Running a Data Centre on the Long Term: Lessons Learnt from 30 Years of CDS History

Françoise Genova, François Ochsenbein, Marc Wenger, Mark Allen, Olivier Bienaymé, Thomas Boch, François Bonnarel, Laurent Cambrésy, Sébastien Derriere, Pascal Dubois, Pierre Fernique, Soizick Lesteven, Cécile Loup, André Schaaff, Bernd Vollmer[1], Gérard Jasniewicz[2], Emmanuel Davoust[3], and Daniel Egret[4]

[1] CDS, Observatoire Astronomique de Strasbourg, UMR CNRS/ULP 7550, 11 rue de l'Université, 67000 Strasbourg, France, genova@astro.u-strasbg.fr, WWW home page: http://cdsweb.u-strasbg.fr
[2] GRAAL, CC 72, Université de Montpellier II, 34095 Montpellier Cedex 05, France
[3] LAT, Observatoire Midi-Pyrénées, 14 avenue Edouard Belin, 31400 Toulouse, France
[4] Observatoire de Paris, 61 avenue de l'Observatoire, 75014 Paris, France

Abstract. The Centre de Données astronomiques de Strasbourg (Strasbourg astronomical Data Centre – CDS) provides value-added, leading-edge services which are widely used by astronomers world-wide in their daily research work. The CDS was created in 1972, and this paper details lessons learnt from more than 30 years of successful activities serving the scientific community.

1 Introduction

The Centre de Données astronomiques de Strasbourg (CDS) was created in 1972 by the *Institut National d'Astronomie et de Géophysique* – INAG, which is now *Institut National des Sciences de L'Univers* – INSU. CDS was initially called *Centre de Données Stellaires* (Stellar Data Centre), and its initial charter can be summarized as follows:

- collect 'useful' data on astronomical objects, in electronic form;
- improve them by critical evaluation and combination;
- distribute the results to the international community;
- conduct research using these data – at the time of CDS creation, the original objective was to gather stellar data to study the galactic structure.

Fundamental keywords which still structure CDS activity and development strategy have thus been present from the very beginning: the Data Centre main objective is to provide science tools to the community – although it also has many aspects of a data curation centre; it deals with *electronic data* – which, taken as a goal as early as 1972, shows the long term vision of Jean Delhaye,

then Director of INAG, who founded CDS; it intends to build a local expertise on data; it aims to serve the international community. Its activities are regularly examined by a Scientific Council, half the member of which are foreign scientists.

In 1983, the CDS target was extended to all astronomical objects (outside the solar system), and its name was changed to the present one. Since the advent of the World Wide Web, CDS has been a major player in the development of on-line information services and in the networking of these services: in particular, thanks to the collaboration between CDS, ADS, NED, the journals, and major observatory archives, all major bibliographical services are now networked. In recent years, CDS has also been a major player in setting up the International Virtual Observatory.

The organisation and activities of CDS are described in Section 2. The way it deals with technical evolution, the importance of partnership, and the CDS relations with the scientific community, are discussed respectively in Section 3, 4 and 5.

2 CDS Organisation and Activities

The CDS goals can be summarized as *collect, homogenize, distribute, and preserve astronomical information, for the usage of the whole astronomical community.*

In the WWW era, CDS deals with information, which includes data, but also published results, metadata, links, resource registries, know-how about data, etc (see Genova et al. [1]). Its activities have different aspects:

- the development, maintenance and on-line diffusion of reference, value-added services, in particular SIMBAD, the reference database for identification and bibliography of astronomical objects (see Wenger et al. [2]), VizieR, the reference database for information in tabular formats (catalogues, tables published in journals, observation logs, surveys – see Ochsenbein, Bauer and Marcout [3]), and the Aladin sky Atlas (see Bonnarel at al. [4]), with also additional services such as the *Dictionary of Nomenclature of astronomical objects outside the solar system* (see Lortet, Borde and Ochsenbein [5]) which is an important by-product of SIMBAD. This covers different types of activities, in particular the building of the content, software development, and operations. Data is selected from publications, catalogues, reference images. Numerous links are built with other resources, in particular observatory archives, electronic journals, and other on-line services, for which metadata are retrieved and maintained.
- the development and maintenance of standards and generic tools. Several examples of standards will be given in Section 4, as an illustration of the importance of partnership with other actors. In terms of generic tools, Aladin has evolved to become in particular a widely-used Virtual Observatory portal; client/server packages and XML/SOAP Web Services are provided to allow the usage of the CDS services from other services. Building on this

