International
Virtual
Observatory
Alliance

Characterisation DM: Complements and new features
Observation quality and variability - complex datasets
Version 2.0

IVOA Working Draft September 6, 2012

This version:
http://www.ivoa.net/Documents/WD/Char2.0/Char2.0-20101207.html

Latest version:
http://www.ivoa.net/Documents/latest/Char2.0.html

Previous versions:
http://www.ivoa.net/Documents/WD/Char2.0/Char2.0-20100711.html

Editor(s):
François Bonnarel

Authors:
François Bonnarel, Igor Chilingarian, Mireille Louys, etc ..

Abstract

The Astronomical Dataset Characterization Data Model (CharDM) defines and organizes all the metadata necessary to describe how a dataset occupies multidimensional physical space, quantitatively and, where relevant, qualitatively, in such a way that they become interoperable. We present here a new version of the characterisation data model, with description of data interpretation aids from variablity of observations, together with new representations

for polarisatiioxin and redshift axes. Complex datasets are also tackled.

Status of this Document

This document is an IVOA data Model Working draft.

Acknowledgments

François Bonnarel and Mireille Louys thank EURO VO for funding of participation to conferences.

1 Introduction

Data Models in the VO aim to define the common elements of astronomical data and metadata collections and to provide a framework for describing their relationships so these become inter operable in a transparent manner.

The Astronomical Dataset Characterization Data Model (CharDM, [1]) defines and organizes all the metadata necessary to describe how a dataset occupies multidimensional physical 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. The Characterization Data Model is an abstraction which can be used to derive a structured description of data and thus facilitate its discovery and scientific interpretation (see figure 1).

Various other VO Data Models are making reference to the CharDM, in particular, ObsCoreDM (Observation Core DM), ObsProvDM (Observation and Provenance DM), SpectrumDM, SSLDM (Simple Spectral Line DM), PhotDM (Photometry DM).

As with most of the VO Data Models, CharDM makes use of STC, Utypes, Units and UCDs. CharDM can be serialized with a VOTable

CharDM v1.13 became an IVOA Recommended standard in March 2007. The history of Characterisation and data model development can be found in Appendix A. Use cases for data analysis (section 2) have been considered and emphasize the need to detail the definition of the place holder for variation maps, as well as other specific features.

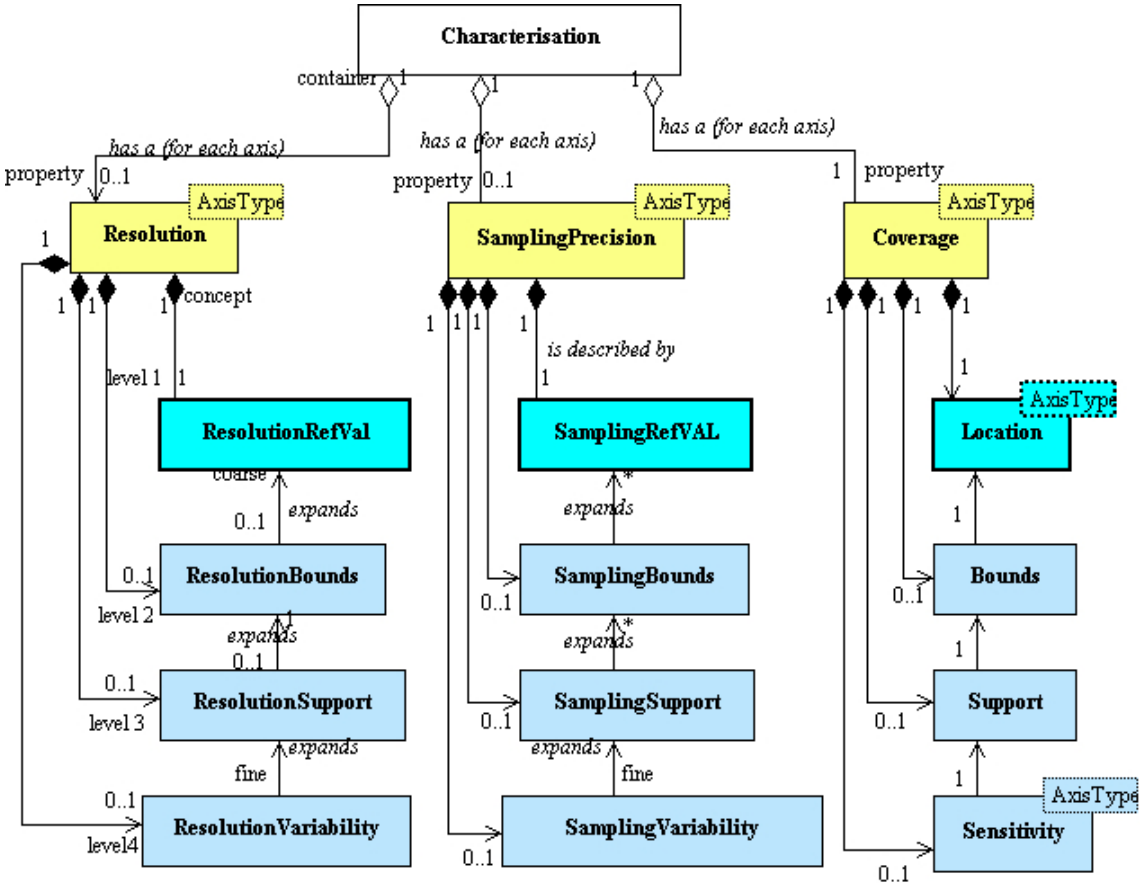

Figure 1: Characterisation DM version 1.13 UML schema

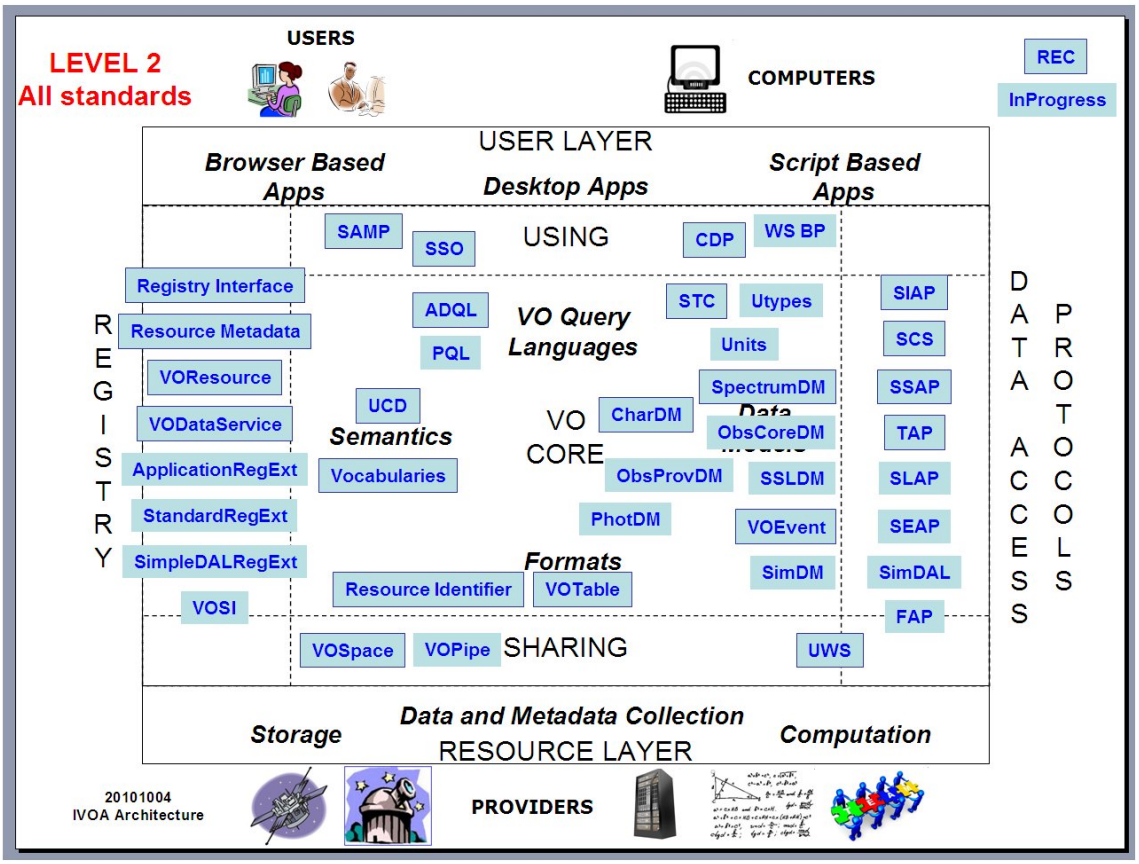

Figure 2: Characterisation in the global VO architecture

- version 2 proposes a detailled description of Level 4.
- version 2 includes possibilities to describe "peculiar" axes such as polarisation axis , and makes it possible to attach an instrumental response function (point-spread function / PSF in case of 2D-images or line-spread function / LSF in case of 1d-spectra)
- it provides mechanisms for handling nested metadata required to deal with composed datasets, i.e. datasets containing several [sometimes rather independent] segments

1.1 Architecture

All this is illustrated by figure 2. Characterisation DM is part of ObsDm, and Spectrum DM, makes use of STC classes... It is extensivly used by DAL protocols such as SIA, SSA and ObsCore.

