The product-type vocabulary during Datalink RFC

F.Bonnarel, M.Demleitner

CENTRE DE DONNÉES ASTRONOMIQUES DE STRASBOURG

GERMAN ASTROPHYSICAL VIRTUAL OBSERVATORY GAVO

IVOA

The dataproduct type vocabulary – where does it come from ?

- Initial list in ObsCore specification :
  - Fully defined inside the document
  - Describes the type of dataset we are discovering with ObsTAP/SIA
  - Can help the client to manage the dataset herself or send to another tool

Obscore List

3.3.1. Data Product Type

The model defines a *data product type* attribute to describe the high level scientific classification of the data product being considered. This is coded as a string that conveys a general idea of the content and organization of a dataset. We consider a coarse classification of the types of dataset interesting for science usage, covering: image, cube, spectrum, SED, time series, visibility data, and event data.

- **image** An astronomical image, typically a 2D image with two spatial axes, e.g., a FITS image. The image content may be complex, e.g., an objective-grism observation would be considered a type of image, even though an extracted spectrum would be a Spectrum data product.
- **cube** A multidimensional astronomical image with 3 or more image axes, e.g., a spectral image cube, a polarization cube, a full Stokes radio data cube, a time image cube, etc. The most common format for astronomical "cube" data products is a multidimensional FITS image, however other formats are allowed so long as they are adequately described.
- **spectrum** Any dataset for which spectral coverage is the primary attribute, e.g., a 1D spectrum or a long slit spectrum.
- **sed** A spectral energy distribution, an advanced data product often produced by combining data from multiple observations.
- **timeseries** A one dimensional array presenting some quantity as a function of time. A light curve is a typical example of a time series dataset.
- **visibility** A visibility (radio) dataset of some sort. Typically this is instrumental data, i.e., "visibility data". A visibility dataset is often a complex object containing multiple files or other substructures. A visibility dataset may contain data with spatial, spectral, time, and polarization information for each measured visibility, hence can be used to produce higher level data products such as image, spectra, timeseries, and so forth.
- **event** An event-counting (e.g. X-ray or other high energy) dataset of some sort. Typically this is instrumental data, i.e., "event data". An event dataset is often a complex object containing multiple files or other substructures. An event dataset may contain data with spatial, spectral, and time information for each measured event, although the spectral resolution (energy) is sometimes limited. Event data may be used to produce higher level data products such as images or spectra.
- **measurements** A list of derived measurements gathered in a particular original dataset of one of the previous sort after some analysis processing, like a source list, or more generally a list of 'results' attached to such datasets.

The dataproduct type vocabulary – context evolution : DataLink

- DataLink 1.1 introduces new « content_qualifier » field
- DataLink not reserved to dataset discovery
- Also for sources in catalogs
- Then the link may be a dataset (spectrum, timeseries)
- Dataset (#thing) to dataset (link) also possible
- Semantics and format do not code the same concepts

  → new field content_qualifier. « Nature » of the link

  Could be something else than product type in the future
- So the vocabulary must become an IVOA vocabulary independant from ObsCore

The dataproduct type vocabulary – context evolution : other usage

- Could also be used to qualify the content of data collections
  - In the context of the registry
- Could enrich ObsCore dataproduct_type vocabulary in the future (virtuous loop !)
  - New kind of time domain data (dynamic spectra)
  - Radio data extension (velocity fields, rotation measures)?
  - High energy extension ?

Dataproductype : What is it for in practice ?

- It’s a guide for the software to know what to do with the link
  - If TopCAT discovers a link where content_qualifier tells it’s an image , send it to Aladin.
  - If Aladin discovers a link where content_qualifier tells it’s a spectrum, send it to SPLAT or CASSIS
  - If Firefly discovers a link where content_qualifier telles it’s « measurements » send it to TOPCAT

What kind of properties distinguish the dataproduct type concept ?

- 4 different types of properties
  - Which sampled data axes are independant ?
  - Which sampled data axes are dependant (flux, velocity, pol angle ….) ?
  - Which independant axes are sparsed and which are regular ?
  - Table or bitmap organization ?
- 2D image have spatial axes regularly sampled.
- Dynamical spectra have spectral and time axes regularly sampled

What kind of constraints do apply to dataproduct type terms?