long term expertise, the CDS team has been very active from the very beginning in the development of standards for the Virtual Observatory (VO): for instance, the first VO standard, VOTable [5], has been proposed by F. Ochsenbein (CDS) and R. Williams (NVO). Moreover, the origin of the International Virtual Observatory Interoperability meetings is in the OPTICON *Interoperability Working Group* which was proposed and run by CDS before the creation of the *International Virtual Observatory Alliance*: its first meeting was held in Strasbourg in January 2002 and settled the basis for VOTable.

- technological/methodological watch and R&D have been critical: new possibilities offered by technical evolution have to be taken into account as soon as possible, but not too early because technologies implemented in operational services must be reliable and also have to be maintained for a certain time. 'Flash in the pan' fashionable technologies have in particular to be avoided and thus careful evaluation is required before implementing any new technology.
- participation in projects, which is built on local expertise: for instance, Hipparcos used the experience gathered from the building of SIMBAD and of the catalogue collection; the XMM Survey Science Center uses VizieR and Aladin images. More recently, thanks to its expertise in standards, information networking, and the provision of tools, CDS has been one of the major partners of the European Virtual Observatory RTD project *Astrophysical Virtual Observatory*, in which it was in charge of the *Interoperability* Work Package, and which VO portal was derived from Aladin, and of the VO-TECH Design study, in which it is in charge of the *Intelligent discovery* aspects, and participates actively in all other tasks. In turn, functionalities prototyped in the frame of projects are included in the services.
- user support is also a very important task, with two types of consumers: astronomers who are willing to use the services in their daily research, and service providers who are willing to use the CDS services in their own services. On-line documentation and a hot line (**question@simbad.u-strasbg.fr**) are provided, plus demonstrations of the services at scientific and technical meetings.

In the last 33 years, the context of CDS activity has been constantly evolving, and it has in particular to take properly into account the evolution of astronomy and the very rapid technical evolution. CDS tasks include very different activities with quite different time scales, from the building of the database content which is a daily activity on the long term (several decades), to prototyping of new techniques which is a short term activity on a few months time scale. Managing this diversity requires defining properly long term goals, and assessing and adjusting strategy as required to take into account the numerous constraints.

The CDS team integrates staff with different profiles and tasks: astronomers, computer engineers and specialized librarians. The strategy and work program

[5] http://www.ivoa.net/Documents/latest/VOT.html

definitions are the task of a group which includes astronomers and engineers, to make sure that scientific objectives and technical constraints are taken into account. The CDS director chairs this group and has the final cut in the very rare cases in which consensus cannot be achieved.

The data curation task is particularly demanding, and requires a highly skilled staff of specialized librarians ('bibliographers'), working in close collaboration with astronomers who provide the expertise on goals and to solve the problematic cases. Proper assessment of the evolution of contents and of data curation methods is an important part of the CDS strategy. For instance, specific actions to improve and increase X-ray data in SIMBAD and VizieR had been undertaken before Chandra and XMM launches. In addition, staff expertise has to be updated continuously with the help of the astronomers to include the scientific evolution of astronomy. The training period of new 'bibliographers' is long because of the difficulty of the task, and continuity in expertise is essential – departures or retirements have to be replaced as soon as possible, and enough time allocated to proper training by astronomers and peers.

3 Dealing with Technical Evolution

Assessing and prototyping new techniques, and taking properly into account the evolution of the technical context, are major elements of the CDS long term sustainability.

