Astronomical Data Analysis Software and Systems VIII
ASP Conference Series, Vol. 172, 1999
D. M. Mehringer, R. L. Plante, and D. A. Roberts, eds.

ESO/CDS Data-mining Tool Development Project

Patrcio F. Ortiz, François Ochsenbein

Centre de Données Astronomiques de Strasbourg

Andreas Wicenec, Miguel Albrecht

ESO, Garching

Abstract. ESO and CDS are currently developing the data-mining tools which will allow users to access and combine the information stored in the forms of catalogs at CDS, and catalogs and data at ESO. CDS counts with about 5000 tables handled by the **VizieR** system (the data-mine, based on Sybase), including catalogs which appeared recently in journal articles. ESO has the data collected in the EIS program (deep survey with NTT) and will have in the future data from VLT's observations and other archived instruments in La Silla.

The main goal has been to build a "friendly" user interface so that users located remotely can either submit their own data "tables" (as ASCII files) for comparison or extract information from either ESO or CDS to perform cross correlations in all the parameter space provided by the data catalogs -not restricting the correlations to positional ones. We have decided to use HTTP as the exchange protocol between the Data Mining Facilities (DMF) located at CDS and the ESO servers, as well as between the user and the DMF.

1. Introduction

The objectives of this joint project between ESO and CDS are to build tools which will allow astronomers to access a large volume of information in the form of electronic data with the purpose of cross correlation. Remotely located users can either submit their own data "tables" (as ASCII files) for comparison or extract information from either ESO or CDS to cross correlate by position in the sky or by any of the parameters provided by the data catalogs. An important point has been the development of knowledge structures with the purpose of facilitating the description of the data to provide highly flexible data-mining options.

This paper outlines the main features users will find at the DMF and comments on features which have allowed us to optimize the consulting time. In addition we discuss the mechanisms used to organize the information in order to accept complex queries regarding the selection of the catalogs to use for comparison based on the nature of the objects contained, the "quantities" stored, wavelength coverage, sky region, etc.

2. The Data Mine at CDS

CDS hosts, under the **VizieR** system, around 2000 catalogs with more than 5000 tables, all of them accessible, described and stored using a very uniform system (Ochsenbein 1998; Ochsenbein et al. 1999). VizieR already allows queries by position and other criteria to individual catalogs.

The catalog content consists of: compilation catalogs, surveys, observing logs (from space missions), and tables from journal articles, with both observational and modeled data.

There are other facilities at CDS which could be used as a complement to data-mining: bibliographic references, name resolver (via Simbad), web resource locator (GLU), etc.

Knowledge-detection structures were developed to complement the meta information in the system. These structures were based on **column content** and also on **astronomical object type**.

3. The Knowledge-Detection Structures.

Two knowledge detection structures were developed: one for astronomical object types, and the other for column content. The structure for object type resembles the structure used in SIMBAD, with a four level hierarchy; the source to assign object types to the catalogs and tables is the standardized description file (ReadMe file) developed at CDS and shared now by other data centers and journals.

The structure related to column content was fully developed for this project. It contains 35 main categories and has a four level hierarchy. Categories such as Photometry, Positions, Spectroscopy, Time and Physical Quantities are amongst the most populated.

A **Unified Content Descriptor** (UCD) is then assigned to each of the columns in each of the tables accessed with **Vizier**. We developed an automatic UCD assignation procedure based on column name, column units, and column description.

Because of the importance of the use of UCD's for datamining purposes, we developed tools to assign UCD's to user provided files.

The existence of knowledge structures implies that it is possible, on one hand, to retrieve the names of all catalogs/tables containing any given set of UCD's, and on the other hand, to perform cross correlations with catalogs/tables containing the same UCD's as the ones describing the content of a user file.

