Information integration and retrieval: the CDS hub

Françoise Genova[a], François Bonnarel[a], Pascal Dubois[a], Daniel Egret[a], Pierre Fernique[a], Gerard Jasniewicz[b], Soizick Lesteven[a], François Ochsenbein[a] and Marc Wenger[a]

[a] CDS, Observatoire Astronomique, UMR 7550, 11 rue de l'Université, 67000 Strasbourg, France
[b] GRAAL, UMR 5024, Université de Montpellier II, CC 72, 34095 Montpellier Cedex 05, France

ABSTRACT

The *Centre de Données astronomiques de Strasbourg* (CDS) develops a set of value-added services, widely used for information retrieval, observation preparation, data interpretation, ... SIMBAD, VizieR, Aladin, and the 'Dictionary of Nomenclature', integrate heterogeneous, selected information from observatory archives, sky surveys and publications. Each service organizes information in a different way (astronomical objects, tables, images with overlays, nomenclature), and the CDS hub allows versatile information retrieval, e.g. looking for known information in a given region of the sky, including observations of ground- and space-based instruments, or searching by criteria in large data sets. Links among the CDS services, and with other reference on-line information systems, such as observatory and survey archives and publications, permit comprehensive searches in a wide variety of resources. Shared exchange standards and generic tools such as the GLU are essential for the building of links. XML is a tool for further information integration, and Aladin is a precursor of an integration tool, relying on FITS and XML. New functionalities will be developed at CDS in the context of the Virtual Observatory, e.g. for data mining and management of very large catalogues. A prototype set of interoperable archives will be implemented in the frame of the European *Astronomical Virtual Observatory* project.

Keywords: Information retrieval, information networking, astronomy virtual observatory, standards

1. INTRODUCTION

The progress of information networking in astronomy has been tremendous in the last years, taking the best advantage of the possibilities offered by the WWW. Information is on-line, from all origins – from Agencies to individuals. The WWW allows integration of data with documentation, thus facilitating the reuse of data. It also allows the building of links between distributed on-line information, which has been the basis of the rapid construction of the astronomical information network.

The quality of distributed information and the suitability of the services for the diversity of scientific needs are key assets. Major actors develop different kinds of services, in their respective fields of expertise:

- observatories implement archive services, which contains data, and information about data quality, instruments, observation procedures, ...;
- specialized data centers, in particular the disciplinary NASA data centers and the ADS, distribute data and tools adapted to their domain;
- journals have very rapidly implemented electronic versions, in close connection in particular with data centers;
- data centers such as the CDS develop value-added services to help astronomers to retrieve and use information from the wide variety of available resources.

Networking of information has been made possible by partnership between the actors, in particular for the definition of a set of *de facto* exchange standards[1], many of them developed well before the advent of the WWW, which are essential for the building of links.

2. THE CDS ACTIVITIES

The *Centre de Données astronomiques de Strasbourg* was founded in 1972 as the *Centre de Données Stellaires*, to tackle information in electronic form and develop expertise on these data, with the objective to provide the international astronomical community with tools to enable scientific research. At that time the main topic was studying galactic stellar populations and dynamics. The CDS became *Centre de Données astronomiques de Strasbourg* in 1983, and since then it has been dealing with all kinds of astronomical objects and information. CDS is operated under joint responsibility of the *Institut National des Sciences de l'Univers* (INSU) and *Université Louis Pasteur*, Strasbourg.

The CDS activity has several facets: development and maintenance of services, in all their aspects – contents, data management, user interface, operations; definition of standards and generic tools; participation to projects. To secure the quality of the service contents, selection, validation and homogenization of information is the long-term, daily task of a highly specialized team. On the other hand, to follow the very rapid pace of technical evolution in these domains, technological watch and R&D activities are carried on at a sizable level, and taken into account in decision making about new developments.

A detailed description of CDS was published in Ref. 2. All services developed and hosted by CDS are accessible from the CDS home page at http://cdsweb.u-strasbg.fr/ (Fig. 1).

Figure 1. CDS Home Page (June 2001).

3. THE CDS SERVICES

The CDS services integrate information selected from heterogeneous sources: publications, catalogues, archives, and large surveys. Each service organizes information differently, permits flexible queries, and implements links with the other CDS services and with other distributed information holdings (archives, data centers, journals, databases).