- Do not break behavior driven by ObsCore definitions if we change them
- Organize some hierarchies in terms
- Allow multiple parents (SKOS vocabulary)

- However do not try to reflect all relationships with four properties
  - we would need new terms for each set of values of properties

Adding new terms/ changing definition creating hierarchies

- We let « event » and visibilities alone

There is a new proposal :
https://www.ivoa.net/rdf/product-type/2021-11-18/product-type.html

- Look at spectrum :

ObsCore definition : *Any dataset for which spectral coverage is the primary attribute, e.g., a 1D spectrum or a long slit spectrum.*
2021 proposal : *Fluxes or magnitudes given as a function of a spectral coordinate.* (too restrictive)
2023 proposal (Markus) : *A dataset with a spectral axis. This can be a classical spectrum if each spectral point is mapped to a flux. For more complex data (e.g. dynamical spectra or spectral cubes), narrower terms should be used.*

Narrower terms : #sed,#dynamical_spectrum,#spectral_cube

Adding new terms/ changing definition creating hierarchies

- Look at timeseries:

ObsCore definition : *A one dimensional array presenting some quantity as a function of time. A light curve is a typical example of a time series dataset. spectrum or a long slit spectrum.*

2021 proposal : *One or a few scalar observables as a function of time.* (too restrictive)

2023 proposal (Markus) : *A dataset with a temporal axis. When time instants are mapped to something non-scalar (e.g. dynamical spectra or temporal cubes), narrower terms should be used.*

Narrower terms : #dynamical_spectrum, #temporal_cube, #light_curve, #velocity_curve, #motion_curve

Adding new terms/ changing definition creating hierarchies

- Look at image:

ObsCore definition : *An astronomical image, typically a 2D image with two spatial axes, e.g., a FITS image. The image content may be complex, e.g., an objective-grism observation would be considered a type of image, even though an extracted spectrum would be a Spectrum data product.*
2021 proposal : *idem ObsCore*
2023 proposal (Markus) : *A dataset with at least two spatial dimensions mapping, typically, the sky or some celestial body.*

I would propose : *A dataset with at least two spatial dimensions regularly mapping, typically, the sky or some celestial body.*

Alternatively = *« exactly two spatial dimensions »*

Narrower terms : #cube ? #velocity_field, #rotation_measure-map ?

Adding new terms/ changing definition creating hierarchies

- Look at cube:

ObsCore definition : *A multidimensional astronomical image with 3 or more image axes, e.g., a spectral image cube, a polarization cube, a full Stokes radio data cube, a time image cube, etc. The most common format for astronomical "cube" data products is a multidimensional FITS image, however other formats are allowed so long as they are adequately described.*

2021 proposal : *idem ObsCore*

2023 proposal (Markus) : *A multidimensional astronomical dataset with 3 or more image axes,*

I would propose : *A multidimensional astronomical dataset with 3 or more regularly sampled image axes,*

Narrower terms : #spectral_cube, #temporal_cube,

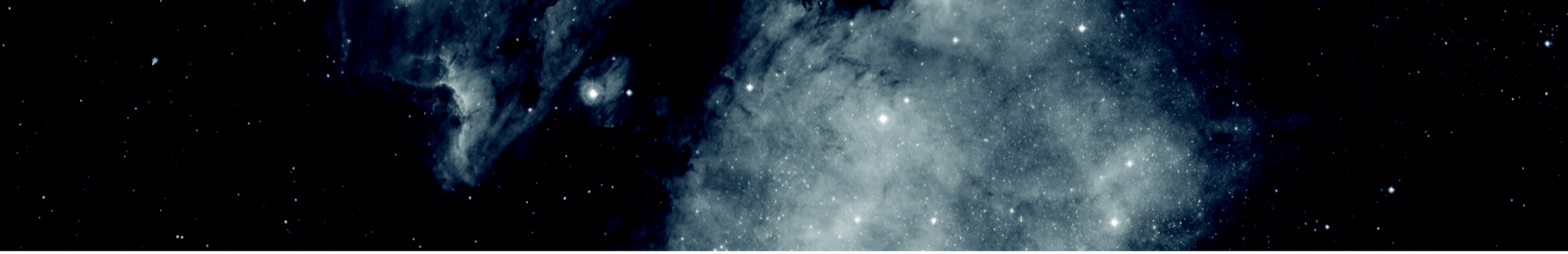
Adding new terms/ changing definition creating hierarchies

