Astronomical Data Analysis Software and Systems XVI
ASP Conference Series, Vol. 376, 2007
R. A. Shaw, F. Hill and D. J. Bell, eds.

The Many Faces of SIMBAD

Françoise Genova

CDS, Observatoire astronomique de Strasbourg, Strasbourg, France

Abstract. SIMBAD (*Set of Identification, Measurements, and Bibliography for Astronomical Data*), which begun as the *Catalogue of Stellar Identifications* in 1971, is one of the essential elements of the astronomical bibliographic information network. This paper describes how SIMBAD has been built and is maintained, in terms of software and content, its evolution in time and its perspectives, and the complementarity with the other CDS services. Critical elements for long term sustainability are also discussed.

1. Introduction

SIMBAD is, together with VizieR and Aladin, one of the main services developed by the *Center de Données astronomiques de Strasbourg* (CDS, Genova et al. 2000). CDS was created in 1972 by the French agency in charge of ground-based astronomy, which was at that time called INAG (*Institut National d'Astronomie et de Geophysique* - National Institute for Astronomy and Geophysics) and is now INSU (*Institut National des Sciences de l'Univers* - Universe Sciences National Institute), as a joint venture with Strasbourg Louis Pasteur University. CDS was then the *Centre de Données Stellaires* (Stellar Data Center). The initial CDS charter is still relevant and still guides the data center strategy:

- collect useful data on astronomical objects, in electronic form—a far-seeing vision in 1971!
- improve them by critical evaluation and combination
- distribute the results to the international community
- conduct research using these data

At that time the objective was to collect stellar data to study the galactic structure. Since the beginning, CDS thus not only dealt with data curation, but also with building local expertise on data and creating and maintaining added-value services in order to provide science tools for the astronomical community. In 1983, the decision was taken to include all objects beyond solar system in SIMBAD, and CDS became the *Centre de Données astronomiques de Strasbourg*, with the high level objective of *collecting, homogenizing, preserving and distributing astronomical information for the usage of the whole astronomy community.* Note that the word "data" from the initial charter was then replaced by the more general word "information."

This paper will describe SIMBAD history (§2) and its recent and foreseen evolution (§3). The keys for long term sustainability will be discussed in §4.

2. SIMBAD History

The development of SIMBAD began in 1971, even before the official creation of CDS, as the *Catalogue of Stellar identifications* (CSI: Jung & Bischoff 1971; Jung & Ochsenbein 1972). The CSI aimed at cross-identifying a few fundamental stellar catalogs (HD, SAO, GC, etc.). In parallel, the *Bibliographic Star Index* (BSI: Cayrel, Jung & Valbousquet 1974) gathered star bibliographies from published papers. SIMBAD, for *Set of Identification, Measurements, and Bibliography for Astronomical Data*, was created in 1981 from the merging of the CSI and of the BSI, and was extended to extragalactic objects in 1983 to provide an homogenized view of astronomical objects across astronomy sub-disciplines.

2.1. SIMBAD Hardware and Software Systems

Since 1971 SIMBAD hardware and software have of course evolved a lot. An overview of SIMBAD is given by Wenger et al. (2000), and a detailed description of this evolution can be found in Wenger et al. (2006). From the beginning the system was designed to be remotely queriable, by means which evolved in time from batch punch-cards, to interactive modes, through client/server and the Web. There were four main releases, in 1971, 1981, 1990 and 2006 (SIMBAD 4), with hardware that evolved from IBM mainframes, to Unix stations, to Linux PCs, and a database system that was initially IBM-dependent, then homemade, then an object-oriented DBMS, and now PostgreSQL. During its long life SIMBAD progressively gained independence from hardware, operating systems, DBMS, vendors, and also from its developers.

SIMBAD 4, available in a beta-test version since the beginning of 2006 September, provides more flexibility to include new kinds of data and to implement new functionalities. All parameters, instead of only a few, are searchable. In spite of the major software evolution and new capabilities, it has been decided to keep the user look-and-feel not too different from SIMBAD 3. Another improvement, hidden from the users but very important for the SIMBAD team, is the implementation of a graphical interface for updating, which will facilitate the work of SIMBAD librarians and astronomers compared to the previous line-by-line, interactive updating.

2.2. SIMBAD Contents

SIMBAD contents are built through the daily work of a team of specialized librarians (the French name of this specialty, *documentalistes*, tells well that they specialize in extracting information from, and building metadata for, documents) and astronomers who provide scientific expertise when needed—e.g., to clarify new astronomical concepts, or to help deciphering the too-frequent, badly written, unclear papers, or lax object nomenclature.