3.1. SIMBAD

SIMBAD[3] contains information about astronomical objects. In particular, it keeps track of all the names of one object in different catalogues and published tables ('identifiers'), and of all the papers in which a given object has been cited. It also contains basic data such as position, object type, proper motion for stars and redshift for extragalactic objects,... , and some measurements. In July 2001, SIMBAD contains 2,800,000 objects, 7,750,000 identifiers, 122,000 bibliographical references,

and 3,560,000 citations of objects in papers. SIMBAD permits different query modes: by object name, around a position (given by coordinates or object names), by reference, by lists of positions, object names or reference, by constraining the basic data (magnitude, color index, spectral type, ...), by checking the presence of an identifier from one catalogue or of citations in a given range of years,... An XML output is being developed. The top of the default result page is shown in Fig. 2 (for NGC 4151).

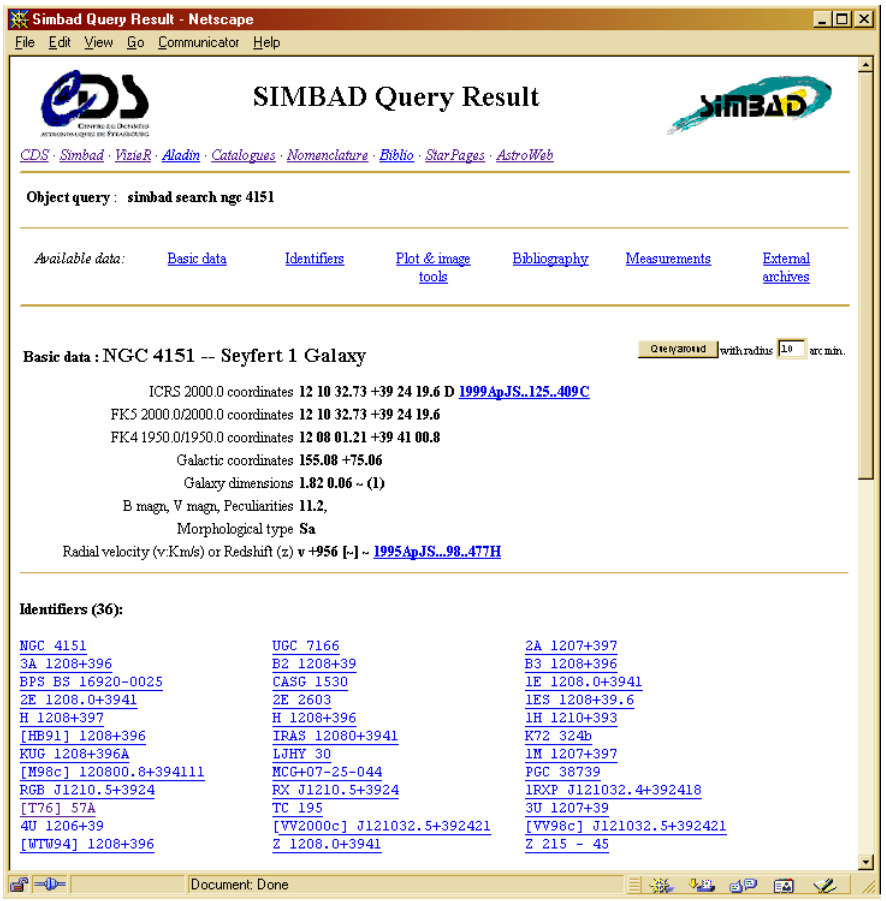

Figure 2. Example of a SIMBAD result page for NGC 4151 (fist part of the result page). The default result page displays basic data, identifiers, links to plots and image tools, bibliography, measurements, links to 'external' archives', including the catalogues in VizieR.

SIMBAD gives access to complementary information from the other CDS services: full catalogue data from VizieR (since March 2001), images from Aladin, information about identifiers from the Dictionary of Nomenclature. It also links from objects to the HEASARC[*] and IUE[†] data, and from bibliographic references to the full set of ADS[‡] information and to the original paper in the electronic journal when available.