1.2 Correlated changes on XML schema and utype list

The modification in the model implies related changes in the XML schema and utype list.

It was a good opportunity to clean up the XML schema and utype list. IVOA has established new rules of writing XML schemata clarifying the relationship between the UML data model descriptions and their XML schema serialisations. This requires specific corrections in the previous Characterisation DM schema.

Eventually, usage of Characterisation (and Observation) DM utypes leads to simplification in the model attributes and attribute names. However utype changes will occur in very specific levels. This is required for optimal backward compatibility of the model. Current usage of the Char DM utypes, as in spectrum DM and SSA protocol is focused mostly on a small subset of Data model items for which utype names are stable and won't be simplified.

1.3 Organisation of the document

The document is composed as following: Section 2 describe science use cases motivating version 2. Section 3 describes new features and changes of the model. Section 4 is presenting the new xml schema. Appendix B is giving the detailed description of the reusable Access package. Appendix C gives the list of characterisation utypes.

2 Science use-cases for Characterisation v.2

One of the principal improvements of Char-2 over the first version is the detailed description of the most advanced 4th level of metadata.

2.1 Crowded-field photometry using multiple PSF-fitting.

An important use-case of the 4th level of the Characterisation DM metadata is connected to imaging data. In so-called crowded fields (e.g. open or globular star cluster, dense regions in the Galactic plane or resolved nearby galaxies) stars are so close to each other in the image plane that aperture photometry (e.g. as performed by SExtractor) does not provide satisfactory results because more than one source often falls inside an aperture of the size enclosing major fraction of the point source flux.

In this case a different approach is used including two steps. At first, a source detection algorithm (usually as simple as the threshold detection of a convolved image) is used to find approximate position of stars and get more

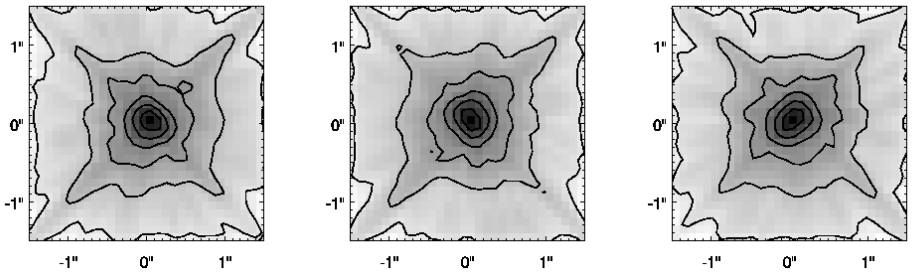

Figure 3: Example of the TinyTim generated PSF of the Hubble Space Telescope Wide Field Planetary Camera-2 in different positions inside the field of view.

accurate positions using weighted centroids (or other similar algorithms). Then, on the second step, these stars are fitted simultaneously using multiple point-spread-functions (PSF) at the positions found at the first step by varying only their amplitudes and sometimes also allowing to adjust the coordinates, although this significantly decreases the stability of the technique. Finally, the best-fitting amplitudes obtained from this multiple PSF fitting are used as photometric measurements. If the fitted PSF is wrong, all the photometric measurements will be biased.

Normally, the PSF shape is determined before processing of the crowded-field photometry by measuring shapes of relatively bright stars located in different positions inside a field-of-view (FoV) using not too crowded calibration fields. In all imagers, the PSF shape and average width changes across the FoV. In some instruments, these changes reach a factor of two or more in the PSF width across the field. Therefore, it is important to precisely take into account the PSF variations in order to avoid systematic differences in the photometry of sources located in different FoV parts.

This description of the PSF variations across the field of view can be achieved using the 4th level of the Characterisation DM. Several ways of the PSF representation can be foreseen:

- A two dimensional array containing a PSF model can be attached to every pixel of the image or to some larger image regions where the PSF variations can be neglected. This is the best way of representation in case of complex PSF shape and is model-independent, however, the volume of characterisation metadata in this case will exceed the volume of the real data by a huge factor (figure ...).
- Another possibility is to adopt some model of the PSF, e.g. a two-dimensional Gaussian with a free positional angle, if this represent well

the real situation for a given instrument. In this case, the coefficients of the representation (σ_a, σ_b, and θ in a given example) can be attached to every pixel, while at the higher level description the actual representation will be presented using, e.g. MathML. In this case, the volume of the characterisation metadata is considerably lower than in the previous example, but some systematic photometric errors may arise if the real PSF is very different from the adopted model. *We can predefine some widely used PSF parametrizations like 2D-Gaussian, top hat profile, mexican hat profile etc.*

- The third possibility is to use again the model PSF (as in the previous case), but to take advantage of the fact that PSF usually changes very smoothly across the FoV. Therefore, it should be possible to approximate the behaviour of the coefficients used to represent the PSF (σ_a, σ_b, and θ in case of a two-dimensional Gaussian) across the FoV using some smooth fitting functions, e.g. two-dimensional polynomials or splines (These functions themselves can be described using MathML, which will at the end decrease the volume of the serialisation of the 4th level characterisation metadata to the size comparable to the 2nd and 3rd levels, see examples in section 4). Of course, it may create systematic photometric errors if the PSF exhibits abrupt changes across the FoV and, therefore, is badly approximated by the selected fitting functions. However, in most real cases, this approach should work very well. In case of complex datasets, like HST WFPC2 mosaics, the best solution will be to use separate 4th level characterisation metadata for every quadrant, and then use the composition mechanism proposed to store descriptions of sub-datasets.
- The fourth alternative would be to use an external service returning the PSF model (i.e. TINYTIM for HST images). Then, only a reference to the service with description of its input parameters is required.

2.2 Full spectral fitting algorithms.

Another use-case of the 4th level metadata deals with spectra. There is a family of techniques referred as "full spectral fitting", when a whole spectrum is fitted by some models pixel by pixel in order to obtain some parameters. Among examples of such techniques implemented as publicly-available software packages or VO services are: penalized pixel fitting [2] and NBURSTS full spectral fitting [3, 4]. They are used to extract from absorption-line spectra of galaxies internal kinematics, e.g. Gauss-Hermite parametrization of the line-of-sight velocity distribution -hereafter LOSVD- and a parametrized star formation history (only NBURSTS) represented by several star bursts

