Astronomical Data Analysis Software and Systems XII
ASP Conference Series, Vol. 295, 2003
H. E. Payne, R. I. Jedrzejewski, and R. N. Hook, eds.

Metadata for the VO: The Case of UCDs

Sébastien Derriere, François Ochsenbein, Thomas Boch

CDS, Observatoire Astronomique de Strasbourg, 11 rue de l'Université, F-67000 Strasbourg, France

Guy T. Rixon

Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge. CB3 0HA, U.K.

Abstract. The UCDs (Unified Content Descriptors) were first developed in the ESO/CDS data mining project, to describe precisely the contents of the individual fields (columns) of tables available from a data center. They have been used to describe the content of the 10^5 columns available in the different VizieR tables. Owing to the wide diversity and high heterogeneity of table contents, UCDs constitute an excellent starting point for a hierarchical description of astronomy, for general data mining purposes. We present different applications of UCDs: selection of catalogues, based on their content; identification of catalogues having similar fields; automated data conversion allowing direct comparison of data in cross-identifications. The compatibility of UCDs with semantic descriptions developed in other contexts (data models for space-time coordinates or image datasets) will also be addressed.

1. Introduction

Astronomical tables can come from many different sources, and the original descriptions are therefore very heterogeneous. Automated processing of the contents of these datasets, which is one of the Virtual Observatory (VO) applications, requires a uniform description for the catalogues (with standardized metadata).

The UCDs (Unified Content Descriptors), first developed in the ESO/CDS data mining project (Ortiz et al. 1999), are metadata describing precisely the contents of the individual fields (columns) of tables available from a data center. They have been applied to describe the content of the 10^5 columns available in the different VizieR tables (Ochsenbein, Bauer & Marcout 2000).

Some tools using UCDs have been developed and are available online: `http://vizier.u-strasbg.fr/UCD/`.

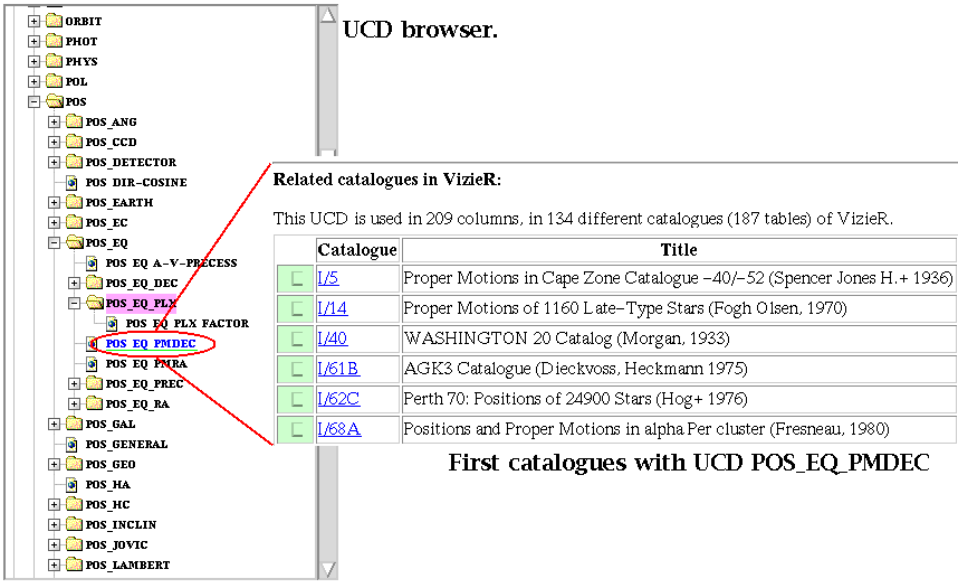

Figure 1. The UCD browser, on the left, is used to locate relevant UCDs in the hierarchical structure. For each UCD, the list of VizieR catalogues containing this UCD in at least one field can be displayed and queried.

2. Usage of UCDs

2.1. Browsing the UCD Tree

The UCDs consist of a 4-level hierarchical structure, with approximately 1500 elements. Different branches of the tree correspond to different domains of the semantic classification (e.g., time, position, instrument).

A tool has been developed to visualize and explore the tree (Figure 1). A javascript and an applet version of the browser are available. The presentation of the tree is similar to a file system browsing engine, with folders being nodes of the UCD tree and documents being the UCD leaves, actually describing the catalogue columns.

Clicking on a leaf gives access to:

- a definition of the corresponding UCD;
- statistics on column labels and units associated to this UCD (Figure 2);
- usage statistics for this UCD in VizieR (catalogues and tables where it occurs).

