LIBRARY AND INFORMATION SERVICES IN ASTRONOMY VI
ASP Conference Series, Vol. 433
E. Isaksson, J. Lagerstrom, A. Holl, and N. Bawdekar, eds.

Information Scientists: Between Editors and Data Centers

M. Brouty,[1] F. Woelfel,[1] C. Bruneau,[1] C. Brunet,[1] H. Claude,[1] P. Dubois,[1] A. Eisele,[1] F. Genova,[1] S. Lesteven,[1] M. Neuville,[1] F. Ochsenbein,[1] E. Perret,[1] P. Vannier,[1] P. Vonflie,[1] and G. Chassagnard[2]

[1]*CDS, Observatoire astronomique de Strasbourg, UMR CNRS/ULP 7550, 11 rue de l'Université, 67000 Strasbourg, France*
[2]*Institut d'Astrophysique de Paris, 98bis Boulevard Arago, 75014 Paris, France*

Abstract. Since the emergence of electronic publications in the early 1990s, astronomy has played a pioneering role in the development and implementation of new capabilities and services.

As a data center, the CDS contributed significantly to this evolution: a synergy between data centers and journal editors started in the 1990s with the publication of large tables and data sets in electronic form and contributed to an efficient linking of publications with existing databases like SIMBAD or NED. This collaborative work, carried out in practice by information scientists, illustrates a new role for us who now have to deal with both editor and database requirements.

After a short description of the CDS, we present our peculiar responsibilities related to the publication process: ensuring, prior to publication, that the link from selected objects quoted by the authors in their papers to the SIMBAD database is correct and maintained in the long term, that the tables and their complete descriptions are accessible through VizieR, and that the data and bibliography are correctly entered in SIMBAD. The *Dictionary of Nomenclature*, which plays an important role in these procedures, is briefly presented. Finally, the skills we developed for these activities are shortly discussed.

1. Introduction

The Centre de Données de Strasbourg (CDS) was founded in 1972 (almost 40 years ago) by INAG (now INSU), the French National Agency in charge of ground-based astronomy, with the role of collecting, maintaining, improving, and promoting the usage of astronomical data in electronic form among the scientific community. The CDS has been on the Internet since 1991, which changed radically the way our data are collected and distributed.

Since its creation, the CDS has maintained an active collaboration with participating observatories and data centers, such as ADS (the SAO/NASA Astrophysics Data System), or NED (the NASA/IPAC Extragalactic Database). And, since 1993 the CDS has been involved in the publication process of electronic tables of the European journal *Astronomy & Astrophysics* (Ochsenbein & Lequeux 1995; Ochsenbein et al. 2003), a collaboration which is now essential for the development and maintenance of the CDS services.

2. The CDS Services

The CDS[1] is best known today for its online services (Genova et al. 2000). The four main services are *SIMBAD*, *VizieR*, *Aladin* and the *Dictionary of Nomenclature of Celestial Objects*.

- *SIMBAD*[2] is the reference database for the identification of and bibliography for astronomical objects outside the Solar System. It provides cross-identifications, basic data, some selected measurements, and a bibliography from 70 different journals for almost 5 million astronomical objects (Wenger et al. 2000).
- *VizieR*[3] is a collection of more than 8,000 catalogs, published tables, and observation logs. It represents more than 7.5 billion catalogued sources with homogenized descriptions and is accessible via standardized interfaces (Ochsenbein et al. 2000).
- *Aladin*[4] is an interactive sky atlas, allowing the user to visualize digitized astronomical images and superimpose entries from astronomical catalogues and databases. Aladin acts as a portal to the Virtual Observatory (Bonnarel et al. 2000).
- The *Dictionary of Nomenclature of Celestial Objects*[5] helps astronomers to find their way through the jungle of names used in the literature to designate astronomical objects (Lortet et al. 1994).