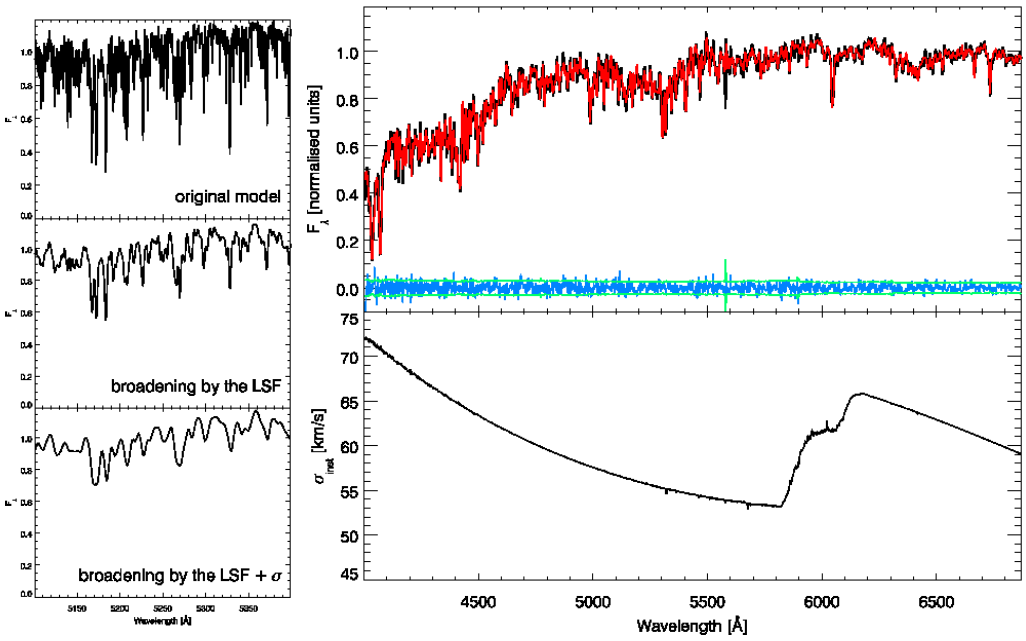

Figure 4: Effects of the PSF and intrinsic stellar velocity dispersion on the absorption line broadening in a galaxy spectrum (3 panels on the left). Example of the full spectral fitting (top right) of an SDSS early-type galaxy spectrum (spectrum is shown in black, its best-fitting template in red, the residuals in blue and 1 σ flux uncertainties in green) and the variation of the spectrograph's instrumental response (bottom right).

events in the galaxy lifetime, each of them characterised by only its age and metallicity, i.e. simple stellar populations.

These techniques use the LOSVD to convolve the models in order to estimate the broadening of spectral lines in galactic spectra due to internal velocity dispersion of stars. However, then the models used in the fitting technique at first have to be corrected for the intrinsic broadening of spectral lines caused by the optics of the spectrograph called a line-spread-function (LSF). As PSF in case of images, LSF may vary along the wavelength range covered by the spectrum. These variations can be very important, especially if an observed spectrum contains several segments obtained from different physical units working at different wavelengths. There is an algorithm which allows to convolve the high-resolution model which is then used to fit a galaxy spectrum with a kernel (LSF) variable along the wavelength range.

In practice, the LSF variations can be estimated from the spectra of twilight sky which are essentially Solar spectra by fitting a high-resolution Solar

spectrum (not broadened) against them in several small segments covering the wavelength range.

To store these variations one can use exactly the same approach as for PSF in case of 2D-images (see above). The only large collection of spectra available in the VO which provides the LSF variation information is SDSS. The LSF is represented as a purely Gaussian function assuming no systematic radial velocity offset (i.e. a Gaussian centred at 0). The value of the Gaussian dispersion (σ) is provided for every pixel and is stored as a vector of the same length as the spectrum itself in the 6th FITS extension in the 1D spectrum files distributed by the SDSS archive. Since SDSS spectra contain two segments, blue and red which are obtained in different units of the spectrograph, there is a sharp break in the behaviour of the LSF parametrization at $\lambda \sim 5900$ Å. *figure 4*.

2.3 Analysing a dataset with complex provenance

Another important use-case for such advanced descriptions developped here is the metadata for complex datasets. By complex datasets we assume datasets comprising several "traditional" sub-blocks. Examples to illustrate this use-case are:

- Data produced by wide-field mosaic images such as CFHT MegaPrime/MegaCam or WFI at ESO/MPI 2.2m telescope working in the optical domain or CFHT WIRCAM or UKIRT WFCAM working in the Near Infrared Domain.
  HST WFPC2, WFPC3 and ACS are other examples of this type of instruments, although strictly speaking they are not wide-field.
  The wide-field imagers have CCD mosaics consisting of several independent CCD chips. Data produced by each such chip can be characterised as a simple CCD image. However, for a mosaic dataset several subtleties arise. Usually wide-field mosaics contain gaps between individual chips and spatial dithering during observations is used to fill them (i.e. shifts of individual exposures with respect to each other). However, each CCD chip has its own characteristics like sensitivity curve and read-out noise, therefore the calibration across the field of view is sometimes non-trivial. Because of the spatial dithering, some regions (close to CCD gaps) in the final processed image may contain signal originating from several individual CCD chips and co-added (after re-normalisation). This will affect the photometric measurements made later in those regions.The scope for Characterisation DM v.2 is to to be able to characterise the end-products of the data reduction

precisely. How the process of such a composition is performed will be sketched out by the Provenance class of the general Observation DM.

- Echelle spectra. Such spectra contain different echelle orders (i.e. little 1D spectra) which are merged together after final data reduction. They may or may not overlap. There are similar difficulties for the data description as for the mosaic CCD images with the only simplification that usually all CCD orders reside on the same CCD chip, so there is no problems with combination of data obtained from the detectors with different intrinsic characteristics. In some cases (e.g. VLT X-Shooter) the echelle orders are projected onto several CCD chips in a mosaic sensitive to different wavelength ranges (i.e. optical and NIR).
- Multi-unit spectrographs. Examples: VIMOS at VLT, MUSE at VLT (has not been delivered yet to the telescope). These instruments actually are sets of independent small spectrographs (units), and the field-of-view of the telescope is splitted between them. In this case beside different detectors like in the case of mosaic imagers, we will also get different dispersers (e.g. grisms) having slightly different characteristics (spectral resolution, blazing angle, sensitivity) and different optical tracks with slightly different distortions. At the end, for example in case of VIMOS-IFU, if the spatial dithering is applied in order to work around dead fibres in the IFU bundle, the same part of the sky (and astronomical object on it) may be taken using different spectral units.

The use-case is to describe the properties of the fully-reduced combined datasets from such systems. Of course, it will be simpler to describe individual segments of Echelle or individual spectra coming from different spectral units, and to let the combination of dithered datasets on the end-user. However, this procedure may be so complex (especially in case of mosaic wide-field imagers), that only specialists of a given instrument will be able to combine *even fully reduced individual observing blocks*, therefore if the dithered observations are not combined, they will have very little interest for a larger user community.

2.4 polarization data

Polarization of light is nearly as significant of physical processes as its spectral distribution. It can be combined with imaging or spectroscopy. Polarimetric observations of stars and AGNs or quasars are relevent for studying the geometry of the atmosphere or outer layers of these objects.

The scattering of light by dust is also a great source of polarized emission.

Polarimetry allows to discover the shape and geometry of the dust grains.

Probably the most common usage of polarimetry is the study of magnetic fields in a wide range of categories of objects. Zeemann effect split spectral lines in two different polarization states (and energy level).

Spectropolarimetry allows to go further the usually low resolution of standard polarimetric imaging. Different lines may have different strengths according to the subregion of the source they are coming from. Spectropolarimetry help to partition the emission in several subsets.

3 Modeling

Modifications from version 1 of the model occur:

- on the general structure (composed data),
- at low level of characterisation where resolution and coverage are extended
- and mainly at level 4 which is extensivly developped.

All this is illustrated by figure 5

3.1 Extension to low level of Characterisation

3.1.1 PSF or resolution arrays

Resolution is considered as a property of an observation along an axis in the Characterization data model. It can be described with an increasing level of precision, starting from a single value (Instrumental function profile FWHM), up to the PSF full profile function itself. Intermediary level of description can be given such as ellipsoidal profiles widths (FWHM). The Characterization model has encoded the reference value content of the Resolution Property as an stc Resolution element, which encompasses circular and ellipsoidal FWHM features. In this version we add the PSF, considered as an array of values as a possible option. The matrix can be given either in an external file (the access of which the model should describe - 3.3.2 describes the proposed solution), or directly in the serialisation. The generic assumption is that the grid of pixel of this local array is aligned on the data axes using the same sampling as the data themselves.

3.1.2 Coverage peculiar case: axis spanned on a discrete set of values

This was elaborated to take into account the polarization case ([5]) but can be easily generalised to other discretly spanned axes such as spectral bands.