3.2. VizieR

The VizieR catalogue browser[4] federates information in tabular format. In July 2001, it contains more than 3,000 catalogues, from the collection of 4,021 catalogues available at CDS (which also includes older versions of catalogues, and more than 300 catalogues not well suited for browsing but which can be copied by ftp): most of them are reference catalogues and tables published in journals. These latter are implemented in close collaboration with the journals[5], and their number is increasing by several hundreds per year. VizieR can be queried around a position (given by coordinates or object name), by a list of

[*] http://heasarc.gsfc.nasa.gov/
[†] e.g. http://godot.u-strasbg.fr/ines/ (copy of the INES database at CDS)
[‡] http://adswww.harvard.edu/

positions, or by constraining the data contained in one or several columns of one or several catalogues. One can also search the whole set of catalogues, or a subset, around a position (Fig. 3). The output format can be chosen from many possibilities (html, ascii, tab separated values – which can be directly used in spreadsheets, XML, ...).

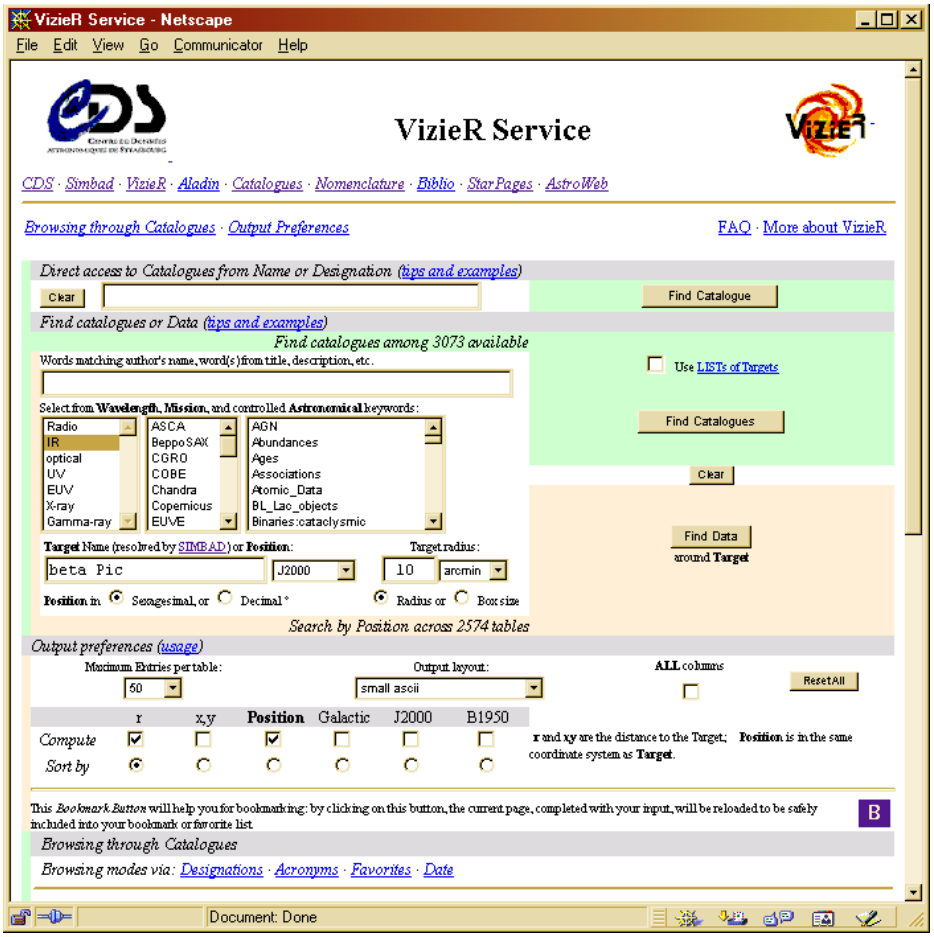

Figure 3. Example of a VizieR query (all objects found in infrared catalogues, within 10 arcmin of beta Pictoris).

VizieR is able to give access to any kind of tabular data, provided that a *standard description*[§] is available. Lists of observations contained in archives are progressively implemented in the system (HST[**], IUE, CFHT[††], ISO[‡‡] are presently on-line), with links to the archival data kept at the observatory site. Similarly, the same interface gives access to survey data (2MASS[§§], DENIS[***], USNO A2.0[†††], FIRST/VLA[‡‡‡], ...). More generally, the list of tables which have links to additional data, images, spectra and time series, is progressively growing[§§§].

