

VOParquet Update

Mark Taylor (Bristol)

Applications WG
IVOA Interop
College Park, MD, USA

3 June 2025

`$Id: voparquet.tex,v 1.12 2025/06/03 10:49:16 mbt Exp $`

Outline

- VOParquet Note status and content
- Known VOParquet implementations
- STILTS (VO)Parquet utilities
- Next steps

VOParquet Status and Content

Note **Parquet in the VO** published 16 Jan 2024

- Compact document (7 page PDF)
- Defines how to store VOTable semantic metadata in a Parquet file
- Uses two IVOA-defined keys in Parquet footer metadata structure

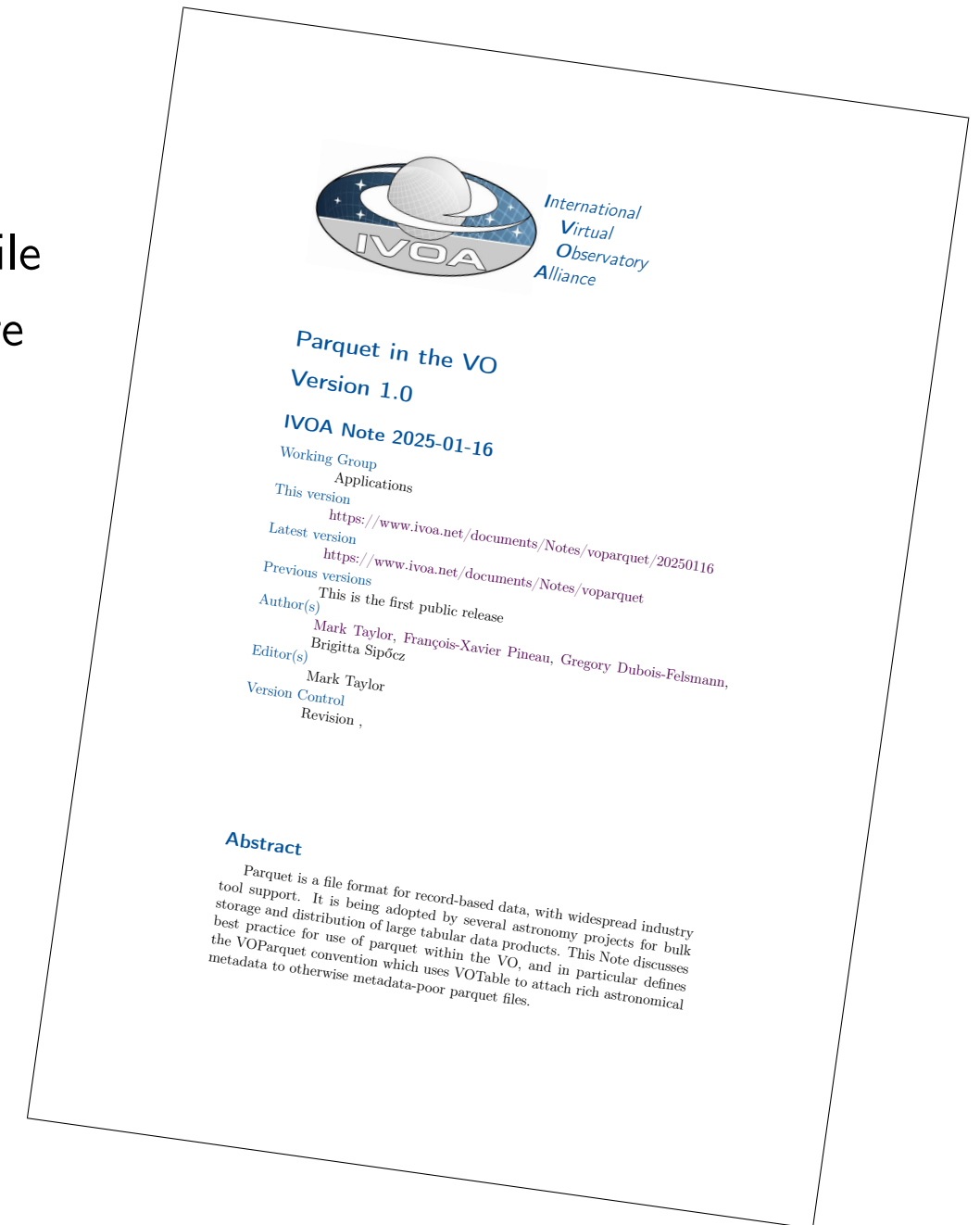
`IVOA.VOTable-Parquet.content: [metadata-only VOTable text]`

