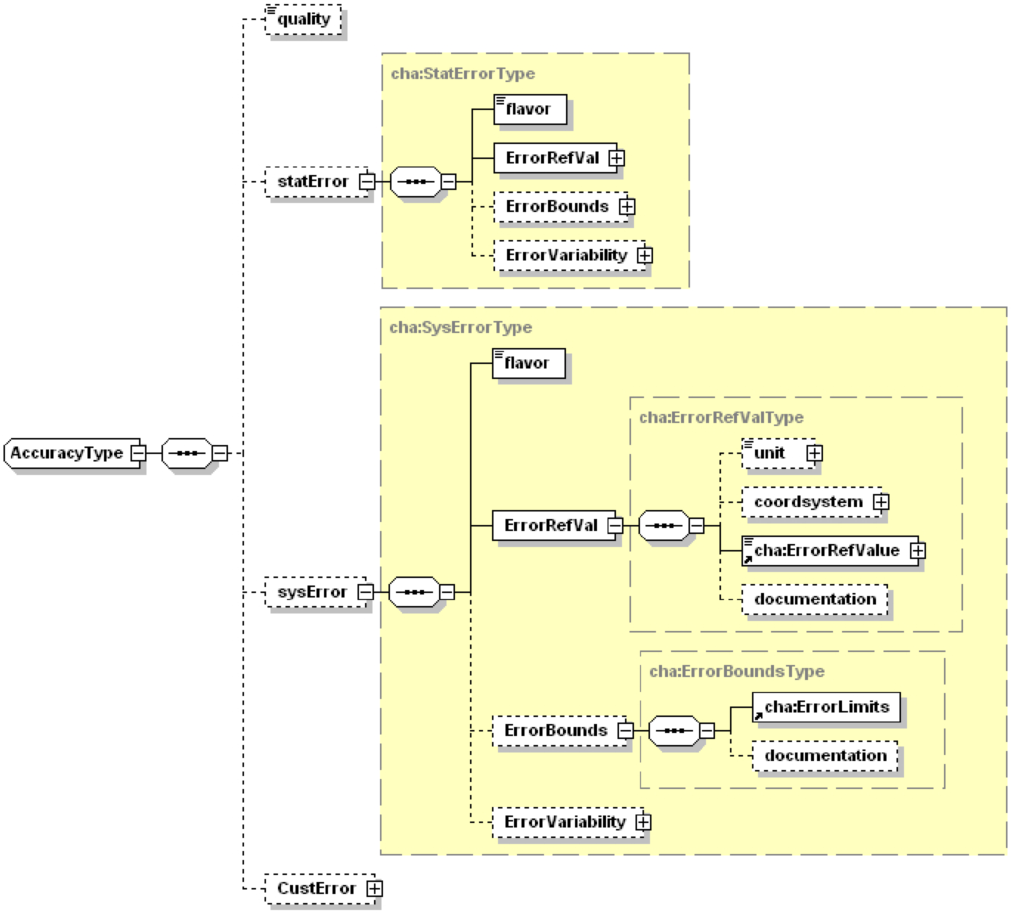

Figure 12: *The accuracy element relies on Errors along the axes and is built up on STC elements.*

difference is that this option may use constrained element (or attribute) values in the Utype path. The IVOA needs to define a single and robust rule to define this concept.

Appendix A: XML serialisation example

An XML instance document representing the characterisation of an IFU data set, taken with the Russian MPFS instrument. It relies on the XML schema mentioned above. See the update Xml document at :
`http://alinda.u-strasbg.fr/Model/Characterisation/examples/MPFS.1.0.xml`.

Appendix B: VOTable serialisation example

An alternative serialisation, using the VOTable format and applying the Utype mechanism to map the various items to the Characterisation Data Model classes and attributes. Utypes are derived from the Characterisation XML schema as mentioned above. See the full Xml document at :
`http://alinda.u-strasbg.fr/Model/Characterisation/examples/MPFSVOt.0.1.xml`.

Appendix C: Characterisation of various dataset properties

Axes /Properties	Spatial	Temporal	Spectral	Flux /Observable
Coverage				
Location	Central position	Mid-time	Central wavelength	Average flux
Bounds	RA, Dec [min,max] or Bounding box [center, size]	Start/stop time	Wavelength [min, max]	Saturation, Limiting flux
Support	FOV as array of polygons	Time intervals (array)	Wavelength intervals (array)	
Sensitivity	Quantum efficiency (x,y)		Transmission curve (λ)	Function property e.g. Linearity
Filling factor	Effective/ Total area	Live time fraction		
Resolution	PSF (x,y) or its FWHM	Duration per image	Band FWHM	FluxSNR (stat error)
Sampling Precision	Pixel scale (x,y)	Duration per image	Band FWHM	(1 ADU equivalent = Quantization)

