Data services at the Centre de Données astronomiques de Strasbourg. Metadata and Added-values.

Marc WENGER[1]

[1]Centre de Données astronomiques de Strasbourg

11, rue de l'Université. 67000 STRASBOURG

wenger@astro.u-strasbg.fr

Abstract - The Centre de Données astronomiques de Strasbourg (CDS) has already gathered 30 years of experience in developing and maintaining several, continuously updated, reference databases in astronomy:
- SIMBAD had to evolve using different hardware, operating systems and software over 30 years of existence, having to manage 350,000 objects as early as 1972. Today, SIMBAD has 3,000,000 objects, and, beyond the basic data, astronomers and librarians at CDS have patiently gathered several added value information, such as object bibliography and cross-identifications between different catalogues.
- Vizier, a 3000-catalogue database, containing about 100,000 columns of every kind of astronomical data. They were used to create a set of (nearly) exhaustive metadata for designation of astronomical information.
- Aladin, an image and data integration tool, is becoming more powerful than its components by grouping images, catalogues and database information in the same frame. Aladin is also an excellent prototype of a "Virtual Observatory portal", accessing heterogeneous, distributed resources.

Résumé - Le Centre de Données astronomiques de Strasbourg (CDS) a déjà 30 ans d'expérience dans le développement, la maintenance et la mise à jour de plusieurs bases de données fondamentales pour l'astronomie:
- SIMBAD a du évoluer au travers de différents matériels, systèmes d'exploitation et logiciels au cours de ses 30 ans d'existence, ayant déjà 350000 objets à gérer en 1972. Aujourd'hui, SIMBAD contient 3.000.000 d'objets, et, au-delà de ses données de base, les astronomes et bibliographes du CDS ont patiemment rassemblés plusieurs valeurs ajoutées, tels qu'une bibliographie par objet astronomique et des identifications croisées d'objets entre plusieurs catalogues.
- VizieR est une base de données contenant plus de 3.000 catalogues astronomiques, regroupant en tout 100.000 colonnes contenant tous types de données astronomiques. Ces définitions de colonnes ont été utilisées pour fournir une liste quasi exhaustive de tous les types de données (métadonnées) utilisées en astronomie.
- Aladin est une base de données d'images, et aussi un outil d'intégration de données. Ses fonctionnalités sont décuplées par l'utilisation conjointe, dans une même interface, d'images et de données provenant de catalogues et de bases de données d'objets astronomiques. Ces possibilités d'accès à des données hétérogènes et distribuées font aussi d'Aladin un excellent prototype de portail d'accès à l'Observatoire Virtuel.

1. Introduction

Astronomy is an observation science, in which large amounts of data are collected since ages. The first known astronomical catalogue was written by Ptolemy (127-141) and contains position and magnitudes (luminosity) for 1047 stars. Between 1918 and 1924, Annie J. Cannon and Edward C. Pickering published a catalogue of 225,000 stars, with position, magnitudes and spectral types. Today, any astronomical ground- or space- mission produces several terabytes of data, and catalogues with up to a few billions of objects. Such data has to be stored and made available over the long term.

The CDS (Centre de Données astronomiques de Strasbourg) was created in 1972 to collect published data in astronomy, improve them through critical evaluation and distribute them to the astronomical community. To achieve these goals, several data systems were created : SIMBAD, a database compiling information for individual astronomical objects; VizieR, a database giving access to all published astronomical catalogues; and Aladin, an image database and data integrator.

Long term preservation, quality and adding value to these data has always been a main concern at CDS, as shown through a few examples below.

2. SIMBAD

In 30 years of existence, SIMBAD evolved through three successive versions: the first one was called CSI (Catalogue of Stellar Identification) and ran on an IBM mainframe. Queries were submitted through keypunched cards and the query result came back in printed version. The database contained up to 450,000 objects, and was only accessible through batch procedures for the French astronomical community. The second version, renamed SIMBAD (Set of Identifiers, Measurements and Bibliography for Astronomical Data), ran on Univac mainframe computers between 1981 and 1990. Through the packet switching networks implemented in most of the countries in the 80s, SIMBAD became available to virtually any astronomer in the world. Telnet accesses, and command line interface were the only ways to query the database, which contained at the end about 750,000 objects. Finally, since 1990, the third version has been running on Unix work stations, and it contains today (Sept. 2002) 3,000,000 objects. Access modes evolved from command line to X/Motif interface before the implementation of client/server protocols and of an access through the Web. Due to a modular design, several concepts that did not exist at design time could be integrated when needed, especially client/server communication.