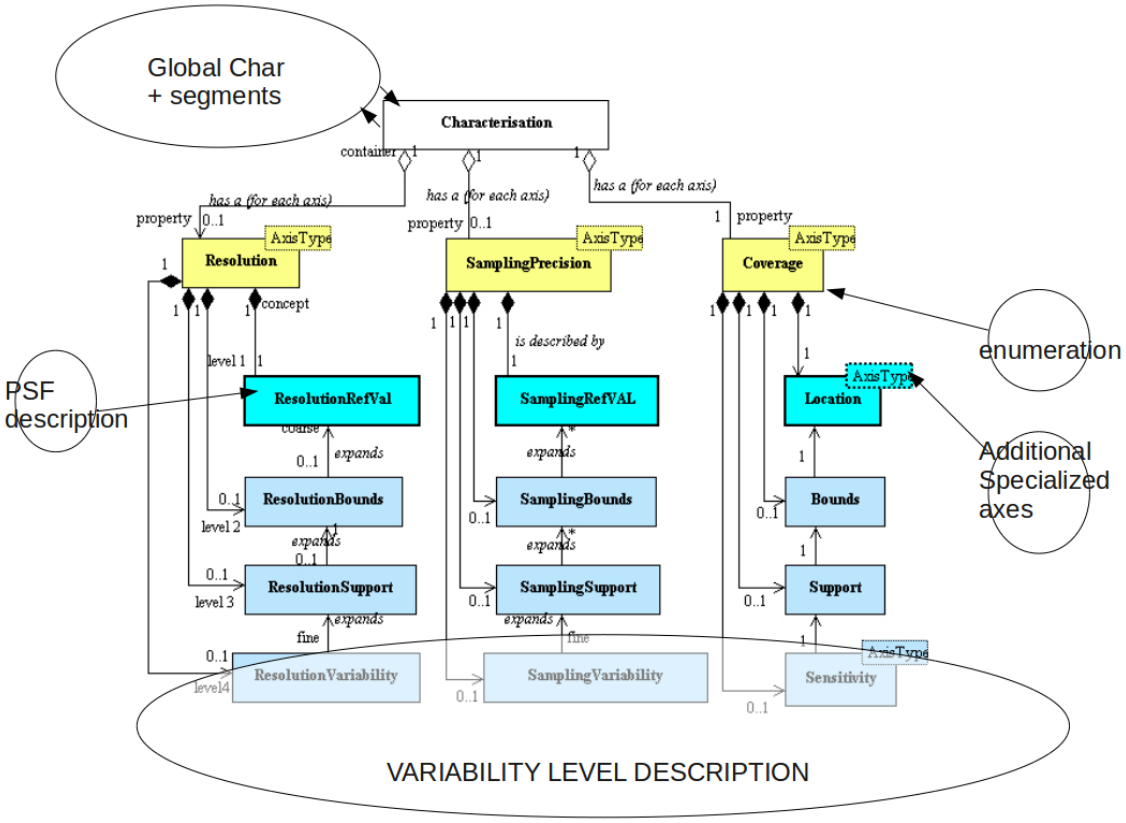

Figure 5: Modifications of the characterisation model with version 2: where does it impact the structure?

The various polarization states (such as Stokes parameters) can be described as different values on a characterisation axis. This axis is peculiar in that it always consists of a discrete set of literal values (XX,XY,YY.. or I,Q,U,V, etc...). This is partly analogous to a spectral axis containing several planes, spaced irregularly and of different spectral widths (often the case, for example, when preparing SEDs) which can be expressed as a set of labels of observing bands. *Full Characterisation of the polarization axis is thus accomplished by listing the polarization states present in the dataset in the stateList attribute of the polarizationAxis.*

On the flux axis, the different polarization states can have different detection limits, and ranges of values (some time the spatial and spectral axis properties, such as resolution, may also differ for each polarization state). This makes it difficult to characterise the flux axis in detail. However It may be sufficient to set outer bounds at the coarsest level (usually taking the maximum total intensity as the upper bound and its minimum as the lower bound). In order to describe in detail each polarization state it may be usefull to duplicate the observable axis (and even spatial or spectral axes in some cases) for each state. This leads to the concept of the characterization of composed data described in next subsection.

3.2 Composed data

The general scheme of characterisation DM version 1.0 allows to express resolution, coverage, sampling (called "Properties" in the characterisation context) on all physical Axes and expands on four levels in a quite efficient way. This simple scheme works for a very wide variety of datasets with the assumption of independant axes. For more complex cases however it will not suffice . Imagine for example an IFU (see above) with spectral range varying with the position. How can we describe the spectral support (Union of intervals where the data are significant, see [1] for definition) of such a dataset?

The underlying problem is the dependence or independence of the axes and properties. *In principle it could be possible to tackle these interdependances by using combined axes. A "combined axis" is defined as an axis which integrates dimension coming from several standard axes (to take into account strong coupling between them). For example, an IFU with complex support and spectro/spatial dependancies (see Chilingaryan for details) could have a support described as a polyedron in the 3D spectro/spatial combined axis.*

However a simpler strategy to solve it for most use cases is to look for subparts or segments of the observation where the independence can be con-

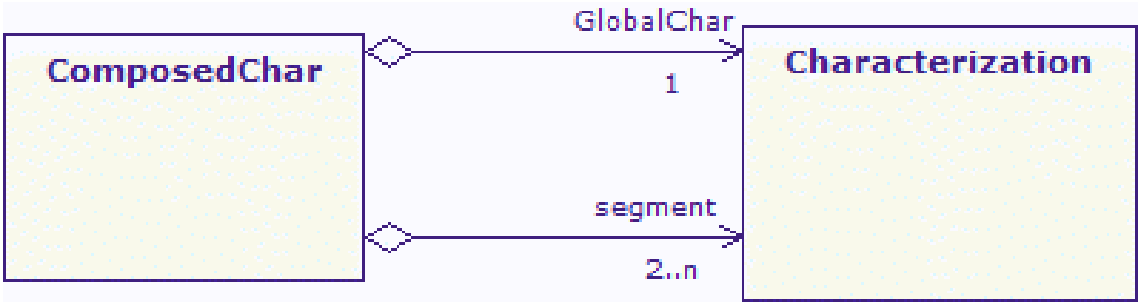

Figure 6: global and segment characterisation UML diagram.

sidered as a good approximation... In the case of an image obtained with a CCD mosaic, it makes sense to consider the spatial resolution to be different for each CCD chip used to record the observation. While the total range for the data set (i.e. the resolution bounds) can be given for the whole, it would be more significant to associate a given reference value to each CCD, i.e. per support ([1]) segment.

In the case of polarimetry, the range in fluxes and the spatial resolution generally depends on the polarimetric state we are considering. The coverage along the Flux axis can be different for each polarimetric state as well as the spatial or spectral resolution.

In practice we will consider the observation as a composed one made of the aggregation of several sub observations. It is then possible to characterize the whole observation roughly or each sub observation in more detail.... We introduce two new different roles for the characterization model: globalCharacterization and segmentCharacterization.

All but one of them play the role of characterisation for a given segment, while the latter plays the role of a global characterisation of the whole dataset. Generally the global characterisation will give a rough level 1 or 2 description while segment characterisation will gather a much finer description (level 3 or 4)...(see figure 6) The globalCharacterizationn of an observation could be computed from all the segmentChacterization of its sub-observations characterisations.

Table 1 illustrates the example of an HST WFPC2 image... Table 2 illustrates the use of a VLA NVSS polarization cube.
(`http://www.cv.nrao.edu/cgi-bin/postage.pl?Equinox=J2000&PolType=IQU&RA=10+10+10`

Table 1: Global characterisation and segments for an HST image

flavor	SpatialAxis.Coverage.Location.refval	SpatialAxis.Coverage.bounds	resolution.bounds	
global	308.633+60.146	(308.645+60.173 308.621+60.116)	1.0 0.2	
flavor	SpatialAxis.Coverage.location.refval	SpatialAxis.coverage.bounds	SpatialAxis.coverage.support	SpatialAxis.resolution
segment 1	308.604+60.148	(308.610+60.173 308.633+60.133)	(308.610+60.173 308.633+60.133)	1.0
segment 2	308.635+60.157	(308.645+60162 308.519+60.145)	(308.610+60.173 308.633+60.133)	1.0
segment 3	308.658+60.140	(308.668+60.155 308.656+60.127)	(308.610+60.173 308.633+60.133)	1.0
segment 4	308.631+60.138	(308.633+60.144 308.628+60.131)	(308.610+60.173 308.633+60.133)	0.2

Table 2: Global characterisation and segments for a VLA/NVSS polarized dataset

flavor	SpatialAxis.coverage.location.refval	SpatialAxis.coverage.bounds	FluxAxis.coverage.bounds	PolarizationAxis.list
global	152.54166+10.16944	(152.29+9.92 152.79+10.41)	(-0.00180 0.24638)	StokesI StokesQ StokesU
flavor	FluxAxis.name	FluxAxis.unit	FluxAxis.coverage.bounds	FluxAxis.error
segment 1	StokesI	Jy/beam	(-0.00180 0.24638)	
segment 2	StokesQ	Jy/beam	(-0.00092 0.00096)	
segment 3	StokesU	Jy/beam	(-0.00092 0.00464)	

3.3 Classes at work for level 4

In order to support multiple cases of variability along the axes we have enriched the description of variation maps.

