ELECTRONIC PUBLISHING: THE NEW ROLES OF CDS

F. GENOVA, J.G. BARTLETT, F. BONNAREL, P. DUBOIS, D. EGRET, P. FERNIQUE, G. JASNIEWICZ, S. LESTEVEN, R. MONIER, F. OCHSENBEIN AND M. WENGER
Centre de Données astronomiques de Strasbourg (CDS)
Observatoire de Strasbourg, URA CNRS 1280, Université Louis Pasteur, 11 rue de l'Université, 67000 Strasbourg, FRANCE

Abstract. The Centre de Données astronomiques de Strasbourg (CDS) has dealt with bibliographic information for many years. References of publications, published observational data related to objects, data tables, nomenclature, have been homogenized and organized into information retrieval systems: SIMBAD, the reference database for the identification and bibliography of astronomical objects; the catalogue service and the VizieR catalogue Browser, for data tables; the Dictionary of Nomenclature of Astronomical objects, which is now maintained by the CDS. Evolution in recent years has brought the Data Centers closer to the publishing process. General standards for astronomy, for the description of references and of data tables, have been proposed and implemented. Data tables from papers are now directly published in electronic form, and distributed on-line by the Data Centers. The emergence of fully electronic publication paves the way to innovative new services, linking the journals to other sources of informations (data bases, tables, then images, data archives), and making use of new methods for information retrieval. This also has an effect on the publishing process, with the possibility to implement new checks and links from text to other kinds of information (from objects names, positions, etc.). The CDS will bring some of the key features in the evolution towards a fully linked astronomy information system, in close collaboration with the journal editors, the ADS, the other Data Centers, and the data providers.

1. Introduction

The *Centre de Données astronomiques de Strasbourg* was founded in 1972, by the French *Institut National des Sciences de l'Univers* (INSU, which was then the INAG - *Institut National d'Astronomie et de Géophysique*). At that time, it was called the *Centre de Données Stellaires*, and its objectives were:

- to collect 'useful' data about astronomical objects, in electronic form;
- to improve these data by critical evaluation and combination;
- to distribute the results to the international community; and
- to conduct research programs using these data.

For this purpose, the CDS was from the beginning imbedded in a research environment at the *Observatoire de Strasbourg*. During the first years, it dealt solely with stellar data (hence Stellar Data Center) to study the galactic structure.

The CDS aims were somehow extended in 1983, when it began to deal with all astronomical objects (except the solar system ones), and was then renamed *Centre de Données astronomiques de Strasbourg*. The CDS objectives can now be summarized by: collect, homogenize, distribute, preserve astronomical information, for the usage of the astronomical community all over the world. It is worth noting that the primary important functions defined at the very beginning, have been kept in the somehow more general present definition of the CDS activities: to deal with information in electronic form, to develop and implement expertise about this information, and to serve the international scientific community in astronomy.

Another important point is that the CDS does not deal directly with observational data, but rather builds 'metadatabases', containing high-level selected, organized, homogenized information.

J. Andersen (ed.), Highlights of Astronomy, Volume 11A, 520–525.

In the following, the CDS activities linked to the bibliography, and their evolution with the rapid development of electronic journals, will be described.

2. Bibliographic information services at CDS

A general description of the CDS services can be found e.g. in Egret *et al.* (1995) and Genova *et al.* (1996), or at the CDS website,[1] which is regularly updated and gives access to the on-line WWW services. In the following, one will mainly describe the service contents linked to the bibliography. Bibliographic information are displayed in two of the main CDS services: SIMBAD and the catalogue/VizieR service.

2.1. SIMBAD

SIMBAD is the reference database for the identification and bibliography of astronomical objects (outside the solar system). It contains identifications, 'basic data', bibliography links, and some measurements, for more than 1,500,000 objects (September 1997). This information is extracted from published papers and selected catalogues, through a large set of collaborations. The scanning of bibliography is a very large effort, involving several French institutes: 90 journals are systematically screened, in partnership with the *Institut d'Astrophysique de Paris*, and the Paris (DASGAL) and Bordeaux Observatories. References, object names, 'basic data' such as position, object type, magnitude, and so on are systematically recorded by experienced staff which scan the published papers (Laloë *et al.*, 1993; Laloë, 1995). In parallel, long tables in electronic form are now more and more often entered by semi-automatic procedures.

In practice, when dealing with a published paper, the SIMBAD 'bibliographers' will for instance:

- find the objects cited in the paper;
- interpret the origin of the acronyms mentioned in the paper (an object called "M 13" is not always from the Messier catalogue!);
- identify new lists of objects in the paper, then create the corresponding acronym, and describe it in the *Dictionary of Nomenclature of Astronomical Objects*;
- identify errors on object names in the paper, and collect this information in a note appended to the reference in SIMBAD.