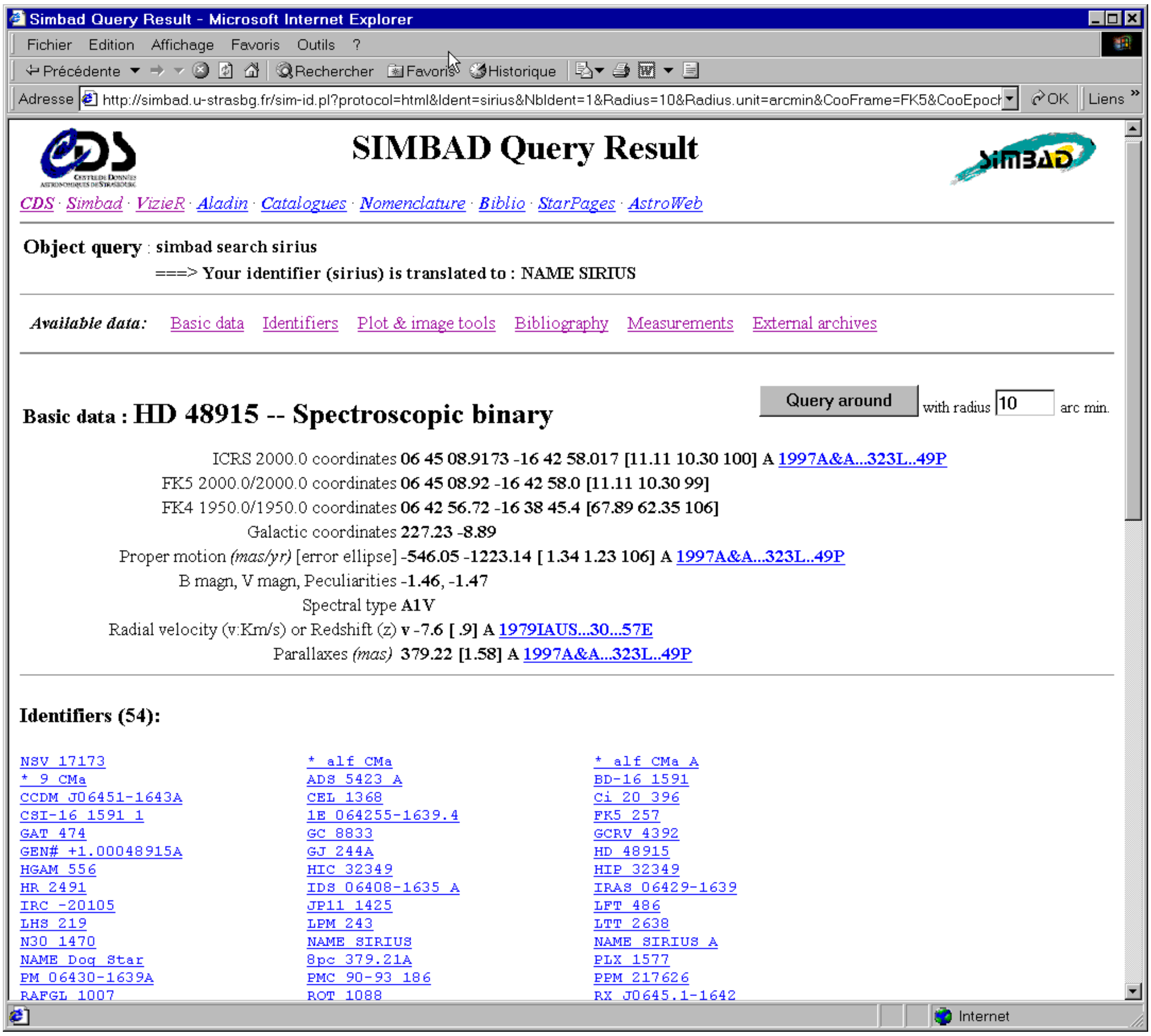

Fig 1. SIMBAD display on the web : basic data and identifiers.

