UCD and VOUnit Usage in the VO

Mark Taylor (Bristol)

Operations IG
IVOA Interop
Online

3 November 2021

$Id: uuc.tex,v 1.17 2021/11/02 18:23:55 mbt Exp $

Outline

- Introduction: UCDs and VOUnits
- UCD/VOUnit validation tools
- State of UCDs in the VO
- State of VOUnit usage in the VO
- Actions?

Column Metadata Vocabularies

Controlled vocabularies for column metadata in the VO:

- **UCD**: Uniform Content Descriptors
- **VOUnit**: Units in the VO

Both have syntax defined by IVOA standards (Semantics WG)

Used for semantic annotation of table columns

- VOTable (attributes `ucd` and `unit` in `FIELD`/`PARAM` elements)
- TAP service metadata (fields `ucd` and `unit` in `TAP_SCHEMA.columns`)
- VODataService instances in the Registry or TAP `/tables` endpoints (elements `ucd` and `unit`)

UCDs

UCD standards:

- There are two incompatible standards, UCD1 and UCD1+.
- UCD1 (CDS): `http://vizier.u-strasbg.fr/viz-bin/UCDs` (any other documentation?)
  - ▷ Examples: `POS_EQ_RA_MAIN`, `PHOT_JHN_B`, `INST_ID`
  - ▷ Inflexible, ancient (pre-2004?), deprecated
  - ▷ Still appears in some very old standards (current Cone Search 1.1, SIA 1.0 but not 2.0)
- UCD1+ (IVOA): UCD1+ 1.3 (REC, 2018), UCD1+ 1.4 (EN, 2021)
  - ▷ Examples: `phot.eq.ra;meta.main`, `meta.id`
  - ▷ Structured and flexible
  - ▷ Processes defined for extending the vocabulary: UCDlistMaintenance 2.0 (2019)

UCD Mandated usage:

- VODataService 1.1: *"There are no requirements for compliance with any particular UCD standard. The format of the UCD can be used to distinguish between UCD1, UCD1+, and SIA-UCD*."*

  * *Untrue — removed in VODataService 1.2*
- VOTable 1.4: *"UCD1+ usage is recommended, but applications using the older vocabulary are still acceptable in this version of VOTable"*
- TAP 1.1: *"the information in the TAP_SCHEMA is equivalent to that defined by VODataService"*

$\Rightarrow$ *UCD1+ usage is not required for TAP/registry, but it's good practice*

VOUnits

VOUnit standard

- VOUnit 1.0 (2014)
  - ▷ Defines recommended *VOUnit* syntax (machine- and human-readable)
    - Rules include: no whitespace; multiplication as "`.`"; exponentiation as "`**`"; no multiple "`/`" at top level
    - Examples: `mag`, `mas/yr`, `1e-23J/(s.cm**2.Angstrom)`
  - ▷ Codifies and *"broadly endorse[s]"* some alternate syntaxes (FITS, OGIP, CDS)

Unit mandated usage:

- VODataService 1.1: *"the unit associated with all values associated with this parameter or table column"*
- VOTable 1.4: *"The syntax of the unit string is defined in [VOUnits]"*
- VOTable 1.3: *"The syntax of the unit string is defined in [`http://cdsarc.u-strasbg.fr/doc/catstd-3.2.htx`]"*
- TAP 1.1: *"the information in the TAP_SCHEMA is equivalent to that defined by VODataService"*

$\Rightarrow$ *VOUnit syntax is not required for TAP/registry, but it's good practice*

Tools

Libraries:

- **Unity**: VOUnit parsing/validation library by Norman Gray (since 2014)
- **Ucidy**: UCD1+ parsing/validator library by Grégory Mantelet (since 2017)

Incorporated in applications (since STILTS v3.4-2, TOPCAT v4.8-2, 10/2021):

- STILTS/TOPCAT: add functions to expression language: `ucdStatus`, `ucdMessage`, `vounitStatus`, `vounitMessage`
- `taplint` TAP service validator: new stage `UUC`, performs UCD and VOUnit checking on TAP metadata declarations
- `votlint` VOTable validator: can perform UCD and/or VOUnit checking on VOTable metadata

Usage:

- I have run these on all registered table columns (via Registry metadata queries) — see below
- You can use them on your own services (via `taplint`/`votlint` or directly)

Tools Cribsheet

Use new STILTS/TOPCAT functions `ucdStatus`, `ucdMessage`, `vounitStatus`, `vounitMessage`

- Check correctness of a single UCD/VOUnit:

```
stilts calc 'ucdStatus("meta.id;meta.main")'
stilts calc 'vounitMessage("Msun/yr")'
```