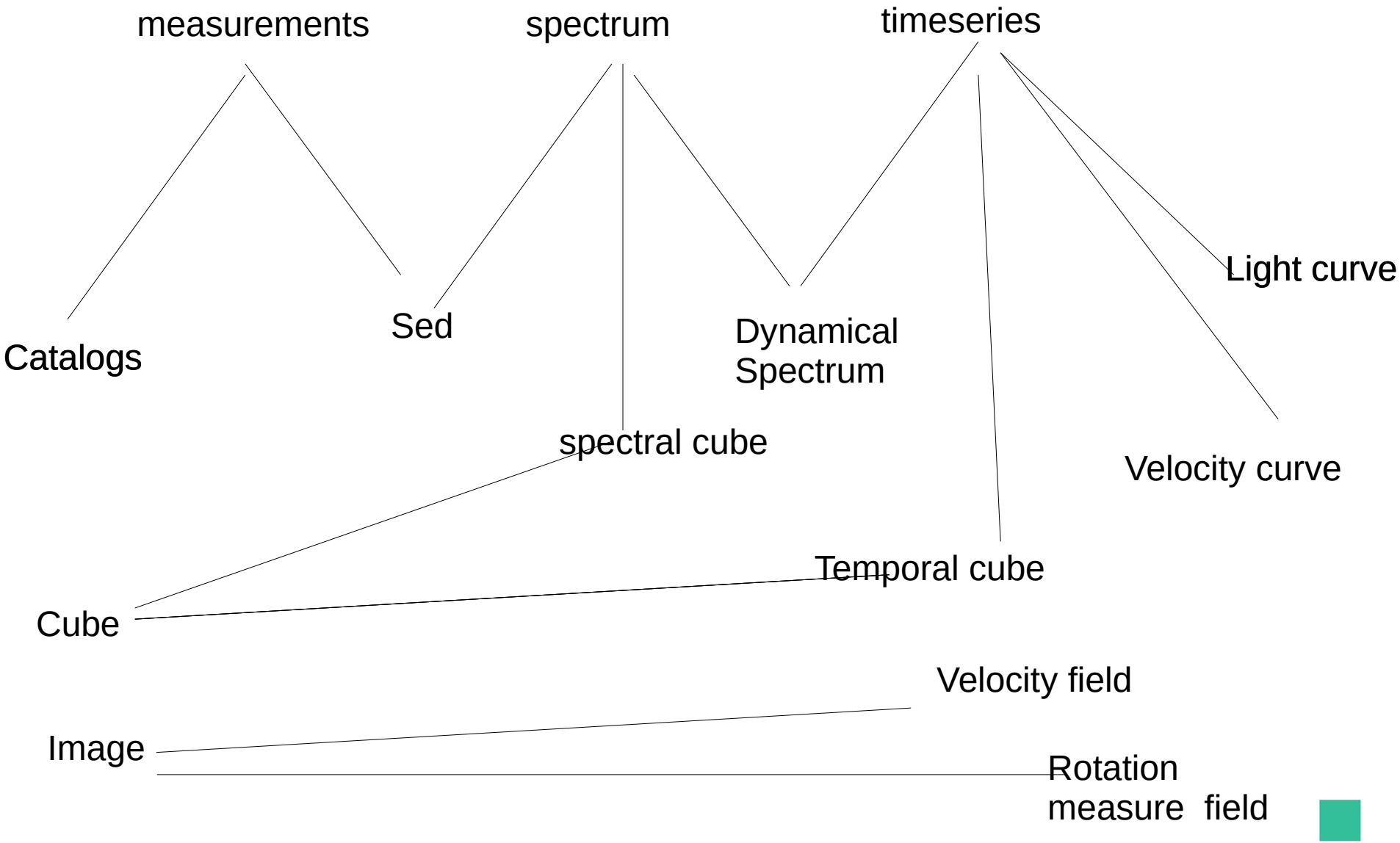
- Look at measurements:

ObsCore definition : *A list of derived measurements gathered in a particular Original dataset of one of the previous sort after some analysis processing, like a source list, or more generally a list of 'results' attached to such datasets.*
2021 proposal : *Generic tabular data not fitting any of the other terms. Because of its lack of specificity, this term should generally be avoided, and new, more precise terms should be introduced instead.*
2023 proposal (Markus) : *none ?*

I would propose : « *Generic multi axis tabular data. Fits for derived Measurements, catalogs tables »*
Or alternatively : *« let ObsCore and add catalog as a wider term ? »*

This branch valid for anything sparsed.




measurements
spectrum
timeseries
Catalogs
Sed
Dynamical Spectrum
Light curve
spectral cube
Velocity curve
Temporal cube
Cube
Velocity field
Image
Rotation measure field

Alternative solution

- The wider terms are renamed (Markus): spectrum → resolves-spec
  timeseries → resolves-time
  image (cube) ? → resolves-space

- Other parents for observable/dependant variable (François): flux mapping, velocity mapping, rotation measure mapping

- #velocity_curve is resolves-time and velocity mapping.

- 1D spectrum is resolves-spec and flux mapping
- Light curve is resolves-time and flux mapping

- Too artificial ?