The *Dictionary of Nomenclature of Astronomical objects*, developed for many years by M.-C. Lortet and her collaborators (Lortet *et al.*, 1994), is now fully maintained by the CDS, in collaboration with the Paris Observatory (DASGAL), as a by-product of SIMBAD bibliography scanning. Older lists cited in recent papers are also tracked back during the process. An on-line version, updated every week, is available in the CDS WWW service, together with the *Specifications concerning designations for astronomical radiation sources outside the solar system*, from the Task Group on Astronomical Designations of IAU Commission 5 which provides basic advices in this topic, and the *Advance registry for acronyms* service.

2.2. THE CATALOGUE/VIZIER SERVICE

This service contains large catalogues and published tables, which are particularly important as 'metadata', since they contain information calibrated and homogenized by the authors. This is a co-operative action between the CDS and the other Data Centers, ADC (GSFC), INASAN (Moscow), NAOJ (Tokyo) and Beijing, with also a regional center for India at IUCAA in Pune. The Data Centers share the same data sets and collaborate in the building up of the contents.

The key to this cooperation is a general standard for the description of the tables, proposed by CDS (Ochsenbein 1994)[2] and now shared by the Data Centers and several major journal editors (*Astronomy and Astrophysics*, the *American Astronomical Society* on its CD-ROM with the tables from the *Astrophysical Journal*, the *Astronomical Journal*, and the *Publications of the Astronomical Society of the Pacific*, *Soviet Astronomy*). The standard description is both fully readable by the user, and an important tool for electronic checks, transformation and exchange of data. It describes all the files which constitute the catalogue, and the byte-per-byte contents of each file. Tools have

[1] http://cdsweb.u-strasbg.fr/CDS.html
[2] http://vizier.u-strasbg.fr/doc/catstd.htx

been developed to manage the units, check that the contents of the tables is coherent with their description (e.g., that declination values are between −90 deg and +90 deg), and transform the tables to FITS format.

In practice, the staff activities for the catalogue service are the following:

- create or improve the description of a table;
- check the consistency of data in a table; and
- discuss with the author when discrepancies are found.

In September 1997, the catalogue service contains more than 2,350 catalogues, of which more than 1,800 are available on-line as full ASCII or FITS file. Several hundreds tables published in journals are added each year, mostly by agreement with the journal editors as explained below. Nearly 1,450 tables are also available through the VizieR browser, developed in 1995 in collaboration with ESA-ESRIN as a follow-up of the ESIS Catalogue Browser, and now fully maintained by the CDS (Ochsenbein, 1996). The standard description is used in the pipeline to include the ASCII tables in the Sybase VizieR database, which makes it very easy to include new catalogues and tables in the system (just build the standard description). Tables in VizieR can be queried by constraining any of their field, or for data around one position, or around one object (the position being then given by SIMBAD).

Plans are now to include logs of ground-based and space-borne observatories in the system, and to build active links from the VizieR access tool to observatory on-line archives.

3. Recent evolutions and perspectives

As shown in the previous section, the CDS has been dealing with information from the bibliography from the very beginning, and SIMBAD (like the NED database for extragalactic objects) and the catalogue service are high value-added services to retrieve checked, homogenized, standardized bibliographic information.

The recent, rapid evolution towards electronic publication has brought the Data Centers much closer to the publishing process. From 1993 on, the CDS has been building on-line tables (which are in general not printed any more) for *Astronomy and Astrophysics* and the *Supplement Series* (Ochsenbein and Lequeux, 1995). On-line abstracts are also prepared for these journals, to which active links to references, objects (from SIMBAD), and soon images (from ALADIN - Bonnarel *et al.*, 1997), are progressively added. In parallel, and by agreement with the *American Astronomical Society*, the CDS is also installing the tables from the AAS CD-ROMs on-line. In addition, the Russian Center is now building the tables from *Soviet Astronomy*, by agreement with the journal editor. As explained above, due to the agreement of all the partners on a common standard, all these tables can be put together in a dataset common to all the Data centers, and can easily be exchanged and transformed. The tables are accessible to users from all over the world thanks to the collaboration between the Data Centers, and information retrieval services such as the VizieR Catalogue Browser also constitute a new, powerful way to access published information.

A very appealing opportunity offered by the WWW, is the possibility to build links between distributed, heterogeneous on-line services. The user can already navigate from one object in SIMBAD, to the references citing the object, then to the ADS services for each reference and to the full paper from the editors when available; or from one SIMBAD object, to the *Dictionnary of Nomenclature* information about one of the object names, to the list of origin of this name in the catalogue service, to the reference containing the list, etc... Reciprocally, from one reference found for instance in the ADS, one can go to the list of SIMBAD objects from this reference, and then to SIMBAD for each of these objects. Or the ADS can be queried by an object name, and then the list of references is retrieved from SIMBAD and NED (through a client/server procedure) and processed by the ADS. The links between bibliographic information and databases are build in particular by regular exchange of information between CDS and ADS (Eichhorn 1997).

