Astronomical Data Analysis Software and Systems XIII
ASP Conference Series, Vol. 314, 2004
F. Ochsenbein, M. Allen, and D. Egret, eds.

UCD in the IVOA context

Sébastien Derriere[1], Norman Gray[2], Jonathan McDowell[3], Robert Mann[4], François Ochsenbein[1], Pedro Osuna[5], Andrea Preite Martinez[1,6], Guy Rixon[7], Roy Williams[8]

[1] *CDS, Observatoire de Strasbourg, France*

[2] *University of Glasgow, UK*

[3] *CfA Harvard, USA*

[4] *Royal Observatory Edinburgh, UK*

[5] *European Space Agency*

[6] *IASF, Italy*

[7] *Institute of Astronomy, Cambridge, UK*

[8] *CACR, Caltech, Pasadena, USA*

Abstract. UCDs (Unified Content Descriptors) are metadata that have been used in the VizieR catalogue service since they were first developed in the ESO/CDS data mining project. Because the catalogues currently described by UCDs cover many different domains, UCDs can be used to describe a large fraction of astronomical concepts. The different Virtual Observatory (VO) projects, collaborating in the IVOA (International Virtual Observatory Alliance) have therefore decided to support UCDs, and to collaborate to define a new set of standard metadata. This new set of UCD will be more coherent and complete than the old one. It uses a new syntax, and a list of atomic keywords from which UCDs are built. The sharing of a core set of metadata is very important to ensure interoperability between the various VO elements. We present how this new list of UCDs has been defined, and how it will be expanded in the future. We also present the mechanisms to assign UCDs to (possibly large) datasets, check for validity of UCDs, and make suggestions for including new concepts in the standard core of UCDs. We also discuss several fields of the VO where UCDs will be used, what kind of applications of UCDs are foreseen, how they can be used and how software using UCDs can be designed.

1. From UCD1 to UCD2

The UCD (Unified Content Descriptors) have been initially developed at CDS, in order to describe homogeneously the contents of the VizieR tables (Ortiz et al.,

1999). This first set of terms (hereafter UCD1) seemed to be a very good starting point for a standard description of astronomy, but had some inconveniences (lack of flexibility, missing concepts) which hampered broader use in the emerging VO projects.

Some tools have already been built to demonstrate the possible applications of UCDs in the VO context (Derriere et al., 2003). In order to generalize the usage of UCDs in the different VO components, it has been decided to build a new set of rules and standard terms, named UCD2. This new set will be validated by a committee at the IVOA level. The list of UCD2 terms will become the reference for use by any VO application.

2. A Controlled Vocabulary

Defining a controlled vocabulary, and getting common agreement is a nearly impossible task, mainly because different people have different views of the same things (e.g. the FITS keyword NAXIS1 can be described as the "number of pixels along the x image axis", or as the "size of the detector").

The main role of UCDs is to describe quantities that are used in practice (and represented as numbers, character strings, ...) at some level between the fuzziness of natural language and the accuracy of attributes of data models. The objective is to ensure interoperability between services in the VO, by the use of a controlled vocabulary that can be interpreted by machines and still readable (and writeable) by humans.

We have tried to reconcile a bottom-up approach (describe what is found in existing datasets) with the ontology-related vision (cf. section 4.). The resulting vocabulary is a compromise, intended to be flexible enough, but less ambiguous than natural language. Of course, the resulting list of words, as any standardization process, is somehow subjective (for example, only one word is kept when there are synonyms of some concept). But words are created to describe the quantities used in practice, taking into account the UCD1 (i.e. the contents of astronomical tables), FITS keywords, etc., so they try to cover the semantic field.

3. Syntax of UCD2

The syntax of UCD1 reflected the underlying hierarchical organization, with different levels separated by underscore characters (e.g. POS_EQ_RA_MAIN). In UCD2, this would be written `ivoa:pos.eq.ra;meta.main`.

There are three reserved characters in UCD2:

- the colon `:` is used to separate the optional namespace from words;
- the semicolon `;` is used to separate composed words;
- the period `.` is used to concatenate atoms to build composed words.

The semicolon is reserved for possible future usage of an optional namespace. The use of a leading namespace should be avoided as far as possible. Standard UCD2 (defined by the IVOA board) can be written with the `ivoa:` namespace, but it is recommended not to write it (e.g. `pos.eq.ra;meta.main`).