[§] http://cdsweb.u-strasbg.fr/doc/castd.htx
[**] http://www.stsci.edu/hst/
[††] http://cadcwww.dao.nrc.ca/cfht/cfht.html
[‡‡] http://www.iso.vilspa.esa.es/
[§§] http://pegasus.phast.umass.edu/2mass.html
[***] http://cdsweb.u-strasbg.fr/denis.html
[†††] http://ftp.nofs.navy.mil/projects/pmm/
[‡‡‡] http://sundog.stsci.edu/
[§§§] http://vizier.u-strasbg.fr/cgi-bin/vizHelp?cats/M.htx

3.3. Aladin

The Aladin interactive sky atlas[6] integrates image information with tabular and database data. Reference images of the sky from DSS-I and -II, scanned by the STScI[****], are completed with full resolution images covering 'crowded' regions, such as the galactic plane and the Magellanic Clouds, scanned by the MAMA digitizing machine at the *Centre d'Analyse des Images* in Paris. The Aladin Java tool allows users to select images, and to overlay SIMBAD, NED[††††], and all VizieR tables - catalogues, published tables, observation logs and surveys (Fig. 4). Images from archives can also be retrieved and displayed, with the same overlay capabilities.

New image sets are being implemented in Strasbourg, in particular the 2MASS one. Moreover, access to distributed image data sets such as the SuperCOSMOS Sky Surveys[‡‡‡‡] will be available in the next version. With the standalone version, users install the software locally, and user catalogues and images (providing that they have a FITS WCS description) can be used with the full interface functionalities.

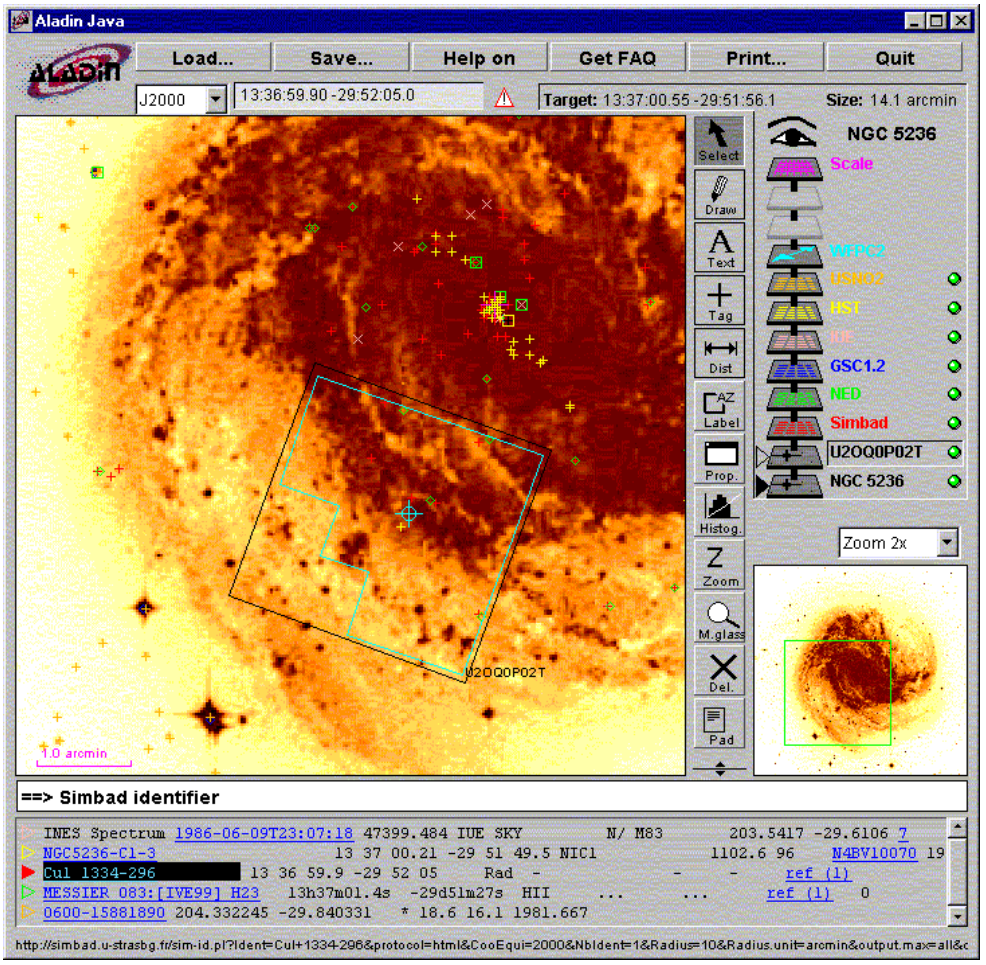

Figure 4. NGG 5236 information displayed in Aladin (DSS image, SIMBAD, NED, GSC and USNO A2.0 surveys, HST observation, with one HST field of view, IUE observations)