The technical history of SIMBAD is a good illustration: it originates in the Catalogue of Stellar Identifications (CSI), which was created in 1971. Renamed SIMBAD (Set of Identification, Measurements and Bibliography for Astronomical Data) in 1981, it managed half a million object on a mainframe with two 28 Mbytes hard disks. The evolution of the hardware, software and languages, and technical constraints, are described in details by Wenger et al. in [6]. In particular, for many years hardware has been dependent on mainframe availability, and the software had to be ported from the French astronomy mainframe in Meudon Observatory (IBM 360/65) to the Strasbourg University Univac 1110, then to another Univac computer in Orsay University. In 1990 local control of the hardware was at last achieved and the database system was ported to workstations (SIMBAD III, which has been operational since 1990): first on Dec/Ultrix, and later to a Sun workstation. For CSI, a major specification was to be able to print the whole database on paper (for local consultation) and on microfiche (for distribution), and updating was an off-line, batch process. The database was then accessible from all French observatories connected to Meudon from leased lines. Updating became a real-time, interactive process in 1981. Interactive mode has evolved from command line language to Web queries. It also became possible, with the decrease of hardware cost and the increase in performance, to focus the development effort on functionalities and software architecture. The database system is currently evolving from a home-made software (which included some features of object orientation) to the open source database system PostgreSQL (SIMBAD IV).

Another example is the evolution of the catalogue distribution process, from tapes, floppies or listings mailed on written request, and later on e-mail request, to massive distribution by ftp and the Web since 1993.

Technological watch, prototyping and R&D actions often have a very positive impact on the services: for instance, Aladin was at the beginning the first exercise of Java implementation at CDS; later many features have been developed as prototypes in the frame of the European RTD project Astrophysical Virtual Observatory, and now of the VO-TECH Design Study. Another example is the Uniform Link Generator (Générateur de Lien Uniformes – GLU), a registry of distributed resources first implemented to avoid hardcoding of links in the CDS Web pages and services (see Fernique et al. [7]). It is now used e.g. to build links from Aladin (or VizieR) to distant archives and services. The *GLU resolver* allows the usage of symbolic names, instead of physical names, for the links; these names are then translated on the fly using the information contained in the *GLU Dictionary*. The GLU has been in usage for many years and is clearly a precursor of a VO registry, with many advanced functionalities (distributed architecture, replication, etc.).

4 The Importance of Partnership and Resource Networking

The context of astronomy is very favourable to the building of partnership between different entities, because there are few commercial constraints: information is in general made freely available (sometimes after a proprietary period). In addition, astronomers begun very early to define standards for data exchange: FITS (Flexible Image Transport System) was defined in 1979 for the exchange of radio and optical images on electronic media, and is currently the standard used everywhere in astronomy not only for data exchange, but also as the format of the observatory archives.

In this favourable context, CDS has been very active in the development of disciplinary standards, e.g. the description of catalogues and tables, first proposed by CDS, and now shared by the other data centres and most of the journals. The cooperation with journals, first with *Astronomy and Astrophysics*, which decided in 1993 that 'large' tables should be published by CDS in electronic form only, has led to a change in paradigm: the numbers which were printed in tables published in journals are now data in electronic form which can easily be downloaded by astronomers, accompanied by rationalized *metadata* which are essential for data usage. These metadata are also used to check the content of the tables (e.g., to detect if a number identified as a declination is outside its known boundaries), a verification which complements the referee's work, and thus improves significantly the quality of the published information. The excellent collaboration between the journals, ADS, CDS and NED, has been the key of the networking of bibliographic information, which is now used by astronomers in their everyday research work. This collaboration has generated

another disciplinary standard, the *bibcode* or *refcode*, to refer to a bibliographic reference (see Schmitz et al. [8]).

Other examples of partnership is the provision of the SIMBAD name resolver, a client-server package which is used by major observatory archives and the ADS to transform object names into the object celestial coordinates, or to list the references of the articles in which the object is cited, or more generally to link together astronomical services and observatory archives.