Two main sources of information are used: the systematic scanning of papers published in about 90 journals, and reference catalogs and tables. The baseline strategy is to enter all objects quoted in the text of published papers, with references to the papers in which the objects are cited. Especially long published tables and catalogs follow another path: they are first documented and entered in VizieR, a step now performed in collaboration with several journals and data centers. Selected contents of this tabular material are then entered

in SIMBAD by semi-automated procedures, with systematic cross-identification with existing objects and a check of the data homogeneity by an expert. Not all tables are entered in SIMBAD because of insufficient manpower, and the decision is taken on a case by case basis.

Systematic "cleanings" of the somehow heterogeneous data entered from the literature are also undertaken when appropriate. For example, all X-ray objects present in SIMBAD were examined a few years ago before the launch of the *Chandra X-ray Observatory* and *XMM-Newton* in order to improve the database, with the aim of helping astronomers interpret observations from these satellites, and to be ready for the inclusion of new information published from their observations.

The database also contains additional information that was recorded when scanning the literature. For many years now the team has been adding notes to bibliographic references to keep track of nomenclature, of tables available in VizieR, of errors found in papers, etc. One significant and recent addition are *notes* attached to objects which contain free text that describes important information about the object.

2.3. SIMBAD in the CDS Hub

SIMBAD has never been an isolated service. From the beginning, CDS has gathered two main types of information: object cross-identifications and bibliography (hence the *CSI* and *BSI*, then *SIMBAD*) on one hand, astronomical catalogs and their descriptions on the other. Catalogs and their description have been produced in collaboration with other data centers, and later also directly with the journals. They have been distributed in the *catalog service*, to which a browsing capability, *VizieR*, was added from 1996 (Ochsenbein, Bauer & Marcout 2000). The *Aladin* visualizer and reference images stored by CDS were made available in 1998 (Bonnarel et al. 2000).

The flow of data about astronomical objects has changed scale rapidly, with the advent of the very large surveys and the fast increase in the number and size of object lists that are published in the literature. This is managed by taking advantage of the complementarity between SIMBAD and VizieR. In addition, Aladin allows visualization and comparison with distributed archives and services. Another service that gathers information from the literature is the *Dictionary of Nomenclature of Astronomical Objects* (Lortet, Borde & Ochsenbein 1994), a by-product of SIMBAD bibliography scanning. It contains complementary additional information applicable to all the objects with a given acronym, in particular the hierarchy (e.g. the cluster name for cluster stars or galaxies), the object type, the telescope used to obtain the data, etc.

2.4. Partnership and Interoperability with Other Services

Building partnerships has been one constant of the CDS strategy in the long term. A tight collaboration with other data centers was set from the early days for the production and documentation of catalogs in electronic form. For instance, the second issue of the *CDS Bulletin* in 1971 December included a paper by J. Mead about NASA Goddard Space Flight Center catalogs, and the third issue in 1972 July a paper by C. A. Murray about RGO catalogs. Later came the collaborations with NED and the ADS, with journals, and with archive

providers. The advent of the Internet added a new dimension, and the capability to build links between on-line services has been widely used, in particular to network bibliographic resources (SIMBAD, ADS, NED, electronic journals). The partnership with electronic journals has been of particular importance in the two core CDS activities: for instance the provision of electronic tables by the journals has permitted us to concentrate the CDS tasks on adding value to published information, such as describing the table content and using this description to check the data, saving OCR (Optical Character Recognition) time. On the other hand, tools provided by CDS such as the SIMBAD name resolver, implemented in 1992, have been widely used by the observatory archives and the ADS. Aladin is now also a common tool for visualization purposes in observation preparation tools and on-line observation archives.

Interoperability had been an issue for CDS long before the Internet, because of data exchange with the partners. For instance, NED and SIMBAD defined in common the bibcode/refcode, a 19-character coding of published references (e.g., `2006A&A...447...89T`) in 1989 (Schmitz et al. 1995). This was later heavily used and extended by the ADS and the electronic journals. The bibcode has been the key facilitator of the very rapid networking of astronomical bibliographic resources, which was implemented several years before publishers finally agreed on a common standard, the Digital Object Identifier (DOI). The bibcode, although it struggles to cover the full complexity of describing references, is still in use in parallel to DOI and keeps the advantage of being human readable.

3. Key Evolution

3.1. Increased Flexibility

One requirement on SIMBAD 4 was to be more flexible than the previous version for users, for software developers, and for the librarians who feed the database. Another objective was to take advantage of this flexibility to offer progressively more information to users. For instance, the object type is stored in SIMBAD when the object is created in the database, and updated when additional information is found in the literature. In addition, every time a new object list is recorded in the Dictionary of Nomenclature, the object type common to the objects in the list is determined and stored in the Dictionary. For each object, the new SIMBAD version displays the list of object types attached to all the object names, which is built dynamically by querying the Dictionary of Nomenclature for all the object acronyms. This gives in particular some insight on the wavelength domains in which the object emits, and helps users to make up their own minds about the object type when there is conflicting information that is not well rendered by the unique SIMBAD object type.