[****] http://stdatu.stsci.edu/dss/
[††††] http://nedwww.ipac.caltech.edu/
[‡‡‡‡] http://www-wfau.roe.ac.uk/sss/

3.4. The Dictionary of Nomenclature

The *Dictionary of Nomenclature of celestial objects (outside the solar system)*[7,3] keeps track of the diversity of astronomical Nomenclature (Fig. 5). New catalogues and lists are recorded, and the different names found in the literature for a given catalogue are recorded. In June 2001, the *Dictionary of Nomenclature* contains more than 9,300 acronyms ('main' acronyms and synonyms), among which more than 4,800 'main' acronyms are included in SIMBAD.

The *Dictionary of Nomenclature* can be queried by acronym, by first author name, by any word contained in the entries,... Access is given to the relevant bibliographical data and to the tables in VizieR when available. Additional information about the data origin (e.g. the mission at the origin of the observations for space data), new versions and additions to original data, is also kept.

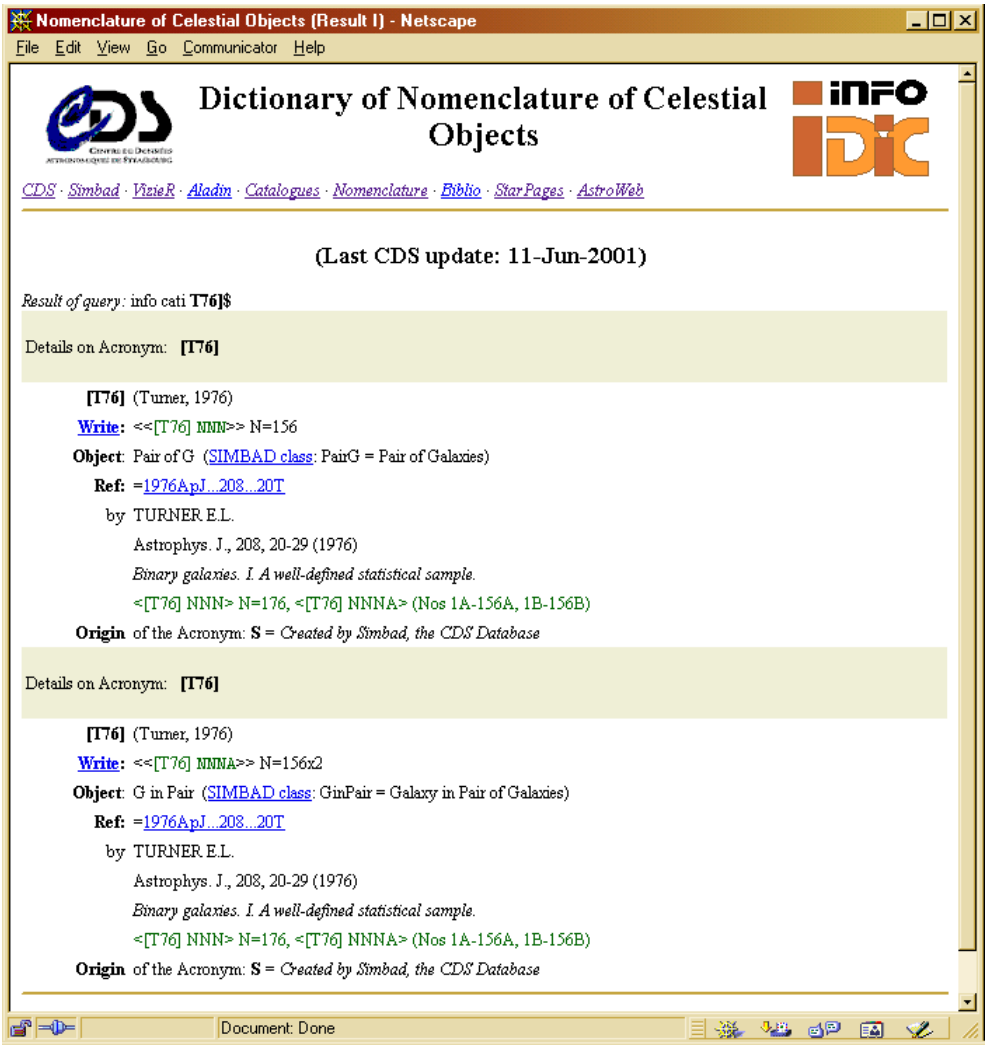

Figure 5. The Dictionary of Nomenclature result page for acronym [T96]