As an example, the following table shows the assignment of UCD's to a few "typical" quantities (or columns) one could use for data-mining:

Column:	R.A.	DEC	Name	Vmag	z
Units:	h:m:s	d:m:s		mag	
UCD's:	POS_EQ_RA	POS_EQ_DEC	ID_MAIN	PHOT_JHN_V	REDSHIFT

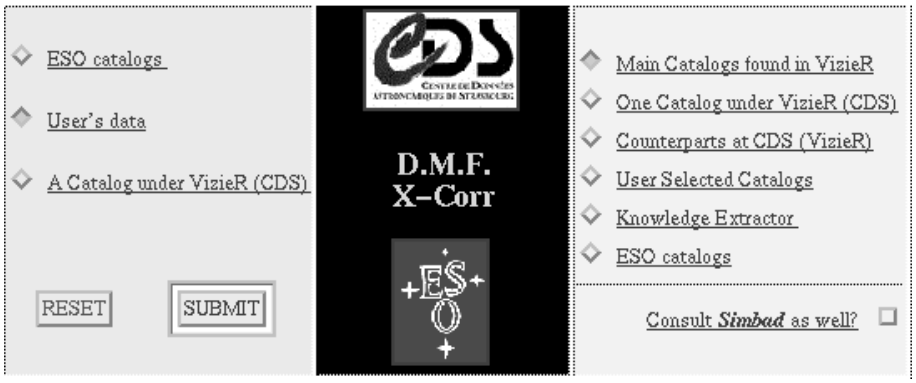

Figure 1. Cross Correlator options.

4. The Cross Correlator

Figure 1 shows the options provided by the cross-correlation interface. One way to envision datamining is a process in which users want to perform cross correlation operations using the data stored in a Data Mine as the comparison side, against their own data or data extracted from the same or other Data Mine. On the Users's side (reference side) there are several scenarios, such as:

- Consulting about single coordinates (user provided).
- Consulting about a list of coordinates and/or other quantities (user provided in the form of a file).
- Using data extracted from a Data-Mine (CDS, ESO, or somewhere else).

On the Data Mining Facility side (comparison side) CDS offers:

- Find counterparts in any catalog under **VizieR** . By far the most time consuming operation.
- Find counterparts in the Main Catalogs under **VizieR** . The main catalogs is a short list comprising some of the most relevant catalogs or surveys in many fields.
- Find counterparts in your favourite catalogs under **VizieR** . Users may submit a list containing the encoded names of what they consider the most relevant catalogs for their research purposes, and counterparts will be sought in those catalogs only.
- Compare data against ONE catalog under **VizieR** . The user knows well the content of a catalog and wishes to correlate against it. Correlation is open to all numerical quantities.
- Compare against catalogs with similar content under **VizieR** . Catalogs with similar "column content" (based on their UCD's) are open for cross correlation in all dimensions.

Some of the important features developed for this project include the following:

Web based engine which provides connection between the user, CDS, and ESO facilities.

Positional Cross Correlator designed to minimize the time to determine whether a catalog contains information in a certain region of the sky or not.

Quick search mechanism takes advantage of the querying system of **VizieR** to determine the matching elements.

High Flexibility to the user so that a number of alternatives can be selected with ease and speed.

Non positional Cross correlation by all numerical quantities stored in catalogs implies that users will be able to perform N-dimensional space analysis with as many numerical quantities as desired.

5. Practical Applications

During the development stage a number of practical applications were used as tests to measure flexibility and performance of the prototypes, some of the most relevant ones are listed here.

- Search for counterparts in one particular direction in the sky amongst all the catalogs under VizieR or a selected set of them.
- Search for counterparts for a user provided list of objects amongst all the catalogs under VizieR or a selected set of them.
- Self consistency test in any of the catalogs stored in the data-mine, e.g., search for double or multiple quasars in the latest list of AGN's by Veron-Cetty & Veron (1998).
- Positional cross correlation between one catalog in the datamine and one or more catalogs, e.g., galaxy clusters' catalog vs X-ray catalog.
- Search for neighbour objects based on non positional criteria, e.g., based on a user list find all objects with differences in redshift and absolute magnitude less than user defined values.

Other applications have been explored and will be part of the manual for the facilities or be part of the on-line examples available to users.

Acknowledgments. We are grateful to Pascal Dubois and Françoise Genova for valuable comments about the interface and the features one expects to find tools like the one presented in this paper.

References

Ochsenbein, F. 1998, in ASP Conf. Ser., Vol. 145, Astronomical Data Analysis Software and Systems VII, ed. R. Albrecht, R. N. Hook, & H. A. Bushouse (San Francisco: ASP), 387

Ochsenbein, F., Fernique, P., Ortiz, P., Egret, D., & Genova, F. 1999, in Future Generation Computer Systems (Amsterdam: North-Holland)

Veron-Cetty, M. P. & Veron, P. 1998, Quasars and Active Galactic Nuclei (8th ed.) (ESO Scientific Rep. No. 18)