The Updating of the Bibliography in the SIMBAD Database

Suzanne Laloë

CNRS, Institut d'Astrophysique de Paris, Paris, France

Abstract

The Institut d'Astrophysique's Library started inputting the bibliography in SIMBAD in 1983. Although the principle is still the same now as 10 years ago – to give users a list of bibliographical references, as comprehensive as possible, for all astronomical objects outside the solar system – we will see that things have changed. The increasing complexity of the objects' nomenclature has led to the development of new tools: the info or dictionary, the creation of acronyms for new designations. The automation of some tasks (inputting the texts of references, scanning long lists of objects) is slowly but surely progressing. But the tremendous increase in the number of pages published each year will make it more and more difficult to remain up to date, mainly by shortage in manpower. [1]

[1] The present "Paris" team includes, in addition to the author of this paper: Angèle Beyneix, Institut d'Astrophysique de Paris, bibliographer. Specializes in cluster stars with many cross-identifications and in inputting objects designated by new acronyms.
Suzanne Borde, Observatoire de Paris, in charge of the "info" and of three journals (Astronomical Journal, Astronomy and Astrophysics Supplement Series, Astrophysical Journal Supplement Series).
Claude Chagnard-Carpuat, Institut d'Astrophysique de Paris, principal bibliographer since 1983. Will retire in July 1995, leaving an invaluable contribution to SIMBAD.
Gratiane Chassagnard, Institut d'Astrophysique de Paris, bibliographer since January 1994. Will be the successor of Claude Chagnard-Carpuat.
Marie-Rose Dulou, Observatoire de Bordeaux, and Nicole Ralite, Observatoire de Bordeaux, specialize in long lists with addition of data.
Most of these people work only part-time for SIMBAD.
The "Strasbourg" team includes astronomers, computer specialists and one bibliographer, Marie-José Wagner.

Preprint submitted to Elsevier Science *5 July 1995*

1 Classical Procedures

1.1 Selecting the Articles

Over one hundred periodicals are scanned, i.e. all those received at the IAP library. Although it is a fairly good library, some journals are missing. In particular, because of lack of funding, new journals published in the Former Soviet Union are presently not taken into account. A number of additional publications, including colloquia and symposia are scanned or have been scanned but not in a systematic way. We do scan some newsletters (Messenger, Gemini, Journal des Astronomes Français, ...) but others are ignored. We used to systematically scan the original soviet publications (not waiting for the cover-to-cover translations) but, in 1995, this might have to be changed. For example, Astrofizika is very difficult to get, whereas the translation Astrophysics arrives regularly, Astronomy Letters arrive in the library at about the same time as Pisma,... Should we change our policy in the future? LISA seems a good place to discuss this question.

Each issue is scanned by one of the bibliographers to detect all papers containing at least one object, even if it is just mentioned in a figure. Then, the bibliographical reference and the quoted objects are recorded in SIMBAD. All this work is done on-line and made immediately available to the whole community. Most papers are treated a few days after the journal has been received in the library. The input of "lists" (papers mentioning more than about 50 objects) takes more time.

1.2 Recording a Reference

The first operation is to create a bibliographical code for SIMBAD (as shown in the poster by M. Schmitz et al. at this conference).

For each recorded paper, in addition to the information given in the bibliographical code (year, abbreviated name of publication, volume, first page of the paper, initial of the first author), we also put in SIMBAD: the last page, the title and the complete list of authors (except for BAAS and other abstracts journals for which we only take the first author if there are more than three).

The recording of the reference is the least scientific task in SIMBAD but requires specific expertise and is certainly very time-consuming. In 1994, the number of characters typed for authors' names and titles was the equivalent of a book over 270 pages! We try to obey rules for difficult authors' names (Chinese names, mention of Jr, III, ...) and for abbreviations in the titles, but the result is certainly not perfectly homogeneous (see section 2.3 for possible improvements).

By a special agreement (see F. Ochsenbein and J. Lequeux, this volume), the bibliographical references and abstracts of papers accepted in Astronomy and Astrophysics and in Astronomy and Astrophysics Supplement Series are put on line by the CDS in advance of publication. The CDS developed the necessary software to introduce the texts of the references in SIMBAD, so for these two journals, we do not have to input them manually.

