Hackathon:
Unicode in VOTable
VOTable 1.6

Mark Taylor (Bristol)

Astro-CC Tech Forum 1
Trieste

7 October 2025

→ **Present: Mark Taylor**
Markus Demleitner
Scige Liu

$Id: univot-summary.tex,v 1.2 2025/10/07 17:43:06 mbt Exp $

Proposal for Unicode in VOTable

Problem:

- VOTable data (and metadata) can't encode non-BMP (non-ASCII?) Unicode characters
- It discusses Unicode in outdated terms (UCS-2)
  - https://wiki.ivoa.net/internal/IVOA/InterOpJune2025Apps/unicode-notes.pdf
  - https://wiki.ivoa.net/internal/IVOA/InterOpOct2014Applications/vot-unicode.pdf

Proposal:

https://github.com/ivoa-std/VOTable/pull/71 — Mark T.
(see also https://github.com/ivoa-std/VOTable/pull/68 — Markus)

- `datatype="char"`:
  - 7-bit ASCII → UTF-8
- `datatype="unicodeChar"`:
  - UCS-2 → BMP-only UTF-16 (no-op)
  - → deprecated
- `arraysize` corresponds to count of **code units** not **characters**:
  - `char`: code unit = 1 octet
  - `unicodeChar`: code unit = 2 octets

Results

Intended consequences of proposal at VOTable 1.6:

- Any Unicode character can be written in a `datatype="char"` column
  - Document encoding in `TABLEDATA`
  - UTF-8 in `BINARY`/`BINARY2`
- BMP characters (no emoji) can be written in a `datatype="unicodeChar"` column
  - Document encoding in `TABLEDATA`
  - UTF-16 in `BINARY`/`BINARY2`
  - … but don't do it, because it's now **deprecated**
- Any **legal** data in `char` or `unicodeChar` of earlier VOTable versions is **still legal** and **still has the same interpretation**
- Illegal non-ASCII (UTF-8) data in older `char` columns will probably get read as (illegally) intended

$\rightarrow$ **Yes, this looks OK**

Results

Unintended corollaries at VOTable 1.6:

- Single (scalar) `datatype="char"` columns can still only contain 7-bit ASCII
  - ▷ … since non-ASCII code points need multiple bytes in UTF-8
- Decoding string arrays (multi-dimensional `char`/`unicodeChar` arrays) requires unpacking to bytes then counting code units (i.e. counting bytes) **not** counting characters
  - ▷ This may be a bit surprising to implementors, but it's not so hard
- You can't specify a string column with a fixed number of characters
  - ▷ you have to specify the length of the UTF-8/UTF-16 **serialization** instead
  - ▷ … unless e.g. you know the column is 7-bit ASCII
- String truncation is not straightforward
  - ▷ Overlength strings may need to be truncated to fit in fixed-`arraysize` strings/array elements
  - ▷ Such truncation has to be done carefully (not in the middle of a multi-octet UTF-8 character)

→ We can cope with these
They are/will be spelled out in the text

Interesting Facts

In all VOTable versions $\leq$1.5:

- `datatype="char"` must be 7-bit clean
- `datatype="unicodeChar"` must be non-BMP

... everywhere, including

- `BINARY`/`BINARY2`
- `TABLEDATA`
- `PARAM`
- and even `INFO`:

  VOTable 1.5 sec 4.8: *"The INFO element is a PARAM element restricted to be of type string (i.e. datatype="char" and arraysize="*" are implied)."*

- Consequences:
  - ▷ My prototype VOTable 1.6-compatible STIL now handles these correctly
    - votlint reports an error
    - STIL maps non-ASCII `char` bytes/non-BMP `unicodeChar` pairs to '?'
  - ▷ Should it do that? There's no reason for the restriction in `PARAM`/`INFO`.
  - ▷ Should we have an Erratum? $\rightarrow$ **yes**

$\rightarrow$ **Propose Erratum:**
All Unicode characters permitted in INFO/PARAM values for all current and previous VOTable versions

Other VOTable 1.6 questions

- Do we go to VOTable 1.6 to accommodate Unicode changes
  - Now? Soon? Sometime? → **yes**
- Who else do we need to engage for Unicode changes?
  → **Astropy, Apps WG, Tom Donaldson; others?**
  → **Rubin already approved PR**
- Other items for VOTable 1.6
  - remove Appendix A issue #53
  - some editorial things
  - MIME type parameter `content=datalink` (issue #26/#15)
  - arrays of variable-length strings, by `xtype` (DALI) or some other way? → **DALI PR#66**
  - ... ?

→ Let's get going on VOTable 1.6:
Report these conclusions at Görlitz
If no objections move towards 1.6 WD→PR
(editor Tom D. or Mark T.)