3.3.1 VariationMaps

Here we complete the design by defining the detailed metadata structure in order to cover the various use-cases exposed in section 2. Here we call "map" any quantity describing a property varying along an axis. For coverage it is generally the sensitivity variation along this axis. ("Sensitivity in a receiver is normally defined as the minimum input signal Si required to produce a specified signal-to-noise S/N ratio at the output port of the receiver"). But It can also be any derived quantity expressing the sensitivity, such as an extended "flat field". In the case of resolution, it can be a map of the variations of the FWHM (flat or circular case), as well as a map of two different X and Y FWHM (elliptical case), or even a map of variations of the PSF over the field.

What is required is both a description of what is encoded in the map (is it a varying FWHM, a transmission factor or whatever) and of the implementation. Several implementations may coexist. The content will be identified by a VariationContent attribute in our generic level 4 model.

As far as the actual description is concerned we may consider several possibilities.

(a) give an array of values, for example as an external file
(b) give a description of various moments characterizing the variation along the considered axis.

(Our variation maps are not made of direct measurements. They are actually a priori *response maps stemming from instrumental calibration parameters. In other words they are distribution functions. For example the coverage spatial sensitivity is the distribution function of the potentially detected events for a uniform input function (along the considered axis). A distribution can be described with the set of its statistical moments. The larger the number of coefficients, the more precise the approximation will be.)*

(c) give a functional description of the variation map.

Figure 7 summarizes the UML class diagram for Level 4 classes. Such a package can be hooked in the general characterisation model either in Sensitivity, in ResolutionMap, or in SamplingMap.

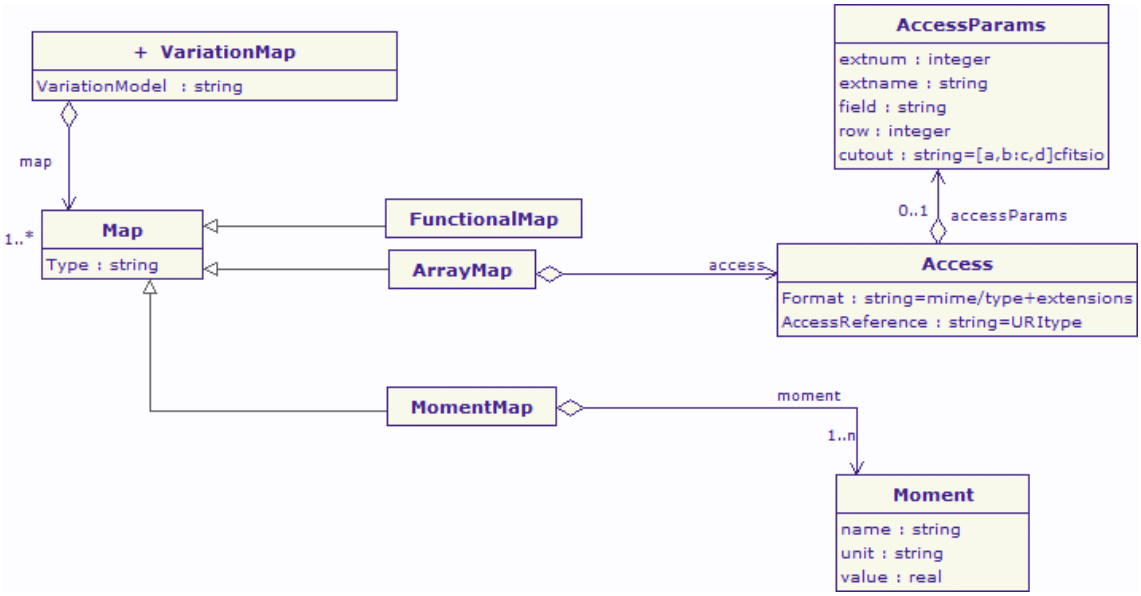

Figure 7: Detailed structure for variation Maps

```
-----------------------------------------------------
Access.format:table/fits
Access.reference:
http://das.sdss.org/spectro/1d_26/1615/spSpec-53166-1615-040.fit
Acces.AccessParams.extnum:6
Access.AccessParams.Field:DISPERSION
Access.AccessParams.unit:km/s
```

Table 3: Access attributes pointing to a table column in 6th extension of a MEF file

3.3.1.1 Array of values:

Basically this can be given either by pointing to a file or by including a matrix attribute in the model.

A Model attribute will give the data model used for the array (which maybe an IVOA data model like spectrum, or a proprietary one) The pointer to an external file will be defined using the Access package described below. For example a spectral resolution variation can be contained in one of the table extension of a FITS/extension file, as described in table 3 for SDSS filter response.

3.3.1.2 Moment description of a distribution:

This is given as a set of structures describing statistical moments. Each of those is giving the name (or range) of the moment, its value and the unit used... For instance, the

```
-------------------------------------------
Moment.name:mean
Moment.unit:m
Moment.value:0.5e-7
Moment.name:sigma
Moment.unit:m
Moment.value:0.1e-7
Moment.name:kurtosis
Moment.unit:m
Moment.value:0.01e-7
Moment.name:4
Moment.unit:m
Moment.value:0.0023e-7
```

Table 4: Approximation of a sensitivity map by moments

sensitivity map on spatial axis can be described with good approximation by the centroid position, the sigma, kurtosis and a couple of higher order moments of the actual sensitivity distribution. See example in table 4.

3.3.1.3 Functional description This is the description of the variation map as a function of the position along the axis. It can be expressed as a C-like expression with a set of variables and parameters or using an external mathematical expression modeling language such as MathMl. For example a psf variation function can be expressed by polynomial variations of a bi-Gaussian function parameter. See example in table 5 for aC-like expression. Actually it is assumed that it is expressed in the same grid of pixels as the data themselves... (The World coordinate mapping to the pixels is inferred from the dataset).

3.3.2 Definition of an Access package to describe URL and structured files.

This package is used in Char Version 2.0 each time we want to bound a file to a VariationMap or PSF description and described Appendix B. As it can be reused in other IVOA models (Provenance model, DataLink model). **we let open the question of where this Access package belongs ? It is probably actually not part of Characterisation 2 an it can be upgraded to the status of a litle reusable model in itself...**

```
-------------------------------------------------------------
Map.type:parametric
Map.function:a*exp(-(x-b)**2/c)
Map.function.variable.name:x
Map.function.param.name:a
Map.function.param.value:50.0
Map.function.param.name:b
Map.function.param.value:0.3
Map.function.param.name:c
Map.function.param.value:1.3
```

Table 5: Functional description . C-like expression

3.4 A full example: visibility data (raw)

As mentionned in [6] radio interferometry visibility data can be described with Characterisation data model. Basically visibility measurements are given at a given time, for a source at a given position, for several spectral channels and polarization feeds, and at several "spatial frequences" or equivalently "baselines". The *uv* plane (spatial frequencies) defines a new "flavor" of the spatial axis, identified by the appropriate ucd, where coverage, resolution and sampling are meaningful and defined up to level 2 or 3. The standard flavor of the spatial axis (with "pos" ucd) will give the pointing (level 1) and the "field of view" is actually a sensitivity map (level 4) showing so much variations that level 2 or 3 are difficult to define for this axis. Spectral axis is generally spanned (data cube) and different polarization states may also be present. For complex visibilities, the Observable axis may actually be split into a "Visibility amplitude axis and a "Visibility phase" axis. A third "visibility weight" axis may be added if necessary. The units for the amplitude could be in Jy or absent depending we are dealing with absolute or relative amplitudes.

