CCLRC / Rutherford Appleton Laboratory
Particle Physics & Astronomy Research Council
Starlink Project
Catalogue Formats 1.1

A C Davenhall (acd@roe.ac.uk)
7 February 2001

Catalogue Formats for Remote Information Retrieval

Abstract

In this talk I'll discuss the formats and particularly metadata required for retrieving subsets selected by some search process from remote catalogues.

Contents

1 Introduction **1**

2 Model for Astronomical Catalogues **1**

3 Tab-Separated Table (TST) Format **5**

4 FITS Tables **5**

5 XML **7**

6 Summary **7**

1 Introduction

In this talk I'll discuss the formats and particularly metadata required for retrieving subsets selected by some search process from remote catalogues. An example is overlaying catalogue objects on an image displayed with GAIA or SkyCat. There is nothing particularly new here; it has been done for some time with existing software which is certainly not a GRIDS system. The user issues a query to a remote catalogue and the rows which satisfy it are retrieved from the remote catalogue. In a GRIDS system details of the catalogue locations etc. might be hidden and the query might involve more than one catalogue. However, the same basic arrangement applies: rows selected according to some query are returned.

I'll discuss: what formats and standards we've got at the moment for such catalogue subsets and what additional standards may be required. The talk is very much in the spirit of a workshop rather than a conference, discussing 'work in progress', technical details etc. The good news, as it turns out, is that many of the required standards and formats are already in place. FITS, of course, is the standard for exchanging astronomical data, with ASCII and binary FITS tables for catalogues. Much work on catalogue standards has been done at the CDS and elsewhere. In a few instances I'll suggest a few additional standards that might be needed. These proposals are a collaboration with Clive Page, Sally Hales and others.

2 Model for Astronomical Catalogues

I'll start by discussing an abstract model for astronomical catalogues. It is more or less the model which Clive Page and I came up with about eight years ago for the Starlink package CURSA. It is intended for whole catalogues, but is equally valid for subsets extracted from them (which are just small catalogues). In this model a catalogue comprises:

- table of values,
- indices,
- column definitions,
- parameter definitions,
- description.

The table of values is just the fields of the rows and columns of the catalogue. Indices typically permit fast searching of large catalogues and are not really relevant for small subsets extracted from a catalogue. The description is a human-readable text describing the catalogue. The column definitions are the information describing each column.

Table 1 is taken from one of the CURSA manuals. It lists the information available for each column. Some of the items are specific to CURSA and an asterisk marks the items that I think are important. The COMM attribute is comments describing the column (for a human). In CURSA they are limited to a single line of text. Ideally, perhaps we would like a description of unlimited length, with hyperlinks to the catalogue description and thence to external Web pages, papers in the ADS etc. However, I'm a pragmatic soul and will settle for a single line of text.

Attribute	Name	Data type	Mut--able	Mand--atory	Default	
Name	NAME	_CHAR		•		*
Genus	GENUS	_INTEGER			physical: `CAT__GPHYS`	
Expression	EXPR	_CHAR			' '	
Data type	DTYPE	_INTEGER		•		*
Character size	CSIZE	_INTEGER			20†	*
Dimensionality	DIMS	_INTEGER			scalar: `CAT__SCALR`	*
Size§	SIZE	_INTEGER			1	*
Null or locum	NULL	_INTEGER			HDS: `CAT__NULLD`	
Exception values	EXCEPT	_CHAR			' '	
Scale factor	SCALEF	_DOUBLE			1.0D0	
Zero point	ZEROP	_DOUBLE			0.0D0	
Order	ORDER	_INTEGER			none: `CAT__NOORD`	
Units	UNITS	_CHAR	•		' '	*
External format	EXFMT	_CHAR	•		varies with data type	
Preferential display	PRFDSP	_LOGICAL	•		true	
Comments	COMM	_CHAR	•		' '	*
Modification date	DATE	_DOUBLE	•		0.0D0	

† The size of character strings; other data types have CSIZE = 0.

