Towards a working definition of a Quantity: Use-cases

Justification for use-cases

We have assembled a number of use-cases so that it is possible to better answer the question what is a quantity and what is it used for?. Naively, one might think that this kind of effort is overkill as the term/word quantity exists in any English dictionary you might care to pick up. But closer inspection reveals that the definition is not precise enough for our uses in the VO. Take, for example, the definition of quantity given by Websters II New Riverside Dictionary (1984):

> **Quantity** : 1. a. A specified or undetermined number or amount.
> b. A large number or amount
> c. An exact amount or number.
> 2. The property or aspect of a think that can be measured, counted, or compared.
> 3. *Math.* The object of a mathematical operation.
> 4. The length of a vowel or consonant sound expressed in terms of its temporal duration.
> 5. *Logic.* The exact character of a proposition in reference to its universality, singularity, or particularity.

Clearly some of the choices for defining quantity are in conflict with the others (such as 1, which suggests that a number, or string (like large) may be quantities, but 3, which involves only mathematical operations might preclude strings, whereas 5. appears to imply that boolean values like true may also be quantities). So, if we clearly can't apply all choices from the dictionary, how are we to choose the correct choice of definition? Must we choose one? Are more than one of these choices correct for our usage? Might we blend choices in the above definition to create our own, working, definition? And even if we might have worked out a dictionary definition of quantity, will it be sufficient for building software (e.g. we might have to be more specific in order to define the VO quantity for software development).

It is questions and issues like these which drive us to adopt a more formal approach to defining the quantity. In modern software design, this formal approach is encapsulated through the development of use-cases. Each use-case, to quote from Erikkson, 1998 ["UML Toolkit", Wiley Computer Publishing", pg 45]:

> "A use-case is used to describe what a system should do. A use-case model is built through an iterative process during which discussions between system developers lead to a requirement specification on which all agree."

Thus, we present a relatively broad sampling of the use-cases that the community have put forth as to the workings/definition of what the quantity should do.

Quantity Use-cases

The following use-cases have been gathered from various documentation and discussions held online and at face to face meetings between IVOA members. These use-cases have been numbered so that they may be easily referenced. The numbering of use-cases should not be necessarily take to imply the relevance of one particular use-case over another.

Use-case 1. (Search/Access)

We need a simple generalized mechanism for describing objects (read subjects now) and for access or query to them. The idea is that objects are described by a set of properties/quantities with values known (measured) to be within a range. The known data on each object is describable by a set of quantities.

O = O(Q1, Q2, ... QN)

These Qs need to hold values, accuracy, and meta-data Qs. And each Q needs to allow for arguments of other Qs to fully specify properties like flux(position) or M(<r) or densities(r) or column_densities(R), because otherwise one is not fully describing the Object. To retrieve objects with properties within a certain range, one places constraints on all objects to return a set of objects satisfying:

O(return) <-- (C(Q12) and/or C(Q28) and/or C(Q132))

Where the constraints can be on the values V, meta-data values Qm, or accuracy a. By a constraint we mean the following in the case with a quantity with scalar values and one piece of meta-data,

C(Q_i) <-- O(c1 < V_i <c2) and O(c5 < a_i < c6) and C(Qm) etc

In the case where the values are strings, as in the class name, then one constrains by regular expression or string equivalence. The strings that represent the Objects I call literals, and are somewhat distinct from quantities.

Use-case 2. (Search)

Consider the search use-case that an astronomer is looking for all flux measurements of the visual band. Of course, the search should find those "atomic" quantities that are V-band fluxes, but it should *also* be able to find fluxes from other flux quantities which are described in terms of wavelength (e.g. F(lambda)) and where some part of their wavelength range overlaps the visual band (300-700 nm). For these quantities F(l)=> F(300-700nm) == F.

Use-case 3. (Support higher-level DM constructs)

The quantity should form the basis for describing the data held in higher level components of the VO data model. This makes it easy to cut/paste/compare or use parts of these components in other software. Some example components (which use quantities, but are not themselves quantities) include things like:

sky coverage
coverage on the sky, if applicable
e.g., a circular or rectangular region or aperture on the sky
Can be used to estimate the WCS but the full WCS should be
defined elsewhere.

time coverage / bandpass
time of observation
refValue, hiValue, loValue, fillFactor
refValue is the mean time of observation, e.g., mid-point

spectral bandpass
range of spectral frequencies in data
id, refValue, hiValue, loValue, units, fillFactor
id is user-defined bandpass name, e.g., "V", "SDSS_U", "K-Band", etc.
refValue is the characteristic frequency of the bandpass

spatial bandpass
range of spatial frequencies in data
hiValue, loValue, units, fillFactor (no refValue)
loValue is also known as the spatial resolution

flux bandpass
range of flux values in data
hiValue, loValue, units, fillFactor (no refValue)
loValue is also known as the limiting flux or magnitude
hiValue is saturation limit or maximum flux

Use-case 4. (search/access)

The model should provide a common way to express quantities associated any physical phenomenon so to:

a. aid users and developers who might see instances of the model in recognizing the concept as a quantity,

b. enable the use of common software for manipulating quantities independent of the physical phenomenon they represent.