However the structure of Visibility data files as decribed in [7] maybe very complex because the information necessary to characterize the data may be mixed with a lot of provenance information and also because the data file may gather information from several sources and several FREQUENCY SETUPS. Composed data feature of characterization (see above 3.2) is obviouly needed. Table 6 shows how we could characterize the sample FITS IDI dataset given by the FITS support office:
`http://fits.gsfc.nasa.gov/registry/fitsidi/BL146_1.fits`
the header of which can be seen in table 7.

3.5 specialized axis

Characterisation version 1 defined three "specialized" characterisation axes beside the generic one: spatialAxis, spectralAxis and TimeAxis. A specialized axis forces the value of the label and the reuse of some specific Stc coordinate classes. Experience showed that it is necessary to define a new set of specialized axes. ObservableAxis is an axis which generally shows a functionnal dependancy with respect to at least one of the other axes. FluxAxis is a specialisation of ObservableAxis. PolarizationAxis details the "sampling" in polarizations states for the observation. It is essentially giving the list of polarization states present in the data set. Data with no analyzed polarisation have only a "Stokes I" value and in that case the PolarizationAxis can be ignored. RedshiftAxis is an important axis for datacubes where Doppler variations of specific spectral lines are sampled. It reuses the Redhshift Coordinate and RedshiftInterval of STC in coverage.location.coordinate and coverage.location.bounds . The coordinate system used for the axis has to contain the RedShiftFrame.

4 New XML serialisation of the Characterisation data model

In this section we present a new xml schema for Characterisation encompassing new definition like variation map and AccessParameters (see sec 3). In addition we applied a new set of xml recommendations mentioned in the VOResource Technical specification, applied in the encoding of the Resource Metadata Model into the VOResource xml schema ([8]). We also took the opportunity to reuse UML to XML mapping recommendation at work in the IVOA as sated in the utypes working draft([9]).

4.1 Applying new IVOA DM rules to the build up of Characterisation xml schema

- The public elements in the XML characterisation schema have been suppressed. All elements in an XML document compliant to charDML v2 must of course reuse XML types from the XML chararacterisation schema. But the document is supposed to define its own elements names following these xml types.
- UML to XML transcription uses the roles in the associations in the class diagram to give their names to the elements (e.g. *refval* in Sampling and Resolution)

FITS KEYWORD	utype	value estimation
RA	Characterization.SpatialAxis.Coverage.location.refval	nominal position
DEC	Characterization.SpatialAxis.Coverage.location.refval	nominal position
TTYPE4 = 'DATE '	Characterization.TimeAxis.Location.refval	convert field value in years and add time
TTYPE5 = 'TIME '	Characterization.TimeAxis.Location.refval	convert field value in years and add to DATE
REF-FREQ	Characterization.SpectralAxis.Location.refval	nominal spectral frequency
CHAN-BW	Characterization.SpectralAxis.Sampling	frequency shift between 2 channels
TTYPE13 = 'FLUX '	Characterization.FluxAxis.bounds	extract min , max of the matrix : either globally or per-axis
STK 1	Characterization.PolarizationAxis.stateList	value determines if we have Linear, circular or stokes IQU pol
UU, VV, WW	Characterization.SpatialAxis.Coverage.bounds (name = uv, ucd =uv)	estimate the uvw ranges from values there
		need calibration to convert baseline units into spatial frequencies

Table 6: Global characterisation and segments for a VLA/NVSS polarized dataset

```
XTENSION= 'BINTABLE'           / FITS Binary Table Extension
BITPIX  =                    8 /
NAXIS   =                    2 /
NAXIS1  =                 1136 /
NAXIS2  =                96843 /
PCOUNT  =                    0 /
GCOUNT  =                    1 /
TFIELDS =                   13 /
EXTNAME = 'UV_DATA '           /
EXTVER  =                    1 /
TTYPE1  = 'UU-L    '           / u
TFORM1  = '1E      '           /
TUNIT1  = 'SECONDS '           /
TTYPE2  = 'VV-L    '           / v
TFORM2  = '1E      '           /
TUNIT2  = 'SECONDS '           /
TTYPE3  = 'WW-L    '           / w
TFORM3  = '1E      '           /
TUNIT3  = 'SECONDS '           /
TTYPE4  = 'DATE    '           / Julian day at 0 hr current day
TFORM4  = '1D      '           /
TUNIT4  = 'DAYS    '           /
TTYPE5  = 'TIME    '           / IAT time
TFORM5  = '1D      '           /
TUNIT5  = 'DAYS    '           /
TTYPE6  = 'BASELINE'           / baseline: ant1*256 + ant2
TFORM6  = '1J      '           /
TTYPE7  = 'FILTER  '           / filter id number
TFORM7  = '1J      '           /
TTYPE8  = 'SOURCE  '           / source id number from source tbl
TFORM8  = '1J      '           /
TTYPE9  = 'FREQID  '           / freq id number from frequency tbl
TFORM9  = '1J      '           /
TTYPE10 = 'INTTIM  '           / time span of datum (seconds)
TFORM10 = '1E      '           /
TTYPE11 = 'WEIGHT  '           / weights proportional to time
TFORM11 = '16E     '           /
TTYPE12 = 'GATEID  '           / gate id from gate model table
TFORM12 = '0J      '           /
TTYPE13 = 'FLUX    '           / data matrix
TFORM13 = '256E    '           /
TUNIT13 = 'UNCALIB '           /
NMATRIX =                    1 /
DATE-OBS= '2007-08-23'         /
TELESCOP= 'VLBA    '           /
OBSERVER= 'GOOFY   '           /
```

```
OBSCODE = 'BL146   '           /
RDATE   = '2007-08-23'         /
NO_STKD =                    4 /
STK_1   =                   -1 /
NO_BAND =                    4 /
NO_CHAN =                    8 /
REF_FREQ=   8.40549000000000000E+09 /
CHAN_BW =   1.00000000000000000E+06 /
REF_PIXL=   5.31250000000000000E-01 /
TABREV  =                    2 / ARRAY changed to FILTER
VIS_SCAL=   1.08991348743438721E+00 /
SORT    = 'T*      '           /
MAXIS   =                    6 /
MAXIS1  =                    2 /
CTYPE1  = 'COMPLEX '           /
CDELT1  =   1.00000000000000000E+00 /
CRPIX1  =   1.00000000000000000E+00 /
CRVAL1  =   1.00000000000000000E+00 /
MAXIS2  =                    4 /
CTYPE2  = 'STOKES  '           /
CDELT2  =  -1.00000000000000000E+00 /
CRPIX2  =   1.00000000000000000E+00 /
CRVAL2  =  -1.00000000000000000E+00 /
MAXIS3  =                    8 /
CTYPE3  = 'FREQ    '           /
CDELT3  =   1.00000000000000000E+06 /
CRPIX3  =   5.31250000000000000E-01 /
CRVAL3  =   8.40549000000000000E+09 /
MAXIS4  =                    4 /
CTYPE4  = 'BAND    '           /
CDELT4  =   1.00000000000000000E+00 /
CRPIX4  =   1.00000000000000000E+00 /
CRVAL4  =   1.00000000000000000E+00 /
MAXIS5  =                    1 /
CTYPE5  = 'RA      '           /
CDELT5  =   0.00000000000000000E+00 /
CRPIX5  =   1.00000000000000000E+00 /
CRVAL5  =   0.00000000000000000E+00 /
MAXIS6  =                    1 /
CTYPE6  = 'DEC     '           /
CDELT6  =   0.00000000000000000E+00 /
CRPIX6  =   1.00000000000000000E+00 /
CRVAL6  =   0.00000000000000000E+00 /
TMATX11 =                    T /
END
```

Table 7: FITS header of UV data table extracted from IDI file

- In the case of polymorphism (for example the dimension dependant structure of spatial position, resolution, sampling) the substitution groups have been eliminated and replaced by the definition of xml extensions of a basic type. Elements defined in the schema by the basic type can be easily replaced by using the xsi:type attribute.
- Some of the names have been shortened: for example, SamplingPrecision has been replaced by Sampling.

4.2 Description of the new features of the model

- For ComposedChar two new restriction of Charectersation type have been defined. One is GlobalChar the other one is segment.