- Check UCDs by column/table in a TAP service:

```
stilts tapquery tapurl='http://example.org/tap' sync=true
       adql='SELECT table_name, column_name, ucd FROM tap_schema.columns'
       ocmd='addcol UCD_STATUS ucdStatus(ucd)'
       ocmd='addcol UCD_MSG ucdMessage(ucd)'
       ocmd='select !NULL_UCD_MSG'
```

- Check UCDs by value in a TAP service:

```
stilts tapquery tapurl='http://dc.g-vo.org/tap' sync=true
       adql='select ucd from tap_schema.columns'
       ocmd='sort ucd' ocmd='uniq -count ucd' ocmd='sort -down dupcount'
       ocmd='addcol UCD_STATUS ucdStatus(ucd)'
       ocmd='addcol UCD_MSG ucdMessage(ucd)'
       ocmd='select !NULL_UCD_MSG'
```

- Check UCDs by column/table in the Registry

```
stilts tapquery tapurl=http://dc.g-vo.org/tap sync=true
       adql="SELECT ivoid, table_name, name AS colname, ucd from rr.table_column
             NATURAL JOIN rr.res_table WHERE ucd IS NOT NULL AND ivoid LIKE '%example.org%'"
       ocmd='addcol UCD_STATUS ucdStatus(ucd)'
       ocmd='addcol UCD_MSG ucdMessage(ucd)'
       ocmd='select !NULL_UCD_MSG'
```

- ... etc

Results: Statistics

UCDs

count	status
928703	OK
209230	
81763	UCD1
30619	BAD_SYNTAX
14266	UNKNOWN_WORD
12249	BAD_SEQUENCE
1565	DEPRECATED
672	NAMESPACE

```
stilts tapquery tapurl=http://dc.g-vo.org/tap
                sync=true maxrec=10000000
                adql='SELECT ucd FROM rr.table_column'
                ocmd='addcol status ucdStatus(ucd)'
                ocmd='keepcols status'
                ocmd='sort status'
                ocmd='uniq -count status'
                ocmd='colmeta -name count dupcount'
                ocmd='sort -down count'
```

Units

count	status
597947	
475792	OK
121025	BAD_SYNTAX
74096	UNKNOWN_UNIT
6305	DEPRECATED
3891	WHITESPACE
11	USAGE

```
stilts tapquery tapurl=http://dc.g-vo.org/tap
                sync=true maxrec=10000000
                adql='SELECT unit FROM rr.table_column'
                ocmd='addcol status vounitStatus(unit)'
                ocmd='keepcols status'
                ocmd='sort status'
                ocmd='uniq -count status'
                ocmd='colmeta -name count dupcount'
                ocmd='sort -down count'
```

Results: Discovering Details

UCDs:

```
stilts tapquery tapurl=http://dc.g-vo.org/tap sync=true
       adql='SELECT ucd, COUNT(*) AS ncol FROM rr.table_column GROUP BY ucd'
       ocmd='addcol status ucdStatus(ucd)'
       ocmd='addcol msg ucdMessage(ucd)'
       ocmd='select status!=\"OK\"&&status!=\"UCD1\"&&status!=\"NAMESPACE\"'
       ocmd='select ncol>20'
       ocmd='sort "status -ncol"'
       ofmt='text(maxCell=80,params=false)'
```

... see output file report-ucd.txt on wiki

Units:

```
stilts tapquery tapurl=http://dc.g-vo.org/tap sync=true
       adql='SELECT unit, COUNT(*) AS ncol FROM rr.table_column GROUP BY unit'
       ocmd='addcol status vounitStatus(unit)'
       ocmd='addcol msg vounitMessage(unit)'
       ocmd='select status!=\"OK\"&&status!=\"DEPRECATED\"'
       ocmd='select ncol>1000'
       ocmd='sort "status -ncol"'
       ofmt='text(maxCell=80,params=false)'
```

... see output file report-unit.txt on wiki

- Each takes just a few seconds
- STILTS v3.4-2 (10/2021) required

UCD Result Summary

- Bad standards text
  - Known issue: `instr.fov` required by SSA (`instr.fov` is S) → discussions on Semantics list
  - Known issue: `meta.ref.url;meta.curation` required by ObsCore/SODA (`meta.curation` is P) → ObsCore, SODA Errata
  - No others known?
- Bad sequence:
  - Secondary used as Primary: `obs.field`, `phys.atmol`, `meta.software`, `stat.fit.omc`, `obs.image`, `meta.code`, `em.*`, ...
  - Primary used as Secondary: `meta.version`, `meta.ref`
- Placeholders:
  - "??" etc.
- Deprecated:
  - `phys.mol.qn`, `time.expo`, `pos.proj`, ...
- Non-existent/made up:
  - `meta.image`, `meta.name`, `time.update`, `em.uv.fuv`, ...
- Typos:
  - `pos.bodyrc.long`, `pos.eq.de`, `src.redsfhit`, ...

