Fig. 1

1. Advanced Column Metadata

Markus Demleitner
msdemlei@ari.uni-heidelberg.de

Blind Discovery: Don't look for "2MASS", look for "Infrared data around Betelgeuze down to 18 mag in K".

Part 1: Enabling registry queries in space, time, and spectrum. See VODataService 1.2 and RegTAP 1.2.

Part 2: Characterising column content (the "down to 18 mag" part).

(cf. Fig. 1)

2. Use Cases

Somewhat more precisely, here are potential use cases for advanced column metadata:

- *Deep Survey* Give me data for the M32 reaching 25 mag in the infrared K band.
- *High Redshift* I am looking for Galaxies with redshifts above 1.
- *High Precision* I need a catalogue of proper motions with errors below 0.2 mas/yr.
- *High Precision advanced* I need a catalogue of proper motions with errors below 0.2 mas/yr at 15^{m} in V.
- *Calibrated* Where are flux-calibrated spectra for stars in globular clusters?
- *Planning* How many rows will `WHERE col<30` return (approximately)?

High Precision advanced and *Planning* will probably be science fiction for a long time to come; they are mainly here to say where I think at this point the boundaries of this endeavour are.

3. Prior Art

VOTable: VALUES/MIN, VALUES/MAX, OPTION

Missing from **VODataService**'s column model.

Missing in the **TAP** columns schema

Then, Grègory's Gaia DR1 publication, with, per float column:

- min_value, max_value
- q1, median, q3
- mean
- filling (number of non-NULLs)

4. Float Columns

Derived from that, I'm now proposing for metadata of float columns:

- min_value, max_value – mainly for VOTable compatibility
- percentile03, percentile97 – for Gaussians, that's pretty much "2σ"
- median
- fill_factor – that's $n_{\mathrm{NULL}}/n_{\mathrm{rows}}$

Absent at this point: Moments (mean, stddev, skewness, kurtosis, ...). I've left them out because many of the distributions interesting here (e.g., magnitudes or redshifts in catalogues) are severely non-Gaussian, and these familiar measures tend to be misinterpreted then. Also, in contrast to the percentiles, they are not linear, so, for instance, with distance r and parallax ϖ, you have $r = 1/\varpi$, but $\langle r\rangle \neq \langle 1/\varpi\rangle$. I'm willing to haggle, though.

5. Discrete Values

Why? Well: "Do you have rows with `o_calib_status=2`?"

Model: A sequence of values (perhaps even: bin centres?) and relative frequencies:

$$[(v_1, f_1), (v_2, f_2), \ldots].$$

Perhaps constrain how large these may become?

6. To Do

VODataService: Add `vs:Stats`-typed element `stats` to `vs:BaseParam`. Continuous stats are attributes of that.

We need serialised values (the v_i, possibly even for median and friends): What XSI type? Proposal: `xsi:type` with VOTable serialisation. But really, I think we need to be guided by implemenation here.

TAP: For symmetry with VOSI tables, `TAP_SCHEMA` should grow $\sim$ the same information. But: efficiently queriable columns must be type-clean, so: no token trick here.

RegTAP: Extend `rr.table_column`? Or rather add tables `rr.stat_num`, `rr.stat_token`, and `rr.stat_discrete` to deal with non-number types?

7. Implementation Status

There's a Note[1] out spelling out these proposals.

Metadata for continous columns is produced and published via a VODataService extension in the upcoming DaCHS 2.4.

The GAVO RegTAP network at http://reg.g-vo.org/tap has a `rr.g_num_stat` table that publishes the harvested information ($\sim$ 1000 records).

Thanks!

[1] `https://ivoa.net/documents/Notes/colstatnote/index.html`