Use-case 5. (Data representation)

It should be possible to render a model instance in a form that is naturally readable and recognizable by scientists, for example, of a simple human-readable model instance could be "10.3 mJy".

Use-case 6. (Search, exchange and fusion)

Quantities need to be able to support search and assembly of simple physical concepts across archives. An example search at this level might be find all tables which describe galaxies and have a complete set of sky coordinate and Infared flux columns and return them to me as a single file. The me might be an astronomer, or in the future, a service which might process this data further to the astronomer's needs.

Use-case 7. (Data exchange)

We need to be able to make data within data-bases, FITS files and other formats interchangeable. Thus, quantities should be able to express the full range of data concepts that exist in astronomical data today (as defined by this target database/file population). Some examples which help frame some bounds on this:

1. Be able to construct a table with 2 columns, each column with different units and UCD (meaning). One column has numbers, the other one has strings. There are systematic errors on the numbers column. Quantities should represent the columns.

2. Create a 2-D spatial quantity of observed brightness (flux/mag) for a CCD camera with axes described in terms of x,y pixels. The brightness values should be described in terms of their accuracy which might include statistical or systematic errors or quality flags.

3. Create a quantity as above but with positional axes, also including background image. The positional axes have values which have errors (statistical) on them.

4. Create a 2-D quantity mapping a vector field (e.g. a velocity field) with axes of position. The vectors have errors on their components.

5. As for '4', but this time axes are also vectors themselves.

6. Create a 1-D flux quantity, which has accuracy and units describing its brightness.

7. Case '6' but with associated sensitivity map.
8. Create a quantity like 4, but allows for alternative descriptions of the axes.

Use-case 8. (Use in higher level models)

VO data access is data model based so some immediate use cases for data models from DAL are:

- o Simple data models for image (SIA) and spectrum/SED (SSA). SIA 1.0 largely side-steps this issue, but SSA will need a more explicit spectral/SED data model. We already have a good start on these.

- o How to characterize datasets such as SIA and SSA, i.e., the coverage issue, component data models, etc. What are the components? This is tricky, for example time of observation is actually a bandpass, spatial resolution is related to spatial bandpass, limiting flux could be an attribute of some sort of "flux bandpass", and so forth.

- o How to represent the component data models in a DAL query response, e.g. VOTable, to summarize the characteristics of a dataset which is a candidate for access. This gets into issues such as UTYPE and data model representation in XML, UCD2, grouping constructs, etc.

- o How to use dataset characteristics to refine a query. Any dataset characteristic which we model will be used both to describe a dataset, and to refine queries looking for such datasets (e.g., can we reduce spectral bandpass to a filter name or coarse bandpass ID such as optical/ir/radio etc.).

- o How to represent actual datasets in various data formats such as XML. We need to map external data into our VO data models and represent such virtual data in XML and (to a lesser extent) other formats. This is closely related to the query response, but some actual data is passed as well. DAL will return data in these formats, e.g., a 1D spectrum or an image cutout. Functions will need to be able to operate directly on data in this form.

Use-case 9. (Data exchange, fusion and representation)

I. Extracting simple measurements with errors. Ensure they have errors and are associated to objects that may be associated with higher level objects, clusters.

#Obtain Iron abundances [Fe/H] for stars in M92.

II. Extracting object holders of properties. Here we ensure bi-directionality of above associations. And introduce a comparison quantity. And we need to hold intermediate arrays and the averages of these arrays. The output is an array of objects, (Gcs).

#Find globular clusters whose stars have average iron abundance < 0.25.

III. Include conditional items in search.

#Same as above but only use cases where more than 2 Iron
lines were included in the analsys.

IV. Extracting meta-data quantity.

#Obtain the assumed Iron transition probabilities assumed in previous example.

V. Extracting line identifications, confidence level etc
#Obtain which lines have been identified in far UV in theISM, their intensities, confidence levels, and central wavelengths.

VI. Extracting 2-d data. One expects either single aperture or CCD data, and everything should carry along errors and quality assessments. Of course the aperture location and pixel sizes must be carried as well.

#Obtain 350 micron emission data of Sgr B2 (galactic center object)
that have resolution < 20".

VII. Extracting 3-d data.
#Obtain, for region XXX, microwave background data from WMAP
in all spectral channels in regions excluding known
galaxies (using the V=25th mag/sq" surface brightness
contour).

VIII. Extracting vectors.
#Obtain from N-body simulation XXX, using mass points in
halos at the present time, the center of mass motion of each
galaxy+recent infall components at z=5.