One has to keep in mind however that with quality in mind, the addition of new information may require some validation, and that implementation may thus require time and effort, even when the functionality is developed. For instance, SIMBAD 4 could display the hierarchy between objects. Information about a hierarchy is found in the Dictionary of Nomenclature (if the list of objects is included in another object, the `Object in...` indication appears). The basic idea is to use this information to code the hierarchy in SIMBAD, but this will

require developing specific validation procedures (e.g., objects "in the field of a cluster" can be cluster members, or not) which are being assessed.

3.2. Dynamical Cross-Match

In addition to the cross-matches gathered in SIMBAD from the literature search, we aim at proposing possible cross-matches dynamically, browsing catalogs that are available in VizieR and user-provided catalogs. A cross-match tool has been developed in Aladin in the framework of the *Astrophysical Virtual Observatory* scientific demonstrations, and will be soon implemented as a separate tool in the VO-TECH project. A specific development for radio sources is also under way. SPECFIND (Vollmer et al. 2005) is a cross-match tool for radio catalogs, which takes into account the source physical parameters and delivers cross-matches, a hierarchy, and associations between objects. This is being extended in the VO context, again in the framework of the VO-TECH project, to become a *dynamic* general SED builder for radio using VO technologies, namely the Uniform Content Descriptors (UCDs) and the Registry. Again the implementation of this new functionality requires additional work on the database content, since new metadata have to be included in VizieR to characterize the observations (e.g., the instrument observation lobe).

3.3. Other R&D Programs of Importance for SIMBAD

Another important constant of CDS activity on the long term has been the importance given to monitoring technology, and to research and development (R&D). In recent years, the Virtual Observatory context has been an excellent opportunity to develop new R&D activities. In particular, the VO-TECH project offers a European framework for programs in "Intelligent Resource Discovery" (Design Study 5, chaired by S. Derriere, CDS). Two on-going R&D programs of VO-TECH Design Study 5 are of potential importance for SIMBAD. First, the use of an ontology of object types, based on the list of SIMBAD object types, complemented by a full description of relations between the ontology concepts, is being assessed for information retrieval (study led by A. Preite-Martinez, INAF: see Derriere, Richard, & Preite-Martinez 2007). Second, semi-automated recognition of object names in articles is being tested, taking advantage in particular of the information contained in the Dictionary of Nomenclature. If difficulties due to the careless usage of nomenclature by astronomers in their papers are finally overcome, the method will be used to help identify objects cited in published papers for SIMBAD. But it appears already very clear that validation by experts would still be absolutely needed. Another product of the study could be the implementation of proper links between NED and SIMBAD, taking into account the differences in nomenclature (which are tracked, thanks to the collaboration between the two data centers).

4. Long-Term Sustainability

4.1. Dealing with the Increase in Data Volume

One critical point to ensure the long term sustainability of CDS has been to develop an ability to deal with the endless and ever increasing data flow without

sacrificing quality. There was no pre-defined model and several paths have been explored, in particular using the complementarity between the CDS services as explained above. Constant attention is also paid to seize all occasions to evolve the procedures to focus staff work on added-value tasks. For instance, the journals agreed to provide an electronic version of tables contained in published papers shortly after they began to implement an electronic version. This saved typing time and avoided introducing errors into SIMBAD references, which increased the global quality of the database. But all references are still checked with the printed publication to correct residual errors. Similarly the *raccord* program compares information from published tables with information contained in SIMBAD, and suggests information that could be integrated into the database (a new object, a new name for an existing object, etc.); the results are then evaluated by an expert librarian, with additional consultation of an astronomer for difficult cases.

4.2. Dealing with Evolution

Another critical issue in the long term has been to deal with the constant evolution of astronomy, of technology, and of the political context of CDS activities. Very different activities have to be managed, each coming with its own time scale(s) and constraints: building the content; developing the databases and user interfaces; operating a complex system; monitoring pertinent technology, and R&D activities (in spite of the operational pressure on the staff), and dealing with schedules imposed by collaborations with external partners or participation to projects. This means, quoting N. Radziwill's (2007) contribution to this conference, a strategy that is "agile and ready to respond to unexpected situations," characterized by "awareness."