2.2. Data Validation

The wide heterogeneity of the original description of astronomical data is clearly visible when making statistics on the column names and units used to represent a single physical quantity (Figure 2). These statistics help pointing out possible errors in the catalogue description, or in the UCD assignation, and are thus useful for data validation.

UCD **POS_EQ_PMDEC** represents: **Proper Motion in Declination (pmdec)**

Statistics for this UCD:

Column names and units associated to UCD: POS_EQ_PMDEC
(there are 14 different column names and 16 different units).

Frequency: column name	Frequency: unit
185 pmDE	91 mas/yr
6 CvarDE	38 arcsec/yr
5 pmDE2000	21 arcsec/ha
3 pmDE1950	18 arcsec/a
1 pmDER	13 mas/a
1 pmDE-ACr	9 10mas/yr
1 pmDE-NPM1	6 10-4arcsec/yr
1 pmDEJ	2 10-5arcsec/yr
1 pmDE-NPM1r	2 10mas/a
1 pmDE-AC	2 10-2arcsec/yr
1 pmDE-HIP	2 carcsec/yr
	1 marcsec/a
	1 0.01arcsec/yr
	1 mag/yr
	1 10uas/yr
	1 ---

Figure 2. Example of statistics on the different column names and units used in all VizieR tables for one UCD.

2.3. Selection of Catalogues

One of the most important use of UCDs is that they allow to select catalogues which exactly contain a given measurement. Instead of searching all the "infrared" catalogues for a K-band magnitude, all catalogues with a Johnson K magnitude can be retrieved instantly.

This selection can be done with the browser (see Figure 1). It is also possible to translate plain text into relevant UCDs. One provides one or several terms to describe in natural language the desired quantity (e.g., 'proper motion'). The answer is a list of corresponding UCDs, tentatively ordered by relevance. These can be used to select the relevant catalogues.

2.4. Automated Data Conversion

If two fields in two tables are described by the same UCD, these fields can be compared because they contain the same quantity. Automated data conversion can then be applied if these fields are expressed in different units (Figure 3).

2.5. Finding Similar Catalogues

Because UCDs precisely describe the contents of catalogues, they can be used to find similar catalogues. Given a reference catalogue, the list of UCDs which are present in this catalogue is used as criteria to perform a search among all other catalogues: similar catalogues are those that will have many UCDs in common with the reference one.

I/146/ppm1 Positions and Proper Motions – North (Roeser+, 1988)
Catalogue PPM–North

RAJ2000	DEJ2000	pmRA	pmDE
"h:m:s"	"d:m:s"	s/yr	arcsec/yr
17 57 24.373	+04 36 09.20	–0.0014	0.032

I/239/tyc_main The Hipparcos and Tycho Catalogues (ESA 1997)
The main part of Tycho Catalogue

RAhms	DEdms	RA(ICRS)	DE(ICRS)	pmRA	pmDE
		deg	deg	mas/yr	mas/yr
17 57 24.42	+04 36 09.0	269.35174824	4.60249678	41.60	37.50

I/239/tyc_main *converted columns :*

recno	pmDE
	arcsec/yr
35715	0.0375
35741	

Figure 3. Example of automated conversion for columns with the same UCD.

3. Possible Evolution of UCDs

Suggestions have been made to improve the current structure of UCDs. The evolution towards an "atomic" rather than hierarchical structure is studied. UCDs could be built by assembling atomic elements (principal nouns, adjectives, complementary nouns) selected among a predefined set of standard atoms. This scheme allows more flexibility in defining new UCDs, avoids dispersion of related quantities in different branches of the tree, and describes the data more completely.

Examples of combinations of atoms (compared to current UCDs):

- angle/declination (current UCD is POS_EQ_DEC);
- length/wavelength/johnson-V (central wavelength of the band, no UCD);
- length/wavelength/extent/johnson-V (bandwidth of the band);
- energy-flux-density/uncertainty/johnson-V (current UCD is ERROR).

4. Conclusions

UCDs are currently used in VizieR to describe the semantics of astronomical content. They offer new ways of selecting relevant datasets, and enable cross catalogue/archive interoperability. Owing to the wide diversity of table contents, UCDs constitute an excellent starting point for a hierarchical description of astronomy, for general data mining purposes. An improved structure relying, for example on atomic keywords, could provide building blocks for the development of astronomical ontologies.

References

Ortiz, P. et al. 1999, in ASP Conf. Ser., Vol. 172, Astronomical Data Analysis Software and Systems VIII, ed. D. M. Mehringer, R. L. Plante, & D. A. Roberts (San Francisco: ASP), 379

Ochsenbein, F., Bauer, P. & Marcout, J. 2000, A&AS, 143, 23