(obtained from the [T96] 57A identifier of NGC 4151, Fig. 1)

3.5. Other Services Available from CDS

CDS hosts mirror copies of reference bibliographic services: the ADS database, the *Astrophysical Journal*, the *Astronomical Journal*, the *Publications of the Astronomical Society of the Pacific*, and since the beginning of 2001 *Astronomy and*

Astrophysics. It also hosts yellow page services (*AstroWeb*[8], the *Star*'s Family*[9]), the Queue Service Observing of CFHT and a copy of their user documentation,... A set of additional bibliographic services (links from object names in abstracts to SIMBAD, graphical visualization of published paper contents through neural network techniques[10]) has also been developed.

A mirror copy of SIMBAD is maintained at CfA by the ADS team, and several mirror copies of VizieR are installed (NASA's ADC-USA; NOAJ/ADAC-Japan; IUCAA-India; CADC/CCDA-Canada; UK Astronomy Data Centre-UK). A copy of the *Dictionary of Nomenclature* is also installed in Japan.

4. INFORMATION RETRIEVAL FROM THE CDS SERVICES

The CDS information hub allows versatile searches in a wide variety of resources, from observations (archive logs and catalogues in VizieR, links to archives in SIMBAD, archive images in Aladin,...) to results published in journals (bibliographic information in SIMBAD, published tables in VizieR, nomenclature in the *Dictionary of Nomenclature*, ...), and also to surveys, reference catalogues and images.

Searching for the available information around an object or in a region of the sky is one of the main query types. It aims for instance at preparing an observation, or at validating and interpreting observations by going back to previously observed data, by multi-wavelength cross-identification of objects,... Querying SIMBAD, VizieR, Aladin (which allows comparison with user data in the standalone version), gives access to a large set of information. The links to distributed resources then go to original observations or publications. In addition, a wide set of on-line services can be queried through *AstroGLU* or from the *AstroBrowse* services[1,11], where a single query retrieves the corresponding page(s) from all the relevant services selected among the available ones. These information retrieval tools rely on the *Générateur de Liens Uniformes* (GLU)[12], developed by CDS, which allows service managers to maintain a distributed dictionary describing resources and links.

In addition, the list of observed objects can be compared to data from SIMBAD and VizieR, through the 'query by list' functionalities. A new service developed in collaboration with ESO in the frame of the *ESO-CDS data mining project*[13], presently in a prototype form, will facilitate these comparisons. In this context, a new set of metadata has been developed, to improve the description of catalogue contents and help the comparison of information from different sources: for instance, it will be possible to identify all the catalogues containing V magnitudes and proper motions, which is not feasible from column names only (V magnitude can be column 'V', 'mag', 'm',...; column 'V' can be V magnitude, velocity, ...).

SIMBAD and VizieR can also be searched for objects compliant with one or several criteria: in SIMBAD, all basic data items can be constrained; in VizieR, any catalogue column (nearly 100,000) can be used for sampling. The new data mining functionalities will also bring new functionalities for these comparisons. Another domain for R&D is cross-correlation of very large catalogues: at present, very large catalogues (more than a few million objects) are organized to allow for rapid positional searches but are not well suited for multi-criteria searches.

More advanced searches can also be performed throughout the CDS information holdings. For instance, all object lists published from observations from a given space mission can be retrieved from the *Dictionary of Nomenclature*. Currently, e.g., in June 2001, more than 300 lists from ROSAT and 25 from Chandra (many of the latter registered before publication through the *IAU Acronym registry*[§§§§]) have been recorded. The *Dictionary of Nomenclature* gives access to the published papers at electronic journal sites, and to the tables if they are available in VizieR. SIMBAD can be queried by any of these acronyms to retrieve all the information about the objects. Alternatively, VizieR can also be searched by any mission name, to retrieve reference catalogues and tables derived from the mission observations. Another example could be to retrieve all gamma ray bursts observed e.g. in 1999 (from SIMBAD), to gather the list of archival wide-field observations of the field (from VizieR – only a few of these observations are available on-line), and to access reference images of the field from Aladin (DSS, HST).