`IVOA.VOTable-Parquet.version: 1.0`

- Discusses possible data/metadata mismatch issues
- Minimum rules to enable data/metadata interoperability
- Leave open details we might want to change later

Result: best of both worlds

- **Parquet data** — efficient storage/processing, industry standard
- **VOTable metadata** — IVOA standards
- VOParquet-aware readers benefit from rich metadata
- Non-VO parquet readers completely ignore VO additions



Implementations

Several implementations can read/write VOParquet-compliant files:

- TOPCAT/STILTS/STIL applications and Java I/O library
 - ▷ In public release since March 2025
- ESAC Science Data Center archives
 - ▷ Working internally, but files not yet public
 - ▷ Will be used for Gaia DR4 bulk downloads
- OpenCADC TAP service
 - ▷ In progress?
- VOTableR R library
 - ▷ Working
 - ▷ Plans to use for 4MOST/WEAVE output

Know others?

Update list on wiki page [ParquetInAstronomy](#)

See also HATS

- HATS Note recommends use of VOParquet
- See Neven's talk later

STIL I/O Handlers

STIL I/O handlers (hence TOPCAT and STILTS) can read/write VOParquet-compliant files

Behaviour is somewhat configurable:

- **Output handler** config options:

`votmeta = true|false` — do/don't write VOTable metadata

`compression = uncompressed|snappy|zstd|gzip|lz4_raw` — column compression format

`usedict = true|false|null` — whether to attempt dictionary encoding

`kvmap = key1:value1;key2:value2;` — doctor footer key-value map (*pre-release only*)

— could add more options on request

- **Input handler** config options:

`votmeta = null|true|false` — use/ignore attached VOTable metadata

`votable = <filename-or-url>` — act as if given VOTable is attached (debug/devel)

Example:

```
stilts tpipe in=x.vot ofmt='parquet(compression=zstd,usedict=false)' out=x.parquet
```

STIL Parquet I/O Notes

Compression formats:

- Supported: Snappy, ZSTD, GZIP, LZ4_RAW
- Not supported: Brotli
- Snappy, ZSTD rely on JNI (platform-dependent) code
 - ▷ Large jar files (implementations for many architectures)
 - ▷ Cross-platform support not guaranteed
- Supported compression formats may be an interoperability issue?

Parquet I/O not currently bundled with all TOPCAT/STILTS releases because of size

- Parquet requires `topcat-extra.jar` (58 Mb) not `topcat-full.jar` (31 Mb)
- May reconsider if parquet usage becomes widespread

parqlint VOParquet Validator

STILTS `parqlint` command

- Checks parquet files against the VOParquet convention
 - ▷ Checks whether a VOTable is present under `IVOA.VOTable-Parquet.content` key in footer
 - ▷ Checks `IVOA.VOTable-Parquet.version` key contains “1.0”
 - ▷ Checks no other `IVOA.VOTable-Parquet` keys are present
 - ▷ Performs full validation on VOTable (à la `votlint`)
 - ▷ Checks that VOTable has no `DATA` element
 - ▷ Checks consistency between VOTable column metadata and parquet data column types
- Use it if you’re publishing VOParquet files!
- Bugs possible — talk to me if you don’t understand/agree with output

Example:

```
% stilts parqlint sky2.parquet
ERROR: IVOA.VOTable-Parquet.content present but IVOA.VOTable-Parquet.version missing
ERROR: [JXC] cvc-complex-type.4: Attribute 'name' must appear on element 'PARAM'.
ERROR: [JXC] cvc-complex-type.4: Attribute 'name' must appear on element 'PARAM'.
WARNING: Column gmag (#5): parquet/VOTable type mismatch, Float != Double
WARNING: Column rmag (#6): parquet/VOTable type mismatch, Float != Double
```