Another important tool for navigation among on-line bibliographic information is the bibcode/refcode, a 'de facto' standard first defined by the CDS and NED, then extensively used by the ADS, and shared by the journal editors. One reference is uniquely described by 19 characters, easily readable (at least for the references from journals) (Schmitz *et al.*, 1995).

In the frame of the rapid development of electronic bibliographic services, more and more linked to the Data Center information, the CDS now hosts the European mirror of the ADS NASA

bibliographic database (the CfA is implementing a mirror copy of SIMBAD for US usage), and the European mirror of the electronic *Astrophysical Journal*. The possibility of hosting the long term archive of the electronic *Astronomy and Astrophysics* is being discussed with the journal editors.

This evolving context offers good opportunities to develop innovative services. Recognition of object names in the published texts, the possibility of linking the text of journals to databases, images (e.g. through the ALADIN service), etc., is certainly very appealing. The first attempts were done in the electronic version of *New Astronomy* (by visual detection and individual check), and in the CDS bibliographic service (for objects in the keyword list, by semi–automatic detection and check). CDS is developing automatic procedures to recognize object names, and one can also recommend the editors to ask the authors to tag the object names themselves when submitting a paper in electronic form. The experience with the scanning of the objects in published papers for SIMBAD, is that the different methods will be complementary, and that a final check by an expert will probably remain necessary for completeness and correctness, mainly because of the complexity of astronomical nomenclature, and also unfortunately because of the carelessness of too many authors in these matters.

In parallel, a method of organizing sets of documents by the similarity of their contents, by using neural network techniques, is being developed (Kohonen maps, Lesteven *et al.*, 1996). The neural network is built from the keywords attached to the papers. The common list of keywords shared by *Astronomy and Astrophysics*, the *Astrophysical Journal*, and the *Monthly Notices of the Royal Astronomical Society*, together with the fact that *Astronomy and Astrophysics* and the *Astrophysical Journal* provide CDS with electronic reference information including the keywords, has proven to be a very important tool to build a service common to several journals.

4. A very practical wish–list

An IAU Joint Discussion is certainly a good opportunity to present a wish–list, so here is a list of very practical points (some of them discussed in the various Commission 5 meetings during this General Assembly), which could help the future development of fully linked astronomy, from observational data to published papers. Many of these points refer to 'de facto' exchange standards, which already play a tremendous role in the existing links and collaborations.

- to have more journals collaborate in the *Astrophysical Journal*, *Astronomy and Astrophysics* and *Monthly Notices of the Royal Astronomical Society* common list of keywords;
- to have more journals collaborate to the on–line published table service;
- to get tags of object names in the text of published papers, from the authors;
- to add the origin of observations (number of observation run, etc.) as an additional 'keyword' to the published papers, to open the possibility of building a direct link between one observation and the resulting published research results. This should also imply the definition of a 'standard description' to identify observations (observation # N with one instrument in one ground–based or space–borne observatory);
- to get the recognition of the bibcode as the 'de facto' standard for describing references in astronomy, and to improve the procedures to maintain this standard.

IAU Commission 5, as shown in the Joint Discussion, and in the meetings during this General Assembly, and the Urania initiative, are certainly good forums in these matters. Commission 5 proposes for instance that the bibcode standard be under the responsibility of its Working Group on Information Handling.

5. Conclusions

For the future, one can certainly expect more links between distributed services. The users can already use the first new paths and new tools to access published information, as shown above. An important objective is now to include access to the observatory archives (i.e. to observational results) in the landscape of the fully linked astronomy.

The management of links (in particular keeping track of their evolution) is certainly a key question here. The CDS has developed a solution to manage its own set of heterogeneous, distributed services and mirror sites, the GLU (Générateur de Liens Uniformes – Wenger *et al.*, 1996), which

avoids storing hard-coded URLs, permitting the resolution of URL names, and which might be used in a wider context.

The coming of electronic publication already produces an important evolution of the journal editor procedures, as explained elsewhere in these Proceedings. The refereeing process is also be expected to evolve since new checks are and will be possible (e.g. the new checks operational already for the table contents). Clearly also, the internal Data Centers procedures will have to be adapted, towards more automated processes, but preserving steadily the expert intervention, and the building of high value-added 'meta-information'. The Data Centers will certainly continue to jump at the opportunity to give access to the wealth of new information available, and to build up innovative information retrieval services. They have some of the keys for the future development of new links, with the 'metadata' services as bridges between different types of information, and information retrieval tools that can be linked to distant services. They will also continue to keep on other essential functions of the Data Center 'mission', push and participate to the definition of standards, and to the development of international cooperation.