A UCD2 is composed of several *composed words*, the most important (carrying most of the meaning) being the first one, that is called *primary word*. The primary word describes a property; it is a first-order description of the quantity. Following words, if present, give additional precision: they can add precision to either the concept to which the property refers, or the context in which it was measured.

The composed words are made of *atoms* separated by periods. They are arranged, only for convenience, in a hierarchical tree. This structure does not imply the existence of an underlying model.

4. UCD and Ontologies

Currently, UCD are not an ontology. In the future, a project called UCD3 will apply knowledge representation methods developed outside astronomy to the controlled vocabulary.

However, to ease the transition between UCD2 and 3, we have tried in the definition of UCD2 to take into account the ideas of concepts, properties, classes and instances.

- The concepts are the top-level elements of an ontology: they are analogous to classes in OO-programming.
- A concept has properties (or slots, or parameters).
- The concept has instances (which are still something abstract to the computer).
- Instances of properties are the real data, that computers deal with.

The thumb rule is that *primary words* refer to *properties*. For example:

- phys.temp;instr.telescope represents the temperature of a telescope (the property being here phys.temp);
- stat.error;pos.eq.ra represents the error on the right ascension.

5. List of Standard Terms

The recommendations for the use of UCD2, and the list of standard terms will be made available on the IVOA web site[1].

The list of proposed roots for the tree of standard composed words currently contains: *meta* (for metadata-related quantites), *instr* (instrument), *obs* (observation), *phot* (photometry), *src* (source), *stat* (statistical), *em* (electromagnetic)...

The *em* branch is a special case: it has been created mainly to indicate in which part of the electromagnetic spectrum a measurement is made. Specific words have been created to identify the frequently used bandpasses and filters. The electromagnetic spectrum has been divided into 8 domains (radio, submm, IR, optical, UV, EUV, X-ray, gamma), in agreement with other IVOA representations. Further divisions are made to define large bands classically used in the different domains. For photometric quantities, the property that

[1] `http://www.ivoa.net/`

is measured is phot.mag, or phot.flux or phot.count, and possible valid UCD2 could be `phot.mag;em.opt.V` or `phot.count;em.x-ray.hard`.

There will be a committee in charge of studying the suggestions for additions of new terms in the list of standard UCDs.

6. Tools and Services

A set of tools for UCD2 will be progressively made available (similarly as for UCD1):

- Tools to list and browse the existing list of UCD2, with their definitions. These could be available as Web Services;
- A validation/resolver service, that will check for the validity of a UCD, and provide the corresponding definition;
- A search engine/assignment tool, designed to suggest the appropriate UCD corresponding to a description in natural language;
- An interface to send feedback, and suggest new UCDs to the board in charge of maintaining the UCD2 standard list.

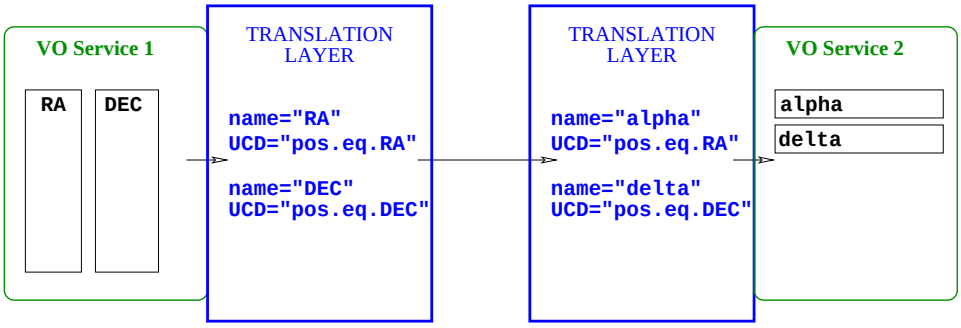

Figure 1. Services use UCDs to exchange information. A translation layer is used to interpret the internal description in terms of UCDs.

It is important to note that data providers do not need to change the internal description of their existing databases to use UCDs. Interoperability only requires a translation layer able to associate UCDs to parameters used internally. This layer is used to make the conversion from (resp. to) UCDs to (resp. from) the internal description, as shown on Fig. 1, in the case of two services exchanging data.

This translation layer only needs to be built once, and the assignment tool can be used to facilitate this step.

References

Derriere, S. et al. 2003, in ASP Conf. Ser., Vol. 295, ADASS XII, ed. H. E. Payne, R. I. Jedrzejewski, & R. N. Hook (San Francisco: ASP), 69

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