5. FROM NETWORKING TO INTEGRATION OF INFORMATION

A few years after the WWW revolution, information in astronomy has thus reached a high level of networking. This is particularly striking for bibliographic information. Thanks to the active collaboration between journal editors and publishers, the ADS database and the data centers, astronomers can navigate to and from the electronic journals, ADS, NED and SIMBAD information. The *bibcode*, a 19-character description of references, first defined by NED and SIMBAD for their bibliographic information exchange, then extended and widely used by ADS, has been a key tool for the bibliographic information networking. Moreover, archives are joining the bibliographic network by beginning to identify publications originating from individual datasets, with links to the ADS for each reference; these links will be referenced in ADS, thus

[§§§§] http://vizier.u-strasbg.fr/viz-bin/DicForm/

enabling links from bibliographic services to archives. Another value-added service, links between object names in published papers and databases, is being implemented in a prototype form: *New Astronomy* has built links to NED and SIMBAD from its first issue, and a few months ago *Astronomy and Astrophysics* began implementing links between object names tagged by the authors and SIMBAD.

The next step is, beyond links between distributed services, to be able to *integrate* results of queries to heterogeneous resources. One ultimate goal would be e.g. to integrate all the information retrieved from an AstroGLU/AstroBrowse query to observatory archives into a single result page, displaying properly all the information gathered from all the different sources.

A first step towards distributed information integration, is to allow one to use data from a distant resource for one's own service needs. A good example is the usage of SIMBAD and NED as 'name resolvers' (i.e., to translate object names into coordinates or reference lists) in all the major on-line archive services and in the ADS, through a client/server mechanism. To avoid software distribution and maintenance, the usage of agreed upon data standards and query mechanisms is being developed. The *astrores* XML description of tables[15], defined by an international collaboration and used (among many other formats) as output by VizieR, is for instance a powerful tool, allowing services to interface easily with VizieR. Browsing of VizieR catalogues is for instance included in the last version of the HEASARC interface and in OASIS[14].

Aladin, used by NED for visualization of their image database, and by the CFHT as image tool for observation preparation in the Queue Service Observing mode, is a powerful integration tool of heterogeneous and distributed information. Images stored in Strasbourg and distributed image archives, catalogues, published tables, surveys, lists of observations in archives, databases such as SIMBAD and NED, can be used together to gather multi-wavelength, multi-technique information on any field of the sky.

New archives can easily be included in VizieR and in Aladin. For VizieR, one only needs to provide a list of observations, which can be regularly updated if needed, e.g. for 'living' observatory logs, and its 'standard description'; for Aladin (for WCS FITS images), access can be built either from the observation list in VizieR or (in the next release) by a direct link to the image archive, provided that the query syntax description is known from the system.

6. CONCLUSION: THE CDS ROLE IN THE VIRTUAL OBSERVATORY

For all astronomy data providers, the evolution in the coming years will be in the context of the Virtual Observatory projects, towards more links and data integration, and towards the development of new services dealing with heterogeneous, distributed information, in particular for statistical study of large surveys, for information extraction from large sets of images, and for data mining.

CDS has largely anticipated the Virtual Observatory concept: SIMBAD integrates heterogeneous information about objects; VizieR federates tabular data – catalogues, published tables, lists of observations and surveys; Aladin is an operational integration tool for images, databases, tables, able to access information from CDS and other data centers, images stored in Strasbourg and in distributed image archives, and user data. These services build a bridge between observations and published results, and include large surveys for positional queries. VizieR and Aladin already implement important functionalities of the Virtual Observatory.

The CDS services will remain essential reference resources for data interpretation, each service being a hub towards the other ones, and towards other distributed resources. Data federation and integration through VizieR and Aladin will be further developed, in collaboration with archive providers and disciplinary centers. In particular, one goal of the *Astronomical Virtual Observatory* project, recently approved in the frame of the European Commission RTD program, is to build an *Interoperability prototype*, based on VizieR and Aladin, for a set of European archives, representative of ground- and space based observations at different wavelegths, in collaboration with all the project partners (ESO, ESO, UK AstroGrid, Terapix, Jodrell Bank). In addition, the implementation of links to the US archives and to the CFHT will continue to be developed.