- Characterization axis includes now an optional stateList element for PolarizationAxis or other special axes definition.
- A new type of element, VariationMap, which is a full new xml hierarchy, is defined and can be included at level 4 as shown in Figure 9 ....
- Several restrictions of CharacterizationAxis have been defined fort specialized axes

Appendix

Overview on IVOA data modelling effort

Modeling of observational metadata has been a long term activity in the IVOA since it was created in 2002. Various modelling efforts like Resource Metadata, Space-Time-Coordinate metadata (STC), Spectrum data model [10], and Characterisation data model [1], have been recommended and are currently used in IVOA services and applications. Historically, models and protocols have been developed in parallel and first focused on simple data types and simple protocols accordingly. However the guide line in the DM WG was to foster full interoperability by covering the full chain of actions a user might want to do for his/her science: data-discovery, data retrieval, data analysis. This work comes now to a more mature state where we need to homogenise the various approaches in order to discover/retrieve/analyse all kinds of observation data products. Although there has been early great successes in the use of some of the data models (Resource metadata, Spectrum with SSA) the general approach described above had long this drawback that it ended up as relatively large data models that people felt difficult to implement and use. This situation was also reinforced by more technical problems:

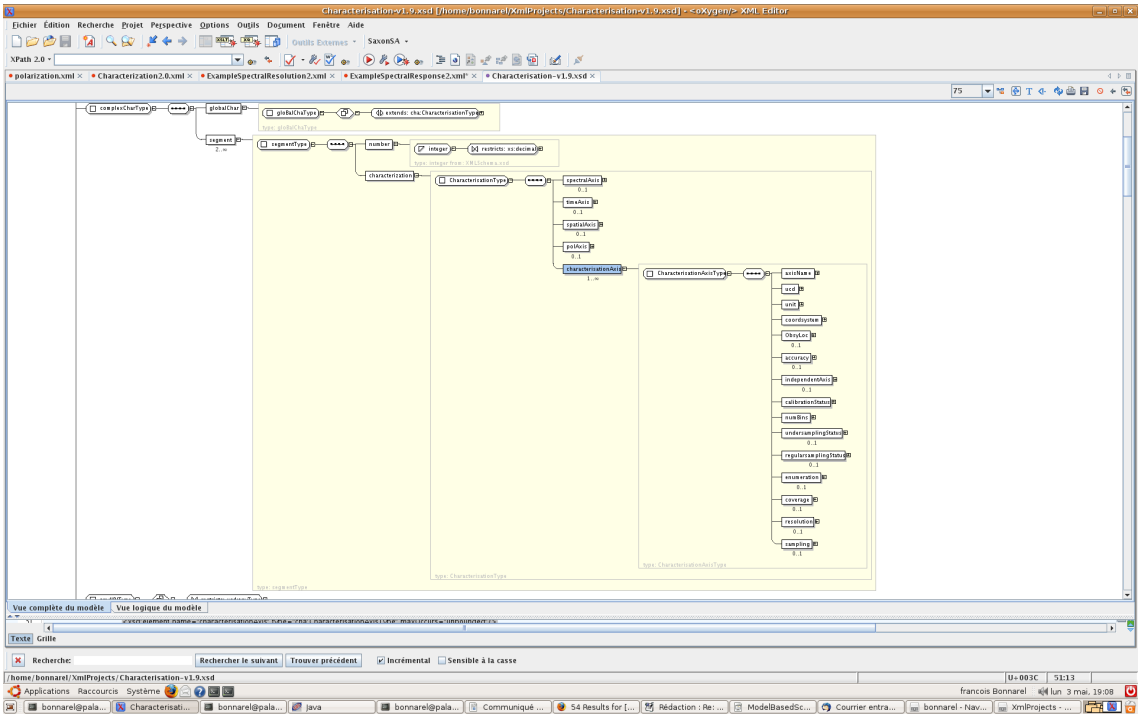

Figure 8: Structure of the ComposedChar class.

serialisation (pure XML or Utypes), and protocols for metadata access were not always available for practical implementation of these data models. A contrario that's probably also why Resource metadata and Spectrum had been implemented by data centres. Their strong linkage with Registry and SSA protocols explained somewhat their early success. In case of Observation and Characterisation data model one obvious family of use cases had long been data discovery. Current effort about OBsTAP and SIAV2 go on along these lines, after the success of SSA.

In the mean time first attempts have been made to use data models in the context of data analysis applications (SED data plugin, [**?**], GALMER [11]) and this experience encountered some limitations from the lack of development of the model itself

The context and history of characterisation metadata modelling

The Observation data model project appeared at the first Data Model forum held at the May 2003 IVOA meeting in Cambridge,UK. Rapidly some main concepts appeared to be necessary to organise the metadata: dataset

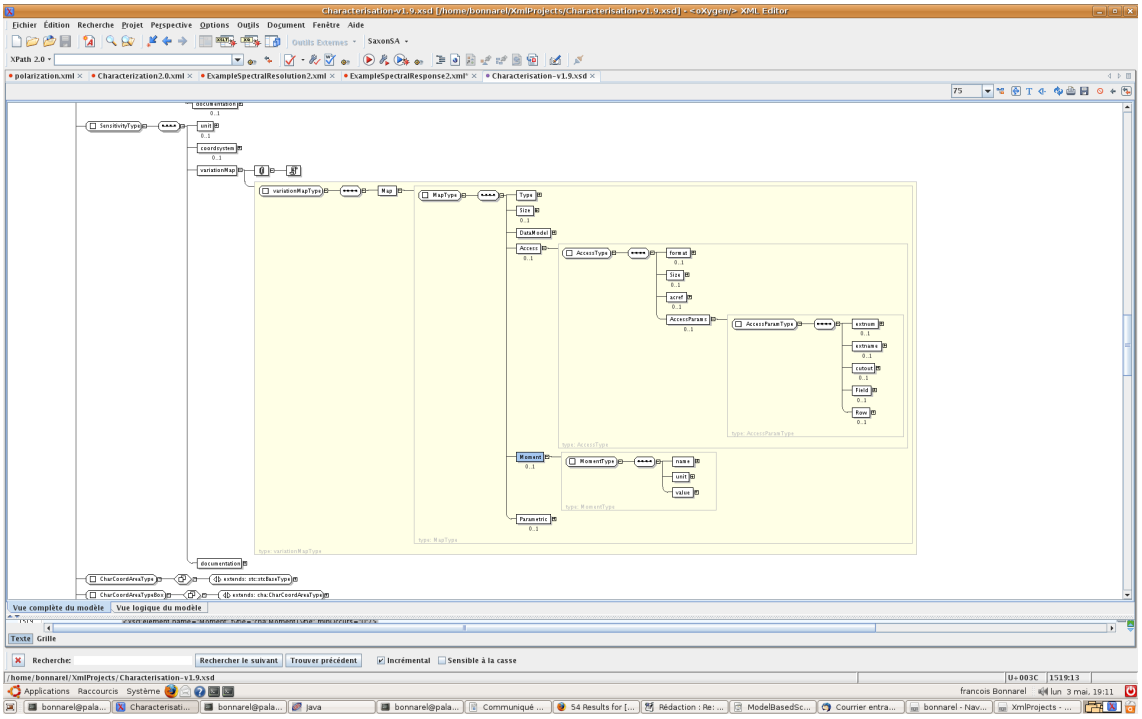

Figure 9: Structure of the VariationMap object.

or observation, identification, physical characterisation, provenance (either instrumental or software) and curation. A description of the early stages of this development can be found in the Observation DM IVOA Note [12]. In parallel, an effort dedicated to spectra was lead by the DM Working group: The Spectrum data model [10] adapted to simple spectra. For the overall Observation Data Model, the physical characterisation has been identified to be on first priority already in 2004. It was completed as an IVOA recommendation after 4 years of discussion which included computer scientists, astronomers and data providers under the lead of J. McDowell ([1])

The Characterisation data model was setting up a logical framework to describe the properties/features of each observation in the VO, but it was lacking a simple DAL access protocol until the emergence of ObsTAP which is solving that not only for the main part of characterisation but for a simple subset of Observation : the so called "core components" [13]. The full Observation data model is still currently developed and will include a detailed description of the Provenance (instrumental and computational) for observed datasets. DataLink is probably the way to give access to exhaustive XML serailizations of the models fully describing the datasets with all the complexity allowed by the model The consistency between these efforts

is a major goal.