5 Relations with the Scientific Community

The relation of the CDS with its users and the astronomical community is at the core of the CDS role. At its beginning, these relations were essentially based on personnal contacts of the CDS Director and his scientific staff with their peers, to convince the astronomers to provide their data on electronic media for a large distribution; a journal, the *Bulletin d'Information du CDS*, was edited and distributed to the astronomical community, with articles presenting scientific results, but also news about astronomical data such as newly available data sets, lists of errata, discussions of incompatible data, etc. The data were distributed essentially on magnetic tapes, but also as microfiches, or as print-outs of extracts of SIMBAD.

The emergence of networks, and later of the web, had a deep impact on the relations of CDS with its users: the usage of the data – at that time essentially SIMBAD data – from terminals and later on the Web brought new kinds of users besides the professional astronomers, such as amateurs or the general public, and it became necessary to improve the data documentation. The usage of the data on a large scale has also an important impact on its reliability: errors are rapidly detected, reported via the hot-line, and corrected. The visibility of the databases also motivates specialists to bring their expertise to clean up important parts of the data, contributing to a better quality.

The visibility of the data at CDS is now considered an important feature by the astronomers: if at its beginning we had to convince the authors to supply a copy of their data to CDS for distribution, we now receive regularly requests from authors to include their results among the dataset distributed by CDS.

6 Conclusions

Since its creation in 1972, CDS has evolved from a service offered to a small number of scientists, with a network of personal or institutional contacts for gathering information, to become one of the major reference services of the astronomical network, by many aspects a precursor and now a major actor of the development of the International Virtual Observatory. The quality, motivation and diversity of the staff, the priority given to the quality of the service contents and software, the consensus-driven management of technical and conceptual evolution, and the constant search for networking and partnership, have

been determining factors for CDS long term success, as well as the long term vision of the Agencies which support it.

References

1. Genova, F., Egret, D., Bienaymé, O., Bonnarel, F., Dubois, P., Fernique, P., Jasniewicz, G., Lesteven, S., Monier, R., Ochsenbein, F., Wenger, M.: The CDS information hub. On-line services and links at the Centre de Données astronomiques de Strasbourg. Astron. Astrophys. Suppl. Ser. **143** (2000) 1–7
2. Wenger, M., Ochsenbein, F., Egret, D., Dubois, P., Bonnarel, F., Borde, S., Genova, F., Jasniewicz, G., Laloe, S., Lesteven, S., Monier, R.: The SIMBAD astronomical database. The CDS regerence database for astronomical objects. Astron. Astrophys. Suppl. Ser. **143** (2000) 9–22
3. Ochsenbein, F., Bauer, P., Marcout, J.: The VizieR database of astronomical catalogues. Astron. Astrophys. Suppl. Ser. **143** (2000) 23-32
4. Bonnarel, F., Fernique, P., Bienaymé, O., Egret, D., Genova, F., Louys, M., Ochsenbein, F. Wenger, M., Bartlett, J.G.: The ALADIN interactive sky atlas. A reference tool for identification of astronomical sources. Astron. Astrophys. Suppl. Ser. **143** (2000) 33–40
5. Lortet, M.-C., Borde, S., Ochsenbein, F.: Second reference dictionary of the nomenclature of celestial objects. Astron. Astrophys. Suppl. Ser. **107** 193–218
6. Wenger, M., Ochsenbein, F., Bonnarel, F., Lesteven, S., Oberto, A.: The SIMBAD database: lessons learnt from 30-year experience. ADASS XV, A.S.P. Conf. Ser. (to appear)
7. Fernique, P., Ochsenbein, F., Wenger, M.: CDS GLU, a tool for managing heterogeneous distributed web resources, ADASS VII, A.S.P. Conf. Ser. **145** 466–469
8. Schmitz, M. Helou, G., Dubois, P., LaGue, C., Madore, B., Corwin, H.G.Jr., Lesteven, S.. NED and SIMBAD convention for bibliographic coding. Information & On-line Data in Astronomy, Kulver Acad. Publ. (1995) 259 – 270