2.1 ASCII data for long term preservation

Through these 30 years, and different host computers, the original data had to be moved and lots of new information were added. One key point for moving an existing database through different hardware and software is the ability to download it in ASCII format. It is the only way to avoid any machine dependence problem. Moreover, such ASCII downloads performed on a regular basis (monthly or yearly for instance) allow to keep track of the data history. Downloads can be kept on optical storage media for instance and will remain readable over the time. The chosen format should remain as simple as possible. XML is e.g. a current possibility. Of course, storage in ASCII format does not prevent to change the type of media when the technology evolution requires it.

2.2 Added value

A database is not only a collection of data. Using this raw material to add new information gives more power and value to the database. SIMBAD is build on data coming initially from catalogues. Beyond this, astronomers at CDS have made a huge effort to improve the added value of the database. The first one consist in finding cross-identification of objects between catalogues. This means assessing identity between objects in different catalogues, when no common name is given. This adds to the object the list of all its names which can be found in catalogues and publications. Some objects have currently more than 60 names. It is then possible to query an object from any of its identifiers. Exploiting these cross-identifications, it was then possible to introduce in SIMBAD another added-value: bibliography of astronomical objects. Articles in journals contain citation of objects by one of their names. Due to the cross-identification work, it was possible to retrieve all publications where an object is cited, whatever name was used in the original paper or catalogue. A unique bibliographical database was build by this way. Of course, maintaining such a bibliography requires manpower: librarians who read papers on a daily basis, and add references to an object for each name found in the articles; astronomers who mainly solve problems linked with cross identifications and object recognition when difficulties occur. SIMBAD bibliography for astronomical objects is used by ADS (Astrophysical Data System, the reference bibliography database in astronomy) to provide, beside the regular queries by keywords, queries by astronomical object names.

3. VizieR

Vizier is a database built from all astronomical catalogues and published tables stored at CDS. It is available in parallel with an anonymous ftp site allowing astronomers to copy the catalogues. Earlier, catalogues where kept on magnetic tapes and copies were sent to astronomers on request.

3.1 Table descriptions

Storing catalogues and tables would be useless without means to retrieve an appropriate catalogue for a given purpose. Furthermore, querying one particular catalogue requires knowledge of its content: column identification, physical units, etc... The key point to this is the existence of a complete description for every catalogue. This description should be both human and machine readable: human readable to allow one to read all the information pertaining to the catalogue; machine readable to be used by the software to incorporate a new catalogue in the database, to populate the metadata database with the characteristics of the new catalogue and to generate the forms required to query the catalogue. In some cases (tables from Astronomy & Astrophysics), these descriptions are provided by the authors, but an important work remains for the CDS staff: checking and often completing, or even rewriting them. The whole proper implementation and further usage of a catalogue depends on the quality of its description.

```
    II/20         UBV Photometry of O & B Stars in Vela     (Denoyelle 1977)
================================================================================
The Spatial Distribution of Young Stars in Vela
     Denoyelle J.
    <Astron.  Astrophys. Suppl. Ser., 27, 343 (1977)>
    =1977A&AS...27..343D
================================================================================
ADC_Keywords: Reddening ; Photometry, UBV ; Colors ; Stars, OB
Description:
    Photoelectric UBV values, derived from observations made at the Boyden
    and  ESO observatories, are presented for  358 early-type stars in the
    Vela section of the southern Milky Way.
File Summary:
       FileName       Lrecl  Records   Explanations
_ ReadMe               80          .    This file
_ data.dat            127       381    Data
Byte-by-byte Description of file:  data.dat
   Bytes Format  Units   Label    Explanations
       1  I1      ---     rec      ? Record number (if several records for star)
   2-  4  I3      ---     Seq      Running number
   6- 11  I6      ---     HD       ? HD number
  13- 20  A8      ---     CPD      ? Cape Photographic Durch. (or CD)
      21  A1      ---     n_CPD    [ *] * = CPDzone,num are from Cordoba Durch.
  22- 23  I2      h       RAh      Right Ascension (1950) hours
  25- 28  F4.1    min     RAm      Right Ascension (1950) minutes
      29  A1      ---     DE-      Declination (1950) sign
  30- 31  I2      deg     DEd      Declination (1950) degrees
 . . . .
 125-126  I2      ---     Nobs     Number of observations
     127  A1      ---     rem      [ *] * = Comment exists for this star
Note on Vmag, B-V, and U-B:
   99.99 indicates that the datum is missing.
Note on Q:
  Q = (U-B) - 0.72 (B-V)
Note on A(V):
  A(V) = 3 x E(B-V)
Note on Z:
  Z = R*sin(GLAT)
References:
  Denoyelle J. (1977) Astron. Astrophys. Suppl. Ser. 27, 343.
================================================================================
(End)
```