The databases behind these services are populated and maintained by the CDS staff. In the old days, this work was a patient exercise in typing and verification: the reference of each article was entered into SIMBAD and the objects studied in the article added or updated in SIMBAD. The huge increase in the number of publications, as well as the amount of data processed and published, led to the development of our current specialisation as "information scientists," some kind of hybrid between a librarian and an expert in astronomical data, having to manipulate the data related to the electronic publication as well as scientific data contained in the CDS databases.

3. The Multiple Roles of Information Scientists

The articles published in the astronomical literature represent the thread of our activity. In order to illustrate this activity, we present in Fig. 1 the path of the data we are dealing with.

The starting point, at the top of the figure, is the author who publishes a paper, which is submitted to an editor, and generates a publication if accepted. At the CDS,

---

[1] http://cdsweb.u-strasbg.fr/

[2] http://simbad.u-strasbg.fr/simbad/

[3] http://vizier.u-strasbg.fr/viz-bin/VizieR

[4] http://aladin.u-strasbg.fr/aladin.gml

[5] http://cds.u-strasbg.fr/cgi-bin/Dic-Simbad

this publication and the astronomical objects studied therein will be entered into SIMBAD (left-hand side of the figure); and the data published in the paper, generally in the form of tables, will be entered as catalogs into VizieR (right-hand side of the figure). The three boxes on the CDS' side, representing the services described in Section 2, are:

- *Catalogs*: the archive of electronically published tables,
- *VizieR*: the service to access these tables,
- *SIMBAD*: connects the published results with what is already known.

The information scientists are represented in this figure by the stars — they are in every step that links the publications to the databases.

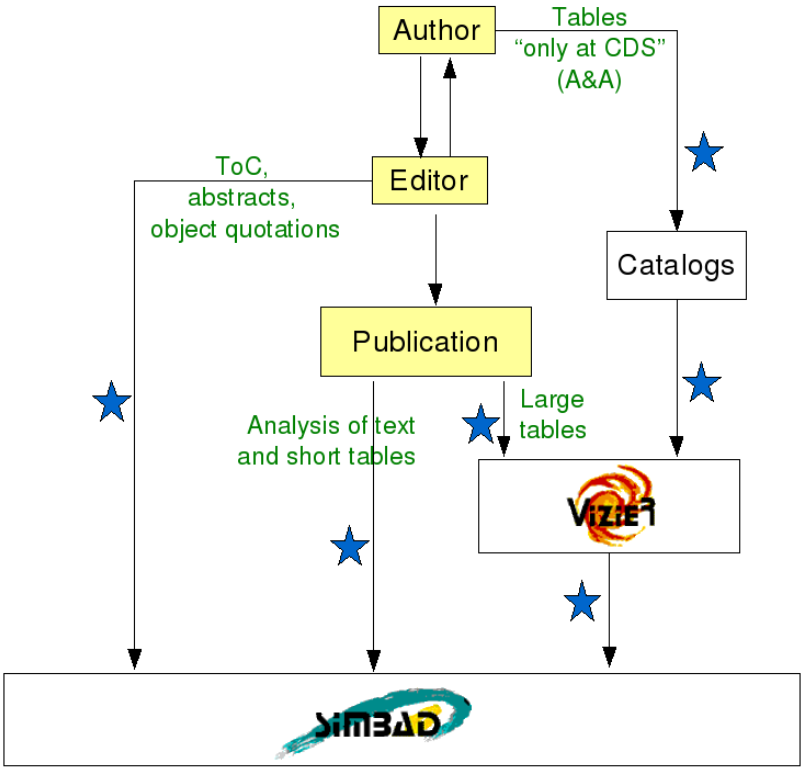

Figure 1. The main actors in the life cycle of an article published in an astronomical journal. The information scientists are represented by the stars.

3.1. Electronic Tables

As part of the publication process, the *Astronomy & Astrophysics* scientific editor selects from a submitted paper the data that are considered useful if distributed in an electronic form. Such data can be tables — large tables are much more useful on a computer than on a printed page — but are not limited to tables. For instance, spectra, images, data cubes, and so on, can be reused in new studies. Therefore, at the editor's

request, the authors send their data along with a description file directly to the CDS (upper right part of Fig. 1).