UCD Results by IVOID

owner	Total	error_percent	BAD_SEQUENCE	BAD_SYNTAX	UNKNOWN_WORD	DEPRECATED	NAMESPACE	UCD1
cds.vizier	768654	0.9	7273		19	1436		12
wfau.roe.ac.uk	421858	11.4	4002	30605	13577			81607
nasa.heasarc	37476	1.2	465					2
org.gavo.dc	10484	0.3	34				49	
archive.stsci.edu	8819	0.1	1		4			46
www.plate-archive.org	3102	2.1	36		28			
lam.cesam	2893	0.2	6		1			
mssl.ucl.ac.uk	2887	0.0						3
lam.cesam.aspic	2804	0.2	6					
eso.org	2646	1.2	32					
gaia.aip.de	2446	8.9	206	12				
astron.nl	1579	2.0	6		26		378	63
helio-vo.eu	1063	0.1			1	3		
tohoku.univ.jp	1026	7.1	3		70	2		
voxastro.org	1025	1.4	4		10		2	
asu.cas.cz	1008	2.6	18		8		97	9
aip.gavo.org	740	0.3			2	2		
ia2.inaf.it	703	5.1	2		34	16		
sao.ru	605	1.3	5		3			
idoc	510	9.8	12		38	16		
purx	504	3.4	3		14		18	3
chivo	482	6.8	1		32		72	12
vopdc.obspm	425	8.9	2	2	34	12		
src.pas	405	12.3	26		24	8		

VOUnits Result Summary

Errors are diverse: probably mostly not attempting to comply with VOUnits

- CDS sexagesimal markers: `"h:m:s"`, `"d:m:s"`
- Square brackets: `[Sun]`, `[cm/s2]`, `[K]`
- Non-VOUnit exponentiation: `cm-2`, `erg/s/cm^2`
- LaTeX/HTML formatting: `nanomaggies^{-2}`, `10<sup>-17</sup> ergs/cm<sup>2</sup>/s/A`
- Unknown non-quoted units: `ppm`, `Rsun`, `dex`
- Plurals: `counts`, `mags`, `Gyrs`
- Capitalisation: `ADU`, `Arcsec`
- Non-standard unit name: `degree`, `sec`, `Kelvin`, `day`
- Whitespace: `0.01 arcsec/yr`, `10**-10 m`
- Too many slashes: `km/s/kpc`, `10nW/m2/cm`, `W/m2/Hz`
- Not really trying: `nanomaggies`, `Time[Julian Date (day)]` , `Stellar_Radius`, `(R+)`, `de Vaucouleurs magnitude fit`

Unit Results by IVOID

owner	Total	error_percent	ERROR	DEPRECATED
cds.vizier	422286	26.3	110883	60
wfau.roe.ac.uk	215386	34.0	73178	2472
nasa.heasarc	18927	47.0	8902	260
org.gavo.dc	5587	0.0		28
archive.stsci.edu	3249	83.9	2725	
lam.cesam	2466	6.0	149	1737
lam.cesam.aspic	2453	6.1	149	1737
eso.org	1826	12.5	229	2
gaia.aip.de	1488	99.5	1480	
www.plate-archive.org	1158	29.4	340	4
voxastro.org	812	75.1	610	
astron.nl	575	0.0		
tohoku.univ.jp	480	0.0		
asu.cas.cz	373	8.6	32	
sao.ru	352	17.0	60	
ia2.inaf.it	326	7.4	24	
aip.gavo.org	279	45.5	127	
idoc	256	0.0		
src.pas	225	0.0		
vopdc.obspm	198	0.0		
bsdc.icranet.org	184	43.5	80	2
chivo	170	0.0		
uni-heidelberg.de	167	13.8	23	
jacobsuni	159	0.0		

Actions

Service providers

- Use diagnostics
  - ▷ RegTAP/TAP_SCHEMA queries using new STILTS/TOPCAT functions `ucdStatus`, `ucdMessage`, `vounitStatus`, `vounitMessage`
  - ▷ New `taplint` stage `UUC`
  - ▷ New `votlint` parameters `ucd=true`, `unit=true`
- Review results
- Improve declared metadata as appropriate

Semantics WG

- Review results
- Consider if there should be changes to UCD/VOUnits standards

Standards Authors

- Make sure that mandated UCDs/Units are legal!

Operations IG

- Monitor compliance and report changes at future interops