Fig 2. Exerpt of a catalogue description in VizieR

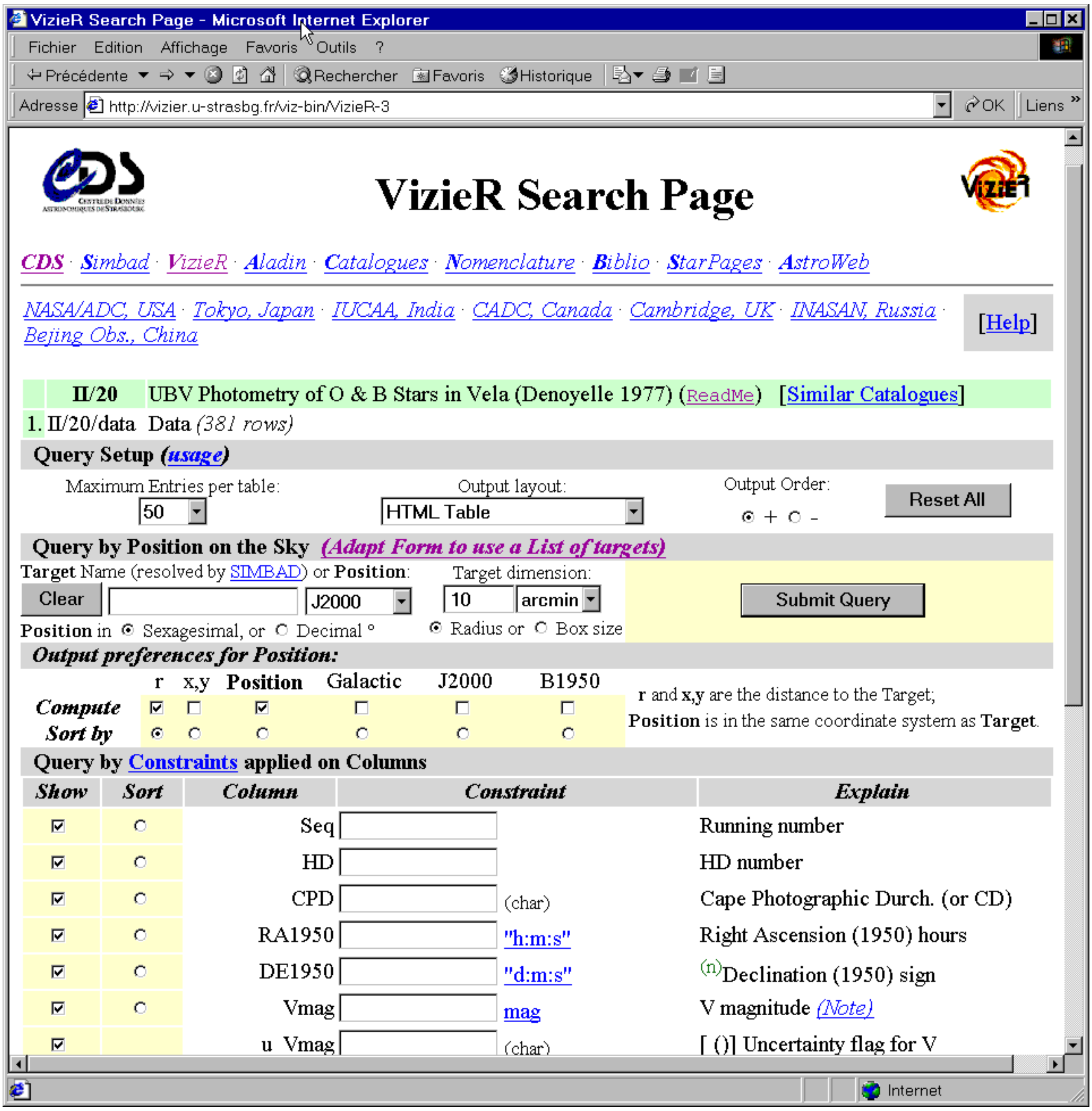

Fig 3. VizieR query page, built on the fly from the information provided by the catalogue description

3.2 Metadata dictionary

All 3000 VizieR catalogues are stored in relational tables. Put together, all these tables cumulate about 100,000 columns. They contain virtually any kind of data existing in astronomy. Unfortunately, many different column names may exist for one particular data type, because these names are given by the catalogue authors. For instance a Blue magnitude can be found in VizieR under 98 different names like **Bmag**, **B**, **<Bmag>**, **Bjmag**, **V30**, **V2**, **V24**, **<B>**, etc …

To allow retrieval of catalogues by column content, a concept linking all the column names to one precise definition designating the type of data was developed: the Uniform Content Descriptor (UCD).

The 100,000 columns were translated into about 1,500 UCDs (Uniform Content Descriptor). The UCDs consist in a hierarchy of definitions: the first level defines main data domains like position, photometry or spectroscopy and each further level designate an ever more precise subtype. For example, the B magnitude in the Johnston system corresponds to the UCD **PHOT_JHN_B**.

Tools are currently developed to help in finding all the catalogues having a particular data type, and also to help data providers to specify the right UCD to characterize a given data.

```
POS                               Position Related Quantities
  POS_ANG                         Angular Position
    POS_ANG_DIST                  Angular Distance and related quantities
      POS_ANG_DIST_GENERAL        Angular Distance Or Separation
      POS_ANG_DIST_REL            Relative or Normalized Angular Distance
      POS_ANG_DIST_SQ             Quadratic Angular Distance
    POS_ANG_VEL                   Rate Of Position Change (drift motion, angular velocity)
  POS_CCD                         Positions Related To CCD's
    POS_CCD_X                     CCD Position Along the X Axis
    POS_CCD_Y                     CCD Position Along the Y Axis
  POS_DETECTOR                    Position Related to the Detector
    POS_DETECTOR_DIST             Distance Or Position In Detector Units