§ SIZE is a single-element array, not a scalar.

Table 1: Attributes of columns

Attribute	Name	Data type	Mand--atory	Default	
Name	NAME	_CHAR	•		*
Data type	DTYPE	_INTEGER	•		*
Character size	CSIZE	_INTEGER		`CAT__SZVAL`†	*
Dimensionality	DIMS	_INTEGER		scalar: `CAT__SCALR`	*
Size§	SIZE	_INTEGER		1	*
Units	UNITS	_CHAR		' '	*
External format	EXFMT	_CHAR		varies with data type	
Preferential display	PRFDSP	_LOGICAL		true	
Comments	COMM	_CHAR		' '	*
Value	VALUE	varies		zero or ' '	*
Modification date	DATE	_DOUBLE		0.0D0	

† The size of character strings; other data types have CSIZE = 0.

§ SIZE is a single-element array, not a scalar.

Table 2: Attributes of parameters

Table 2 is also taken from one of the CURSA manuals. It lists the information available for each parameter. A CURSA parameter is a single item of information which pertains to the entire catalogue, eg. the epoch and equinox of the celestial coordinates. In CURSA almost as many details are stored for each parameter as for each column, arguably more than are strictly necessary. Do we really need the format or data type? But they do reduce the scope for ambiguity. It does, however, seem a good idea to store the units: in the context of this talk it allows a user to interpret the values properly and more generally is also important for long-term curation.

Perhaps surprisingly, all the above is still not quite all that is required to allow a client or browser to automatically interpret a catalogue. Consider some of the common operations carried out on remote catalogues:

- searching to find the objects in a particular patch of sky,
- producing finding charts, image overlays and similar plots,
- pairing or joining two catalogues to find the objects in common,
- converting the celestial coordinates to a new coordinate system (perhaps to a new epoch).

These operations all involve celestial coordinates. Celestial coordinates are ubiquitous in astronomical catalogues and central to the way they are used. We need some means to automatically identify the columns of celestial coordinates in a catalogue. This mechanism which would allow a client or browser to proceed with searches, sky-plots etc. automatically without user intervention. It should also be remembered that a celestial coordinate is more than just a Right Ascension and Declination. The full components of a celestial component are:

- Right Ascension,
- Declination,
- proper motion in Right Ascension,
- proper motion in Declination,
- parallax,
- radial velocity,
- plus equinox, epoch and name of the system.

There is a similar argument that the catalogue should provide hints on how objects in it are to be plotted in finding charts etc. There are various options in common use, including: circles (and other symbols) scaled according to magnitude or flux and an ellipse approximating to some isophote on the sky.

Having outlined the requirements, I'll move on to discuss the formats available.

```
Simple TST example; stellar photometry catalogue.

A.C. Davenhall (Edinburgh) 26/7/00.

Catalogue of U,B,V colours.
UBV photometry from Mount Pumpkin Observatory,
see Sage, Rosemary and Thyme (1988).

ra_col: 1
dec_col: 2

# Start of parameter definitions.
EQUINOX: J2000.0
EPOCH: J1996.35
# End of parameter definitions.


Id<tab>ra<tab>dec<tab>V<tab>B_V<tab>U_B
--<tab>--<tab>---<tab>-<tab>---<tab>---
Obj. 1<tab> 5:09:08.7<tab> -8:45:15<tab>  4.27<tab>  -0.19<tab>  -0.90
Obj. 2<tab> 5:07:50.9<tab> -5:05:11<tab>  2.79<tab>  +0.13<tab>  +0.10
Obj. 3<tab> 5:01:26.3<tab> -7:10:26<tab>  4.81<tab>  -0.19<tab>  -0.74
Obj. 4<tab> 5:17:36.3<tab> -6:50:40<tab>  3.60<tab>  -0.11<tab>  -0.47
     .
     .
     .
```

Figure 1: A simple tab-separated table. Note that in a tab-separated table the list of column names, sequences of dashes and fields in the table are separated by tab characters. In this figure tab characters are indicated by '`<tab>`'.