On the CDS' side, the descriptions of the data are homogenized and standardized; data are checked for consistency with their descriptions. This step may involve interactions with the authors, for example if the description of a parameter is unclear, or if a value seems to be out of bounds.

Once this step is achieved, the tables are then ready and become public upon publication of the article. The tables become also accessible via VizieR, the gateway which allows one to query the data from any of the stored catalogs and tables.

3.2. Bibliography in SIMBAD

The list of new papers is transmitted to the CDS in the form of tables of contents for addition into SIMBAD (upper left part of Fig. 1): the publisher sends to the CDS files that contain the tables of contents, abstracts, and keywords of a journal issue, along with, in *Astronomy & Astrophysics'* case, the object's quotations, i.e. the list of objects studied in a paper and tagged by the authors.

The CDS then checks the tables of contents; the quoted object names, in particular, must be verified because we provide links from the object names as quoted in the paper to SIMBAD.

3.3. Data in SIMBAD

Beyond its bibliographical aspects, SIMBAD is also a database which gathers many parameters describing the astronomical objects stored, like distances, motions, velocities, photometric measurements, etc. There are two paths to enter data into SIMBAD: one is feeding directly to SIMBAD from the published paper, and the other one makes use of the facilities offered by VizieR (bottom part of Fig. 1).

3.3.1. Manual Additions and Updates

This way of inserting into SIMBAD the objects studied in the papers is the traditional way, where our main task is to retrieve the astronomical objects which are studied in the publication as well as their related data. This involves reading each paper, finding out which are the studied objects — stars, galaxies, nebulae, etc. — and adding into SIMBAD the missing data.

So, the first step is to detect and identify the objects, i.e. to find a name that will be understood by SIMBAD. The next step is then to check, for each object, whether it already exists in SIMBAD or not. If it exists, we then check that it is the right object, from a comparison of the details given in the paper about that object with what is stored in SIMBAD — this is what we call the *cross-identification* process.

Since 2008, and for selected journals, a dedicated tool, developed at the CDS, is used to assist the information scientists in the detection of object names in the papers. This tool is called *DJIN* and is described in this conference by Lesteven et al. (2010); *DJIN* however does not retrieve the data related to each new object.

3.3.2. Addition of Large Datasets

The second path to enter data into SIMBAD is more adapted to the addition of large tables; it is illustrated in the bottom right part of Fig. 1. This second way is subdivided into two steps: the large tables go first into VizieR, and then from VizieR into SIMBAD.

1. Insertion into VizieR

   This step is quite similar to the process of adding electronic tables as part of the *A&A* publication process, illustrated in the upper right part of Fig. 1 and described in Section 3.1: the tables of the electronic papers are homogenized and standardized, and a description file is created if necessary. The data are then checked for consistency before the tables become available via VizieR.

2. From VizieR to SIMBAD

   Once the tables are available in VizieR, they can be processed semi-automatically to be entered into SIMBAD.

   The procedure we follow consists of a selection of the relevant data from the VizieR tables that will be compared with what is already known in SIMBAD. Such data include identifiers of the objects, their position in the sky, and other parameters like velocities or redshifts, distances, magnitudes, etc. Once again, it is necessary to identify the objects by finding or giving them a name that will be understood by SIMBAD.

   The data can then be compared automatically with what is already stored in SIMBAD. The result of this process is a list of matches and mismatches which is then analyzed by an information scientist who will make the final decision regarding the cross-identification: typically we have to decide whether this is a new object for SIMBAD, or if it is another name of an already existing object. In the case of a mismatch, we decide whether there is an error in the paper (this may happen) or an error in SIMBAD (this may happen too). Finally, we update SIMBAD accordingly.