    POS_DETECTOR_RADIUS           Radius of an Object Expressed In Detector's Unit
  POS_DIR-COSINE                  Direction Cosine
  POS_EARTH                       Positions On The Earth
    POS_EARTH_LAT                 Earth's Latitude
      POS_EARTH_LAT_MAIN          Earth's Latitude
      POS_EARTH_LAT_VAR           Variation of Local Latitude
    POS_EARTH_LOCATION            Earth Location
    POS_EARTH_LON                 Earth Longitude
  POS_EC                          Ecliptic Coordinates and derivates
    POS_EC_LAT                    Ecliptic Latitude
    POS_EC_LON                    Ecliptic Longitude
  POS_EQ                          Equatorial Coordinates and related quantities
    POS_EQ_A-V-PRECESS            Annual Variation of Precession in RA
    POS_EQ_DEC                    Declination related quantities
      POS_EQ_DEC_3T               Third Term in Declination
      POS_EQ_DEC_MAIN             Declination
      POS_EQ_DEC_OFF              Declination Offset Difference
      POS_EQ_DEC_OTHER            Declination in Non-Standard Units or partial values
      POS_EQ_DEC_PRECESS          Precession Variation in Declination
      POS_EQ_DEC_REL              Relative Declination in a Special Scale
    POS_EQ_PLX                    Relations between Parallax and RA and Dec
      POS_EQ_PLX_FACTOR           Parallax Factor in Declination
      POS_EQ_PMDEC                Proper Motion in Declination (pmdec)
      POS_EQ_PMRA                 Proper Motion in Right Ascension (pmra)
    POS_EQ_PREC                   Annual Precesion Quantities
      POS_EQ_PREC_DEC             Annual Precession In Declination
      POS_EQ_PREC_RA              Precession Variation In RA
    POS_EQ_RA                     Right Ascension related quantities
      POS_EQ_RA_2T                Second Component in right Ascension
      POS_EQ_RA_3T                Third Term In Right Ascension
      POS_EQ_RA_CORR              Correction in Right Ascension
      POS_EQ_RA_MAIN              Right Ascension
      POS_EQ_RA_OFF               RA Offset or Residual In Right Ascension
      POS_EQ_RA_OTHER             Right Ascension in Non-Standard Units or partial values
      POS_EQ_RA_REL               Relative Right Ascension in a Special Scale
  POS_GAL                         Galactic Coordinates and related quantities
```