3 Tab-Separated Table (TST) Format

Figure 1 shows an example TST format catalogue. TST catalogues are basically ASCII text files, with tab characters to separate fields in the table (tab characters are shown as '`<tab>`' in the figure). They were invented, as far as I know, by Allan Brighton and his colleagues at ESO, for use in SkyCat and subsequently GAIA. They are now also used in a few other packages (including CURSA). The documentation is a bit scattered, and I've tried to give a definitive description in SSN/75: *Writing Catalogue and Image Servers for GAIA and CURSA* (contact me if you have difficulty obtaining a copy). The TST format was originally intended for the specific problem of catalogue overlays of images displayed with SkyCat, not as a general format for catalogue exchange. It is somewhat short of metadata, in particular the only information for a column is its name, which is not adequate. The format does, however, have a mechanism for indicating which columns are the celestial coordinates; the parameters `ra_col` and `dec_col`, plus special rules for the units of columns of celestial coordinates. There is also a similar mechanism for specifying plotting symbols.

4 FITS Tables

It would be taking coals to Newcastle to describe the FITS table at this meeting, so I won't. There are ASCII and binary variants. An arbitrary amount of metadata can be included by using keywords.

Catalogue parameters are naturally implemented as keywords. However keywords only have:

name value comments

In particular there is no standard way to store the units of keywords. I may be worrying unnecessarily; FITS tables have been used successfully for years. But it would be nice to have a robust way of storing the units.

The basic keywords for columns include the most important information: `TTYPEn` etc. However, catalogue manipulation software can write additional column keywords if it so wishes; I did this to an extent with CURSA.

In order to interpret celestial coordinates: there are recommended units for storing angles in FITS files (Table 3). As long as columns stick to these units a client should be able to interpret the column correctly. However, these units are not always adhered to: a common, but particularly pernicious practice is to store the degrees (or hours) minutes and seconds as separate columns, which makes it virtually impossible to interpret the coordinates in a general-purpose program.

It is often desirable to display a celestial coordinate in a sexagesimal format, even if it is not stored in this way. We are working on a proposal for a way of allowing sexagesimal display to be specified by allowing special forms in the `TDISPn` keyword. This work is still in progress.

Surprisingly, there is no way to automatically identify the celestial coordinates in a FITS table. Again we are working on a proposal for a set of additional keywords to specify the columns to be used (Table 4) and thinking along similar lines for specifying the plotting symbols.

TUNITSn value	Units
rad	radians
deg	degrees
arcmin	minutes of arc
arcsec	seconds of arc
mas	milli-seconds of arc
h	hours
min	minutes of time
s	seconds of time

Table 3: Angular and time units recommended in the first FITS WCS paper. Only time units which are also used to measure angles are included. Compound units are also permitted.

Keyword	Description	Mandatory?	Suggested column name
RACOL	Right Ascension	•	RA
DECCOL	Declination	•	DEC
PMRACOL	Proper motion in Right Ascension		PMRA
PMDECOL	Proper motion in Declination		PMDE
PLXCOL	Parallax		PLX
RVCOL	Radial velocity		RV

Table 4: Keywords to specify columns of celestial coordinates

5 XML

XML (eXtensible Mark-up Language) is likely to become important because standard, commercial browsers can make some sense of it. It seems a practical way to represent fragments of astronomical catalogues. I know there has been some work in this area. I attempted to draw up a DTD for an astronomical catalogue, deliberately without looking at any of the previous work; it seemed perfectly practicable:

- DTD for an astronomical catalogue,
- example of a very simple catalogue marked up with this DTD.

6 Summary

- Metadata is important; it is hardly necessary to make this point to this audience, though perhaps it is something that we need to emphasise to the wider community.
- FITS tables provide most of the facilities required.
- Better compliance to the existing recommendations of the units for angles is required.
- A few additional keywords are needed to allow the automatic identification of celestial coordinates.
- XML is a practical way of representing fragments of astronomical catalogues and is likely to become important.