4.3. The Good Use of R&D

The key driver of CDS is to provide the astronomical community with useful services. It must be and remain science-driven to gain and keep community support in the long term. But it is also mandatory to implement a proper R&D program, to ensure the long term technical pertinence of the services. The path is narrow and there is a real risk of shifting progressively towards a technically driven strategy: the CDS R&D objective is to develop innovative ways to improve the services. It is very good if this also brings technologically interesting advances but that is not the primary goal. A proper balance has to be found between risk and innovation, and so new technologies have to be implemented not too late, to fulfill users' expectations, but also not too early—only when they are mature enough to be used operationally. For instance, the two projects led by CDS in collaboration with other astronomy laboratories and research teams specialized in information technologies selected in a French national IT Call for Proposals, IDHA (*Images Distribuées Hétérogènes pour l'Astronomie—Distributed Heterogeneous Images for Astronomy*) and MDA (*Massive Data in Astronomy*), have had important spin-offs. IDHA triggered the development of what has now become the IVOA "Characterization" Data Model. MDA triggered work on ontologies that are now being followed up in VO-TECH, and that led to the development of new methods for hyper-spectral data analysis (see the

report on the Ph.D. work of M. Petremand by Genova et al. 2007) and to the development of the *AÏDA* work-flow management system (Schaaff et al. 2006).

5. Conclusion

The main lessons learnt with the passing years about data center management are that *quality is a must*, and that in such a long-term endeavor *routine is the worst enemy*—even more than the ever increasing data flow. The critical importance of *integrating* a team of astronomers, specialized librarians and software engineers has also been demonstrated: the strategy is defined by taking into account the scientific and technical points of view. The added-value tasks of building the databases content has also proven to be an excellent specialization for librarians in a context where they are losing their traditional tasks. One very important driver in the last years has also been the transmission of expertise and the evolution of software and procedures (e.g., the major software evolution of SIMBAD and VizieR) to hand the torch to the next generation.

The present and the future of CDS are also of course shaped by its participation to the Virtual Observatory. The basic keywords of CDS activity, provision of value-added services and tools and interoperability, fit extremely well with the VO concept, which CDS has in a sense anticipated. SIMBAD and the other CDS services are major building blocks of the VO, Aladin is one of the few VO portals, and the team is actively participating in the International Virtual Observatory Alliance, Euro-VO and French VO activities to convey its expertise and collaborate on the VO definition. The VO development is a change in scale and a remarkable new opportunity, bringing new partners with whom to collaborate, more data and more services around which to interoperate, new functionalities in the CDS services that take advantage of the development of the VO interoperability standards, and new types of usage.

Acknowledgments. Thanks to P. Dubois, F. Ochsenbein, A. Preite Martinez, and M. Wenger for their comments on this manuscript.

References

Bonnarel, F., et al. 2000, A&AS, 143, 33

Cayrel, R., Jung, J., & Valbousquet, A. 1974, BICDS, 6, 24

Derriere, S., Richard, A., & Preite-Martinez, A. 2007, in ASP Conf. Ser. 376, ADASS XVI, ed. R. A. Shaw, F. Hill, & D. J. Bell (San Francisco: ASP), 607

Genova, F., et al. 2000, A&AS, 143, 1

Genova, F., Petremand, M., Bonnarel, F., Louys, M., & Collet, C. 2007, in ASP Conf. Ser. 376, ADASS XVI, ed. R. A. Shaw, F. Hill, & D. J. Bell (San Francisco: ASP), 297

Jung, J., & Bischoff, M. 1971, BICDS, 2, 8

Jung, J., & Ochsenbein, F. 1972, BICDS, 3, 6

Lortet, M.-P., Borde, S., & Ochsenbein, F. 1994, A&AS, 107, 193

Ochsenbein, F., Bauer, P., & Marcout, J. 2000, A&AS, 143, 23

Radziwill, N. 2007, in ASP Conf. Ser. 376, ADASS XVI, ed. R. A. Shaw, F. Hill, & D. J. Bell (San Francisco: ASP), 363

Schaaff, A., Bonnarel, F., Claudon, J.-J., Louys, M., Pestel, C., David, R., Genaud, S., & Wolf, C. 2006, in ASP Conf. Ser. 351, ADASS XV, ed. C. Gabriel, C. Arviset, D. Ponz, & E. Solano (San Francisco: ASP), 323

Schmitz, M., Helou, G., Dubois, P., LaGue, C., Madore, B., Corwin, H. G., Jr., & Lesteven, S. 1995, in Information & On-Line Data in Astronomy, ed. D. Egret & M. A. Albrecht (Dordrecht: Kluwer), 259

Vollmer, B., Davoust, E., Dubois, P., Genova, F., Ochsenbein, F., & van Driel, W. 2005, A&A, 431, 1177

Wenger, M., et al. 2000, A&AS, 143, 9

Wenger, M., Ochsenbein, F., Bonnarel, F., Lesteven, S., & Oberto, A. 2006, in ASP Conf. Ser. 351, ADASS XV, ed. C. Gabriel, C. Arviset, D. Ponz, & E. Solano (San Francisco: ASP), 662