Fig 4. Exerpt of the list of UCDs

Because UCDs build an exhaustive set of astronomy data types, they can be used to characterize any data type in metadata definitions and exchange. A first application is to introduce UCDs into the XML standard for data exchange between services. This XML standard, called VOTable, consists of field definitions followed by the data themselves. The UCDs are put as an attribute in the field definition.

```
<FIELD ID="MB" name="MB" ucd="PHOT_MAG_B">
<DESCRIPTION>
B magnitude
</DESCRIPTION>
</FIELD>
<FIELD ID="MV" name="MV" ucd="PHOT_MAG_V">
<DESCRIPTION>
V magnitude, visual magnitude
</DESCRIPTION>
```

Fig 5. Example of VOTable output of SIMBAD

Creation of these UCDs was an unexpected result of the VizieR database. It will not only improve the catalogue discovery tools, but leads to a more general way of characterizing and mining astronomical data.

4. Aladin

Basically, Aladin is an image database. Several collections of images produced by different surveys are stored in files, whereas a relational database contains the image description, as well as a pointer to the image file. It is then possible to display images based on criteria like the position on the sky, the wavelength range or the image or survey name.

The graphical interface, which is a JAVA client application, allows users to display images and to perform several manipulations on them.

Aladin can also access distributed image servers, several collections of data allowing to overlay astronomical objects on images (e.g. SIMBAD and VizieR), or to compute new coloured images by combining different images from the same field, in different wavelength for instance.

Due to this synergy, Aladin has become a very powerful, integrated tool offering many new facilities:

- By overlaying objects coming from catalogues or databases like SIMBAD or NED (Nasa/IPAC Extragalactic Database), Aladin facilitates the correction of errors in these databases. Misplaced objects overlayed on an image show immediately wrong coordinates.
- Overlaying objects coming from different catalogues in different wavelength bands, helps in cross identifying such objects, improving the range of wavelengths known for a particular object
- Adding two or three images, taken in different wavelengths, into the basic RGB colours gives, through a false colours image, new information related with physical processes at work in the objects, which would be hidden in one unique image.
- Two images taken at different epochs, inserted in two different colours of a composite image, can immediately reveal object movements, thus showing easily objects with proper motions.

Thus, far beyond a simple image database, Aladin has become a powerful tool for analysing data.