Table 2: *Property versus Axis description of metadata describing a 2D optical image. This represents a single integration or indivisible stack of exposures, taken in a single broad-band filter, so the spectral resolution is the same as the filter FWHM.*

Properties/Axes	SPATIAL	TEMPORAL	SPECTRAL	FLUX/OBSERVABLE
Coverage				
Location	Central position	Mid-Time	Central wavelength	Average flux
Bounds	Slit RA, Dec [min, max] or Bounding box	Start/stop time	Wavelength [min, max]	Saturation, Limiting flux
Support	Slit as accurate array of polygons	Time intervals (array)	Wavelength intervals (array)	Lowest and highest value
Sensitivity	Response (x,y) along slit		Quantum eff (λ)	Function property e.g. Linearity
Filling factor	Effective/ Total area	Live time fraction		
Resolution	Slit area	Min. extractable interval	LSF or its FWHM	FluxSNR (Stat error)
Sampling Precision	Slit area	Min. extractable interval	Pixel scale in λ	(1 ADU equivalent Quantization)

Table 3: *Property versus Axis description of metadata describing a 1D-Spectrum.*

Properties/Axes	Spatial	Temporal	Spectral	Flux/Observable
Coverage				
Location	Central position	Mid-Time	Central wavelength(all spectra)	Average flux
Bounds	Field RA, Dec [min, max]	Start/stop time	Wavelength [min,max] (all spectra)	Saturation, Limiting flux
Support	Union of fiber footprints on the sky	Time intervals (array)	Disjoint wavelength intervals	Lowest and highest value
Sensitivity	Response(x,y) along the slit		Quantum eff. (λ)	Function property e.g. Linearity
Filling factor	Effective/ Total area	Live time fraction		
Resolution	PSF (x,y) or its FWHM	Min. extractable interval	LSF or its FWHM	Flux SNR (stat error)
Sampling Precision	Pixel scale (x,y)	Min. extractable interval	Pixel scale in λ Quantization)	(1 ADU equivalent

Table 4: *Property versus Axis description of metadata describing 3D IFU data. These are taken using a mask of multiple slits or fibres each focusing a separate spectrum onto a single detector array. The Support comprises multiple discrete intervals in all dimensions, into which data products could be decomposed. The spatial resolution is determined by the telescope aperture (and the seeing) which spreads the incident radiation over several CCD pixels; the resolution and pixel scales impose different constraints on downstream data analysis.*

Properties/Axes	Spatial	Temporal	Spectral	Flux/Observable
Coverage				
Location	Central position	Mid- Time	Central Frequency	Average flux
Bounds	RA,Dec [min,max] or Bounding box [center, size]	Start/stop time	Frequency [min,max]	Saturation, rms noise
Support	Primary beam FWHM (or mosaic polygons)	Time intervals (array)	Frequencies (array)	Peak, 3σ rms
Sensitivity	Smearing limits/ functions (of integ. time/ chan. width)	Gain- elevation	Bandpass function(s) or FWHM(s)	Dynamic range
Filling factor	Fraction of mosaic filled	Live time fraction	Fraction above FWHM sensitivity	
Resolution	Spatial scales (max and min of BMaj, BMin, BPA)	Min. imageable duration	FWHM of Hanning smoothing	RMS noise
Sampling Precision	Pixel scales [min, max]	Integration time	Channel width	

Table 5: *Property versus Axis description of metadata describing a radio image service, potentially mosaiced. The Max. and Min. spatial resolutions arise from the shortest and longest baselines present; any intermediate value may be selected when an image is extracted from visibility data. The spectral resolution may be coarsened by smoothing to mimimise artefacts.*

Properties/Axes	Spatial	Temporal	Spectral	Flux/Observable
Coverage				
Location	Central position (0, 0)	Mid- Time (0)	Central Frequency	Average flux
Bounds	Bounding box [center, size]	Relative start/stop time	Frequency [min,max]	Saturation, rms noise
Support	FOV as array of polygons	Time interval	Frequencies	
Sensitivity	Quantum efficiency (x, y)		Transmission curve	Detector linearity
Filling factor	Effective/ Total area	(100%)		
Resolution	PSF FWHM	Duration	Band FWHM	Noise error
Sampling Precision	Pixel scales [x, y]	Duration	Band FWHM	Quantization

Table 6: *Property versus Axis description of metadata describing a simulated CCD observation in a single band. The spatial coordinates may be expressed in (x, y) independent of celestial position.*

Appendix D: Updates of this document

http://alinda.u-strasbg.fr/Model/Characterisation/characterisationDraftUpdate.pdf includes a previous revision history (very soon).