5 Appendix B: Access package

The package describes the format of the file, the URI pointing to it and is completed by an AccessParams structure. It is an extension of the Access class in SSA data model. Actually in the general case we need to describe not only the file globally but some specific parts in the case of files with complex structure ... Variation maps may sometimes be part of the same FITS multiextension file than the data but in a different FITS extension. We intend to describe all kind of FITS tables or arrays, FITS multiextensions files, and tables in VOTABLE. We also intend to describe internal structures (dataset paths) of tar or zip archives.

Norman Gray recently proposed a mechanism to do the same fine grained access using extended URI.

The AccessParams structure is made of several attributes:

- The **extnum** attribute gives the extension number in FITS/extension file
- The **extname** attribute gives the table name in a VOTABLE file or fits:extension table file
- The **cutout** attribute can apply either to a global array if the considered extension is such an array or to a field if we have an array type cell in the TABLE case. It is a description of the subarray limits and sampling in FITSIO syntax.
- FIELD and Row attributes allow to select the corresponding features in the considered table

References

[1] M. Louys, A. Richards, F. Bonnarel, A. Micol, I. Chilingarian, J. McDowell, and the IVOA Data Model Working Group. IVOA Recommendation: Data Model for Astronomical DataSet Characterisation. *ArXiv e-prints*, November 2011.

[2] M. Cappellari and E. Emsellem. Parametric Recovery of Line-of-Sight Velocity Distributions from Absorption-Line Spectra of Galaxies via Penalized Likelihood. , 116:138–147, February 2004.

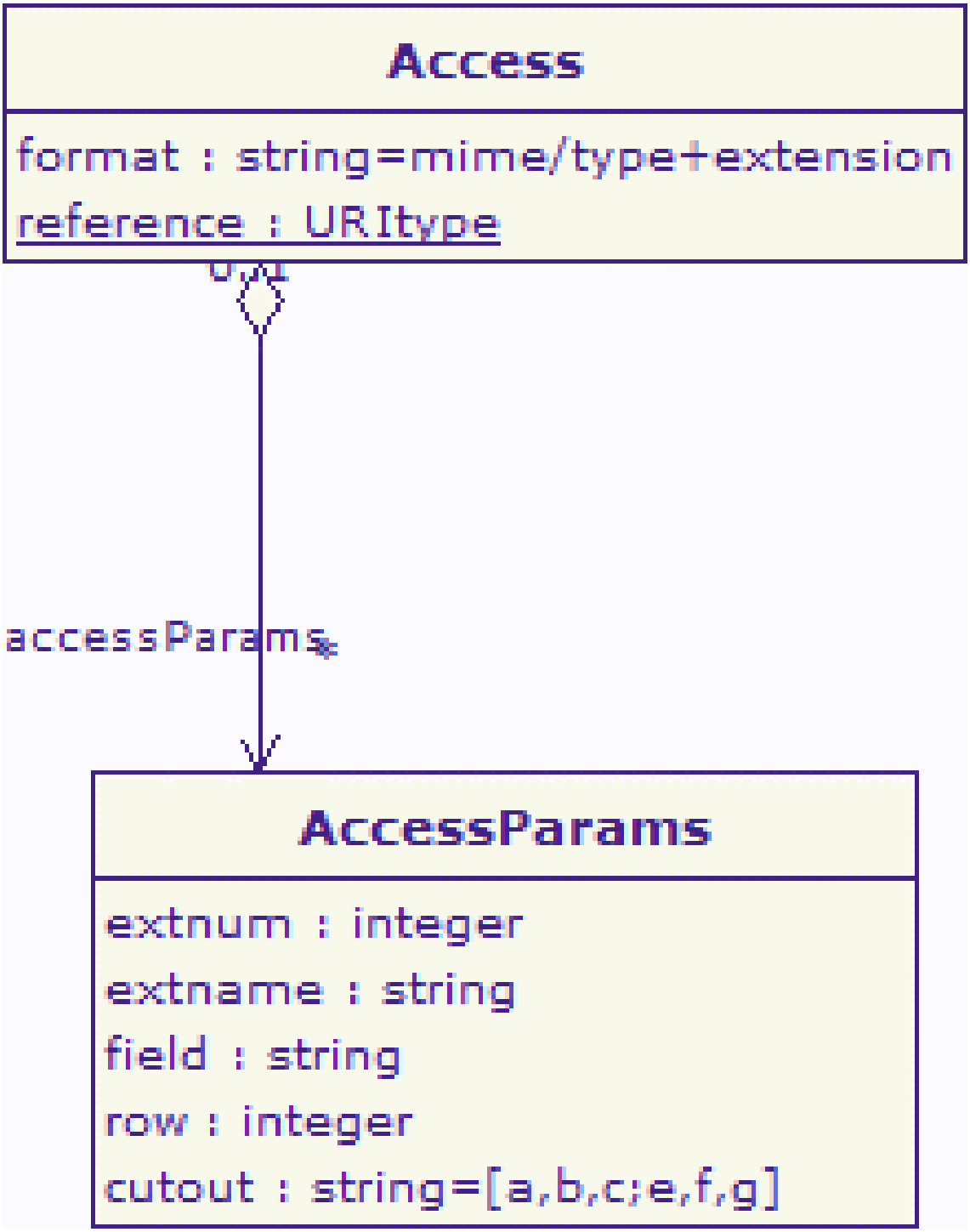

Figure 10: Reusable Access package

[3] I. V. Chilingarian, P. Prugniel, O. K. Sil'Chenko, and V. L. Afanasiev. Kinematics and stellar populations of the dwarf elliptical galaxy IC 3653. , 376:1033–1046, April 2007.

[4] I. Chilingarian, P. Prugniel, O. Sil'Chenko, and M. Koleva. NBursts: Simultaneous Extraction of Internal Kinematics and Parametrized SFH from Integrated Light Spectra. In A. Vazdekis and R. F. Peletier, editors, *IAU Symposium*, volume 241 of *IAU Symposium*, pages 175–176, August 2007.

[5] A.M.S Richards and F. Bonnarel. Ivoa note: Note on the description of polarization data, 2010. `http://www.ivoa.net/Documents/Notes/Polarization/`.

[6] A.M.S Richards. Radio interfeometry data in the vo, 2010. `http://wiki.ivoa.net/internal/IVOA/SiaInterface/Anita-InterferometryVO.pdf`.

[7] Greisen E.W. The fits interferometry data interchange convention - revised, 2012. `http://www.aips.nrao.edu/FITSIDI.pdf`.

[8] R. Plante, K. Benson, M. Graham, G. Greene, P. Harrison, G. Lemson, T. Linde, G. Rixon, A. Stebe, and the IVOA Registry Working Group. IVOA Recommendation: VOResource: an XML Encoding Schema for Resource Metadata Version 1.03. *ArXiv e-prints*, October 2011.

[9] M. Louys, O. Laurino, L. Michel, D. Tody, M. Demleitner, F. Bonnarel, A. Micol, G. Lemson, M. Cresitello-Dittmar, and J. McDowell. Utypes: a standard for serializing data models instances. version 0.7. ivoa dm wg internal draft, 2012. `http://wiki.ivoa.net/internal/IVOA/Utypes/WD-Utypes-0.7-20120521.pdf`.

[10] Jonathan McDowell et al. Ivoa spectral data model. *http://www.ivoa.net/Documents/latest/SpectrumDM.html*, 2007.

[11] I. V. Chilingarian, P. Di Matteo, F. Combes, A.-L. Melchior, and B. Semelin. The GalMer database: galaxy mergers in the virtual observatory. , 518:A61, July 2010.

[12] J. McDowell et al. Observation data model for astronomical dataset. *http://www.ivoa.net/Documents/latest/DMObs.html*, 2005.

[13] M. Louys, F. Bonnarel, D. Schade, P. Dowler, A. Micol, D. Durand, D. Tody, L. Michel, J. Salgado, I. Chilingarian, B. Rino, J. de Dios Santander, and P. Skoda. IVOA Recommendation: Observation Data Model Core Components and its Implementation in the Table Access Protocol Version 1.0. *ArXiv e-prints*, November 2011.