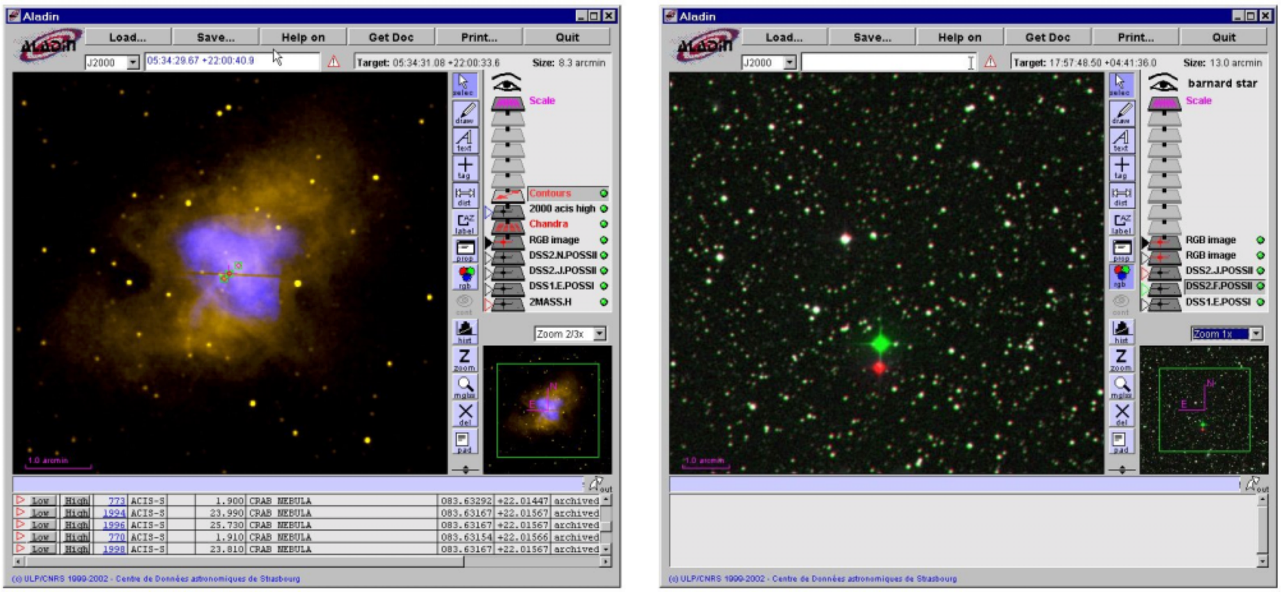

Fig 6. Examples of Aladin colour combinations :
left : crab nebula in visible and X-ray wavelength
right : Barnard Star (high proper motion) at two different epochs

5. Conclusion

Through its long experience in developing several databases in astronomy, CDS has become conscious that just providing raw data is only one part of the job. Creating added value, by combining data together and developing new information and services from this confrontation is the key to a more powerful usage of information. Of course, this requires manpower, not only computer engineers, but also scientists able to deeply understand the managed data, and librarians dealing with bibliography. This is the price to pay to fully exploit the available information.

A huge effort in currently undertaken world-wide in the astronomy domain to improve the usefulness of data, through interoperability between distributed, heterogeneous services. This leads to the concept of Virtual Observatory.

Acknowledgments: The UCD development was first initiated in the frame of the “ESO-CDS data Mining Project”.

References

[1] *The Centre de Données astronomiques de Strasbourg (CDS)*. http://cdsweb.u-strasbg.fr/

[2] M. Wenger, F. Ochsenbein, D. Egret, P. Dubois, F. Bonnarel, S. Borde, F. Genova, G. Jasniewicz, S. Laloë, S. Lesteven, R. Monier. *The SIMBAD astronomical database.The CDS reference database for astronomical objects*. Astronomy & Astrophysics suppl., 143,9 2000.

[3] F. Ochsenbein, P. Bauer, J. Marcout. *The VizieR database of astronomical catalogues*. Astronomy & Astrophysics suppl., 143,23 2000.

[4] F. Bonnarel, P. Fernique, O. Bienayme, D. Egret, F. Genova, M. Louys, F. Ochsenbein, M. Wenger, J. Bartlett. *The ALADIN interactive sky atlas. A reference tool for identification of astronomical sources*. Astronomy & Astrophysics suppl., 143,33 2000.

[5] Uniform Content Descriptor (UCD). http://vizier.u-strasbg.fr/UCD

[6] VOTable: A proposed XML format for Astronomical Tables. http://cdsweb.u-strasbg.fr/doc/VOTable/

[7] Nasa/IPAC Extragalactic Database (NED). http://nedwww.ipac.caltech.edu/

[8] Astrophysics Data System (ADS). http://adswww.harvard.edu/index.html