   This procedure is called "semi"-automatic: it can deal with several hundred objects or more at once, but it still needs a human eye and expertise to decide on problematic cases. More details about this procedure can be found in Woelfel et al. (2007).

4. The Dictionary of Nomenclature

As we have seen in the preceding paragraphs, the designations or *identifications* of astronomical objects is the key aspect of our work.

The designations of astronomical objects used in the literature are often confusing. One object can have more than one name, and the existence of over 20 names for most studied objects is common (the Crab Nebula has over 70 names); and conversely, one name can refer to several objects, depending on the context. As can be imagined, this fundamental ambiguity is a source of much confusion.

Most of the time, astronomical objects come in lists or catalogs. Each catalog is given a short name or acronym (for instance M for Messier or NGC for New General Catalog), and each object in a catalog a number. One object can appear in several catalogs: this is why, for instance, the Andromeda Galaxy is also known as M 31 or NGC 224.

As an aid to help us to find our way in this object designation jungle, the CDS maintains and distributes online the *Dictionary of Nomenclature of Celestial Objects*. This Dictionary collects the designations from the literature, and if no acronym is suggested by the authors, the Dictionary creates one to ensure an unambiguous identifica-

tion. Currently the Dictionary provides keys to over 19,000 acronyms, and is growing at a rate of about 1000 acronyms per year.

To reduce the ambiguities of the designations of the astronomical objects, the authors of new catalogs and lists of objects are encouraged to submit their new acronyms to the IAU Commission V, via a link available on the Dictionary webpage.

The Dictionary is therefore a tool we use daily to solve the ambiguities existing in the literature; it is also an essential input to DJIN, the software used to detect object names in papers (see Section 3.3.1).

5. Conclusion

Now that we have provided an overview of our many activities, we will attempt to characterize the skills needed by the "information scientist." Obviously a certain knowledge of astronomy is essential. We need it:

- to detect and identify astronomical objects in the papers
- to be familiar with the astronomical zoo (stars, binaries, cepheids, clusters of stars, planetary nebulae, interacting galaxies, voids, ...)
- to recognize the parameters relevant for SIMBAD (redshifts, magnitudes, proper motions, spectral types, metallicity, ...)
- to follow the evolution of the discipline: new topics can lead to new object types, as for example when the first extrasolar planets were discovered some 15 years ago.

But we also need information technology skills: we have to be comfortable in a Linux environment, and in order to process large sets of data, we also need to understand and write scripts in languages like AWK, Perl or use other Linux tools.

Sharing our expertise among our team is essential. We also maintain close interactions with the astronomers working at CDS, and regularly discuss complex cases with them. We are therefore a team of people with complementary skills: information scientists, astronomers and software engineers, we all work together towards the same goals.

Our collaboration with journal editors is quite effective, especially for the linking between data and bibliography, and contributes to the value we add to our services which are widely used (currently around 300,000 queries per day).

References

Bonnarel, F., et al. 2000, A&AS, 143, 33
Genova, F., et al. 2000, A&AS, 143, 1
Lesteven, S., et al. 2010, in ASP Conf. Ser., Vol. 433, ed. E. Isaksson, J. Lagerstrom, A. Holl & N. Bawdekar (San Francisco: ASP), 317, (these proceedings)
Lortet, M.-C., Borde, S. & Ochsenbein, F. 1994, A&AS, 107, 193
Ochsenbein, F. & Lequeux, J., 1995, Vistas in Astron., 39, 227
Ochsenbein, F., et al. 2000, A&AS, 143, 23
Ochsenbein, F., et al. 2003, in Proc. Library and Information Services in Astronomy IV, ed. B. Corbin, E. Bryson & M. Wolf (Washington, DC: U.S. Naval Observatory), 257
Wenger, M., et al. 2000, A&AS, 143, 9
Woelfel, F., et al. 2007, ASP Conf. Ser., Vol. 377, Library & Information Services in Astronomy V, ed. S. Ricketts, C. Birdie & E. Isaksson (San Francisco: ASP), 43