New, innovative services such as the data mining/cross-identification tool developed in the *ESO-CDS Data Mining project* will be provided. Multi-criteria access to very large catalogues will also be implemented. In addition, the CDS will continue to participate actively in the definition of metadata and exchange standards (in particular by studying XML usage in different contexts, in addition to tables), and to provide generic tools for interoperability such as the GLU.

REFERENCES

1. D. Egret, and F. Genova, "Interoperability tools for the Virtual Observatory", *Proc.* SPIE 4477, these proceedings.

2. F. Genova, D. Egret, O. Bienaymé, F. Bonnarel, P. Dubois, P. Fernique, G. Jasniewicz, S. Lesteven, R. Monier, F. Ochsenbein, and M. Wenger, "The CDS information hub. On-line services and links at the Centre de Données astronomiques de Strasbourg", *Astron. Astrophys. Suppl.* **143**, pp. 1-7, 2000.
3. M. Wenger, F. Ochsenbein, D. Egret, P. Dubois, F. Bonnarel, S. Borde, F. Genova, G. Jasniewicz, S. Laloë, S. Lesteven, and R. Monier, "The SIMBAD astronomical database. The CDS reference database for astronomical objects", *Astron. Astrophys. Suppl.* **143**, pp. 9-22, 2000.
4. F. Ochsenbein, P. Bauer, and J. Marcout, "The VizieR database of astronomical catalogues", *Astron. Astrophys. Suppl.* **143**, pp. 23-32, 2000.
5. F. Ochsenbein, and J. Lequeux, "The A&A tables and abstracts: an example of collaboration between data centres and editors", *Vistas in Astron.* **39**, pp. 227-233, 1995.
6. F. Bonnarel, P. Fernique, O. Bienaymé, D. Egret, F. Genova, M. Louys, F. Ochsenbein, M. Wenger, and J. Bartlett, "The ALADIN sky atlas. A reference tool for identification of astronomical sources", *Astron. Astrophys. Suppl.* **143**, pp. 33-40, 2000.
7. M.-C. Lortet, S. Borde, and F. Ochsenbein, "Second Reference Dictionary of Nomenclature of Celestial Objects", *Astron. Astrohys. Suppl.* **107**, pp. 193-218, 1994.
8. R. Jackson, D. Wells, H.M. Adorf, D. Egret, A. Keokemoer, and F. Murtagh, "AstroWeb. A database of links to astronomy resources (Announcement of a database) ", *Astron. Astrophys. Suppl.* **108**, pp. 235-236, 1994.
9. A. Heck, "Electronic yellow-page services: The Star*s Family as an example of diversified publishing", in *Electronic publishing for Physics and Astronomy*, A. Heck, ed., pp. 211-220, Kluwer Acad. Publ., Dordrecht, 1997.
10. P. Poinçot, S. Lesteven, and F. Murtagh, "A spatial user interface to the astronomical literature", *Astron. Astrophys. Suppl.* **130**, pp. 183-19, 1998.
11. D. Egret, R.J. Hanisch, and F. Murtagh, "Search and discovery tools for astronomical on-line resources and services", *Astron. Astrophys. Suppl.* **143**, pp. 137-143, 2000.
12. P. Fernique, F. Ochsenbein, and M. Wenger, "CDS GLU, a tool for managing heterogeneous distributed web services", in *Astronomical Data Analysis Software and Systems VII*, R. Albrecht, R.N. Hook, and H.A. Bushouse, eds., *A.S.P. Conf. Ser.* **145**, p. 466, 1998.
13. P. Ortiz, F. Ochsenbein, A. Wicenec, and M. Albrecht, "ESO-CDS Data-mining tool development project", in *Astronomical Data Analysis Software and Systems VIII*, D.M. Mehringer, R.L. Plante, and D.A. Roberts, eds., *A.S.P. Conf. Ser.* **172**, p. 379, 1999.
14. J.M. Good, et al, *Proc.* SPIE 4477, these proceedings.
15. F. Ochsenbein, M. Albrecht, A. Brighton, P. Fernique, D. Guillaume, R.J. Hanisch, E. Shaya, and A. Wicenec, "Using XML for accessing resources in astronomy", in *Astronomical Data Analysis Software and Systems IX*, N. Manset, C. Veillet, and D. Crabtree, eds., *A.S.P. Conf. Ser.* **216**, p. 83, 2000.