parqlook Parquet Viewer Utility

STILTS `parqlook` command

- Reports information about parquet files
 - ▷ parquet schema
 - ▷ key-value metadata items in footer
 - ▷ data block information
 - ▷ column chunk information and stats
 - ▷ VOTable text if present
- Can report all information or just some items
- Other available tools do similar things, but this may be convenient if you have STILTS

```
% stilts parqlook sky1.parquet
```

```
Parquet Schema:
```

```
message table {
  optional double ra;
  optional double dec;
  optional double l;
  optional double b;
  optional float gmag;
  optional float rmag;
  optional double b_r;
}
```

```
Key-Value Metadata:
```

```
Description: Simulated star catalogue, based on Gaia EDR3
HealpixLevel: 5
IVOA.VOTable-Parquet.content: <1375 chars>
name: SimulatedSky-100
writer.model.name: STIL
IVOA.VOTable-Parquet.version: 1.0
```

```
Data Blocks:
```

#	ordinal	rowIndexOffset	rowCount	compressedSize	totalByteSize
0	0	100	5031	5031	

```
Column Chunks:
```

```
Block 0 (100 rows):
```

```
-----
```

#	path	type	count	nullCount	uncompressedSize	compressedSize	codec	encodings	minMax
ra	DOUBLE	100	0	833	833	UNCOMPRESSED	RLE,PLAIN,BIT_PACKED	1.9054392576217651/35	
dec	DOUBLE	100	0	833	833	UNCOMPRESSED	RLE,PLAIN,BIT_PACKED	-80.10884857177734/75	
l	DOUBLE	100	0	833	833	UNCOMPRESSED	RLE,PLAIN,BIT_PACKED	-150.43421936035156/1	
b	DOUBLE	100	0	833	833	UNCOMPRESSED	RLE,PLAIN,BIT_PACKED	-67.16999053955078/77	
gmag	FLOAT	100	0	433	433	UNCOMPRESSED	RLE,PLAIN,BIT_PACKED	15.876356/22.393879	
rmag	FLOAT	100	0	433	433	UNCOMPRESSED	RLE,PLAIN,BIT_PACKED	14.146327/22.20099	
b_r	DOUBLE	100	0	833	833	UNCOMPRESSED	RLE,PLAIN,BIT_PACKED	-0.5077816843986511/4	

```
VOParquet VOTable:
```

```
<?xml version='1.0'?>
<VOTABLE version="1.4"
  xmlns="http://www.ivoa.net/xml/VOTable/v1.3">
<!--
! VOTable written by STIL version 4.3-2 (uk.ac.starlink.parquet.StarParquetWriter$1)
!-->
<RESOURCE>
<TABLE name="SimulatedSky-100" nrows="100">
<DESCRIPTION>
Simulated star catalogue, based on Gaia EDR3
</DESCRIPTION>
<PARAM datatype="int" name="HealpixLevel" value="5">
<DESCRIPTION>Healpix level of aggregated statistics</DESCRIPTION>
</PARAM>
```

VOParquet Hacking with STILTS

STILTS tools `parqlint`, `parqlook` and parquet I/O handler options can be useful to get VOParquet metadata right:

- Check a VOParquet file `x.parquet`:

```
% stilts parqlint in=x.parquet
```

- Maybe see some problems
- Extract the VOTable metadata into a file `x.vot`:

```
% stilts parqlook in=x.parquet items=votable >x.vot
```

- Now look at and maybe edit the VOTable in `x.vot` to diagnose issues, write adjusted VOTable metadata in `x1.vot`
- Try the validation again, with the VOTable metadata taken from the edited file `x1.vot` instead of the embedded VOTable:

```
% stilts parqlint in=x.parquet votable=x1.vot
```

- Iterate... if it's OK then write a new parquet file with adjusted VOTable metadata (*Note: STILTS pre-release needed for this step*)

```
% stilts tpipe in=x.parquet ifmt='parquet(votmeta=false)  
ofmt='parquet(votmeta=false,kvmap=IVOA.VOTable-Parquet.version:1.0;IVOA.VOTable-Parquet.content:@x1.vot)'  
out=x1.parquet
```

- Check the new file is OK

```
% stilts parqlint in=x1.parquet
```

Next Steps

- Any implementation feedback on Note?
- VOParquet **Note** → **Endorsed Note**?
- Maybe a more general Note on using VOTable with metadata-poor formats
 - Parquet, FITS, others?