References

Bonnarel, F., Ziaeepour, H., Bartlett, J.G., Bienaymé, O., Crézé, M., Egret, D., Florsch, J., Genova, F., Ochsenbein, F., Raclot, V., Louys, M. and Paillou, Ph. (1997) The Aladin Interactive Sky Atlas, *IAU Symp. 179, New Horizons from Multi-Wavelength Sky Surveys*, Kluwer Academic Publishers.

Eichhorn, G., Kurtz, M.J., Accomazzi, A., Grant, C.S., Murray, S.S. (1997), Connectivity in the Astronomy Digital Library through the ADS, this volume.

Egret, D., Crézé, M. , Bonnarel, F., Dubois, P., Genova, F., Jasniewicz, G., Heck, A., Lesteven, S., Ochsenbein, F. and Wenger, M. (1995) A global perspective on astronomical data and information: the Strasbourg astronomical data centre (CDS), *Information & On-line Data in Astronomy*, Egret & Albrecht (Eds.), Kluwer Academic Publishers, pp. 163–174.

Genova, F., Bartlett, J. G., Bienaymé, O., Bonnarel, F., Dubois, P., Egret, D., Fernique, P., Jasniewicz, G., Lesteven, S., Monier, R., Ochsenbein, F. and Wenger, M. (1996) CDS as an Astronomical Information Hub, *Vistas in Astronomy*, **40**, pp. 429–437.

Laloë, S., Beyneix, A., Borde, S., Chagnard-Carpuat, C., Dubois, P., Dulou, M.R., Ochsenbein, F., Ralite, N. and Wagner, M.J. (1993) Updating the bibliography in SIMBAD, *CDS Inform. Bull.*, **43**, pp. 57–63.

Laloë, S. (1995) The updating of the bibliography in the SIMBAD database, *Vistas in Astronomy*, **39**, pp. 259–270.

Lesteven, S., Poinçot, Ph., Murtagh, F. (1996) Neural networks and information extraction in astronomical information retrieval, *Vistas in Astronomy*, **40**, pp. 395–400.

Lortet, M.C., Borde, S., and Ochsenbein, F. (1994) Second Reference Dictionary of Nomenclature of Celestial Objects, *Astron. Astrophys. Suppl.*, **107**, pp. 193–218.

Ochsenbein, F. (1994) Adopted standards for catalogues at CDS, *CDS Inform. Bull.*, **44**, pp. 19–28.

Ochsenbein, F. (1996) VizieR, the new catalogue interface at CDS , *CDS Inform. Bull.*, **48**, pp. 47–50.

Ochsenbein, F. and Lequeux J. (1995) The A & A tables and abstracts: An example of collaboration between data centres and editors, *Vistas in Astronomy*, **39**, pp. 227–234.

Schmitz, M., Helou, G., Dubois, P., LaGue, C., Madore, B., Corwin, H.G. Jr and Lesteven, S. (1995) NED and SIMBAD conventions for bibliographic reference coding, *Information & On-line Data in Astronomy*, Egret & Albrecht (Eds.), Kluwer Academic Publishers, pp. 259–270.

Wenger, M., Fernique, P., Genova, F., Bartlett, J.G., Bienaymé, O., Bonnarel, F., Dubois, P., Egret, D., Jasniewicz, G., Lesteven, S., Monier, R. and Ochsenbein, F. (1996) SIMBAD on the Web and Links to other Services, *BAAS*, **189**, #06.02.

Comments

BOYCE: I am impressed with two things you said. 1. The degree to which you have to automate your functions and 2. The advent of new tools, such as Kohonen maps, which are new tools available only in the electronic environment.

GENOVA: 1. Yes, it is very important to automate the functions as much as possible, and to concentrate the work of experienced staff on value-added activities. This is also the only way to keep pace with the increasing volume of published literature.

2. There will certainly be many new tools thanks to the electronic environment, and this might increase the acceptability of electronic publishing for the scientific community.

HJELMING: The comment about supplying accurate postitions of X-Ray sources is related to the management of misinformation. Often errors are large and poorly described. Dealing with this, and changes, is critical. Often only the scientist can make the judgement among diverse data with errors.

GENOVA: Dealing with errors and changes is certainly a very important task for the data centers. For instance, SIMBAD is continuously updated for cross-identification purposes, and notes

and errors are added to the references. Active links can help to give access to checked and/or updated information.

CHEUNG: The Astrophysics Data Facility (ADF) at NASA's GSFC is actively working on the problem of correcting designations of objects identified from NASA missions to astronomical nomenclature, and building tools to help the researcher do object cross identification. See the NASA project (http://amase.gsfc.nasa.gov) homepage.

https://doi.org/10.1017/S1539299600021924 Published online by Cambridge University Press