IX. Extracting tensors.
#Assuming all clusters in the Abell Catalog have mass of
1E14 solar masses, calculate how the acceleration vector
increases with distance from the Milky Way in each of the
3 cartesian coordinates.

X. Instrument as Objects.
#Find a spectragraph at a major observatory with dispersion
<100A/px and throughput
better than 20% over the region 3500-4000A. Display the throughput
curve for each of these.

Use-case 10. (fusion/utility)

It should be possible for a researcher to retrieve information from the VO, and then select various data within the returned structure for processing. These data need to be capable to be used in various basic arithmetic operations, like *, /, ... for non-conformant units, and +, -, ... for conformant units. Use case example is the conversion of photon frequency to energy; another is the conversion of ADUs on a CCD to sky brightness in mag/arcsec^2.

Use-case 11. (Data representation)

What I would like for Christmas is (the ability for quantities to express):

* World Co-ordinates (does that count as two quantities which are angles?)
* Brightness (ie flux, magnitudes)
* Wavelength/Frequency
* Quality... perhaps this is not a Quantity...
* Errors

Use-case 12. (data representation)

The quantity is a container class for _meta-data_ as well as _data_. This is because what is considered meta-data versus data is a contextual issue (e.g. data is what you look for, to be meta-data the information is always given in association with owning data).

So a use-case is that the columns and the keywords within a FITS file should be representable by quantities. In this fashion, it should be possible to easily search for given FITS keywords (and synonyms) not only in registered examples of FITS files in the VO, but incidence of those keywords in databases, ascii text (incl. XML) and perhaps even word documents.

Use-case 13. (data representation)

SsaSurvey provides some examples of data that should be held in quantities:

number: ~100 K - 1 Mio. spectra/data provider; often only several 100s in very specific catalogs
size: few KBs - few MBs/spectrum
wavelength/energy ranges: gamma, X-ray, UV, optical, IR, radio
resolution: 20 - 1.000.000
linear and non-linear/irregularly sampling (for instance in energy domain)
often with WCS
stellar parameters: e.g.: Teff, log(g), [Fe/H]
noise
masks (coded aperture mask data, INTEGRAL)
observation dependent transfer matrix to convert counts to phys. units
spectral features (SLOAN): emission, absorption lines and parameters, emission redshift params. cross-correlation parameters
time resolved spectral information (XTE, BATSE, Swift, HETE)

Columns/data items:
wavelength (better fravergy: frequency, wavelength, energy)
flux (magnitude, flux, flux density, counts)
flux error
quality flag (quality of what exactly?)
variance arrays
photon events
antenna temperature vs. frequency (e.g. SWAS spectra)

Use-case 14. (data representation)

(from SsaSurvey)

As a minimum, data should consist of (x,y) pairs of wavelength-flux values expressed in wavelength/frequency/energy units, and spectral flux density units respectively. The units information should be present as well, in header keywords or table column descriptors or any other suitable form.
Additional information that could be included with the above, (if available) is a third value associated with each (x,y) pair representing the measurement error in flux density. Also, it won't hurt if obvious pieces of information such as the object name, are provided as well.
Uncalibrated data (such as in counts/s units) can also be ingested but not much else can be done with it besides plain plotting.
FITS WCS; useful for well sampled rasters.

Use-case 15. (for realization of UCD's)

Please find below some possible use cases for UCD.

-- How do the different proposed UCD/DM schemes help with these cases?

-- What does the query look like in each case, and how would that query be coded?

-- Can you supply additional use cases? Please try to stay within the realm of the possible here!

-- If you refer to a Data Model, please try to be concrete about who/how/example. Please do not simply invoke "the Future IVOA Data Model".

(1) Cone search. How to decide which columns are the RA,Dec that was used in the search. What frame (B1950, J2000, ...) do these come from? If there are columns with ID, what sort of ID is it, and how do I resolve it?

(2) SIAP search. Find the column that contains the URLs where the images are. Find out if there are other columns that have RA,Dec of the image center.

(3) We have for example a crossmatch service that is clever enough to know about error ellipses. How does it get from a table the most sophisticated error info that is there: (a) position (b) circular error (c) ellipse error.

(4) We want to compare photometry in two tables covering the same star cluster. How do I decide if they share measurements in the same filter? One has R band, the other has Halpha. What happens if fluxes are expressed differently -- eg number / energy /
magnitude / luminosity density.

(5) I want distances to stellar objects measured in meters, so I can make a 3D display for the children. How do I recognize a redshift (z) value, how do I recognize a radial velocity, how do I recognize an actual distance measure?

(6) I am looking for supernovae that have both optical and Xray measurements. Can I (should I) use UCD to help my search?

(7) How do I find 21cm observations (that may be redshifted), which also have polarization information?

(8) I have a heap of data expressing modeled and observed values for a handful of quantities; celestial coords and proper motions would do as an example. I want to read out all the modelled versions of one of these quantities and compare them to assess the spread of the models. Then I want to do the same with the observations. What do I look for?