We are at the moment trying to incorporate the references from Astrophysical Journal Letters without retyping them, but this has not been done so far. This step is certainly the one which should change most in the near future and I do hope that more standard formats (like SGML) will make possible the transfer of bibliographical data from the original author's text directly to the database.

1.3 Link with the Objects

Once the bibliographical reference exists in SIMBAD, it is attached to each object mentioned in the paper. Therefore it is necessary to identify correctly the objects cited in the paper. **This is the most delicate phase of the operation** (and we deal each year with about 100,000 objects!). Different situations could occur:

(i) *The identifier is recognized by SIMBAD*

This is the easiest situation. However, it is important to remain vigilant because some confusion could still exist, e.g. the object M 31 could be the star 31 of a list of Merrill (1971, Obs. Astron. Univ. Nac. La Plata, 38, 1) and not the Andromeda Galaxy Messier 31. Figure 1 shows an example of the variety of cases that we can find in the same paper.

(ii) *The identifier is not recognized by SIMBAD*

Up to about 1990, most of these objects where ignored. The situation has now changed and will be explained in section 2.2. Some of these objects used to be input with a name built on their equatorial coordinates (e.g. EQ 1234-332) or under provisional designations following the "Dictionary of nomenclature" by M.C. Lortet et al. These provisional designations were preceded by the letter "p" and, although you can still find some in SIMBAD, they are progressively changed to definitive designations.

Some objects, not coming from a list or catalogue, but designated by a kind of "nickname", were already and are still designated by NAME. Ex. NAME ROSETTE NEBULA, NAME STEPHAN'S QUINTET.

In some cases, the objects remain unidentified and are not recorded in the database (see section 2.2). A comment is then added to the reference.

1.4 Addition of Data

In addition to the bibliographical reference, other fundamental data found in the paper may be added to the object: position, magnitudes, spectral type or morphological type. These data are added to SIMBAD only if there is no previous information of this kind in the database. Systematic corrections to the fundamental data are made at the CDS. The problem for both the bibliographers and the users is that there is no indication of where the fundamental data were found and how reliable they are. So, the addition of a reference and a reliability evaluator to the fundamental data is planned.

Adding the coordinates by hand is a long, tedious and "risky" task (mistypings are difficult to detect), but it is still done for a large number of papers. The use of X-windows terminals combined with the availability of some tables in electronic form is beginning to allow the bibliographers to "click" the coordinates with the mouse rather than copying them from the paper. Adding the magnitudes V and B usually requires that the bibliographers make a preliminary small calculation, V and B–V being usually given in the literature, rather than V and B.

Adding the object type is the least boring but one of the most difficult addition to the data. For most acronyms, this is done automatically (ex. any BBW object will be recorded as a HII region), but for general surveys or for the NAME designations, the object type has to been added by hand. The list of object types has appeared in SIMBAD in 1993 (superseding the old limited choice: *, G, ?) and is regularly updated.

2 Recent Developments

2.1 New Acronyms and the Info Tool

The main improvement of SIMBAD in the last years has been the systematic introduction of new designations.

Every week, at the IAP bibliographers meeting, we collect an average of 7 new designations. These designations are sometimes clearly given by the authors but most of the time we have to build a new acronym. Presently, we follow the rules used by the NASA Extragalactic Database (NED): an acronym is composed with the initials of the three first authors followed by the last two digits of the year, all between brackets, e.g. [WIE93] for 1993MNRAS.261.185W by Warren S.J., Irwin M.J., Ewans D.W. et al., but [M94a], [M94b], [M94c] if there are several papers with only one author beginning with M.

New acronyms are immediately registered in Paris in the "info" and, then, made available in SIMBAD at the CDS within about a month. In 1994, several hundred acronyms which were on a "waiting list" were made available. This led to going back to about 500 articles and to input the objects which had previously been left aside because they had no identification recognized by SIMBAD.

The use of "info" (or "info –l" for detailed information) is an absolute must to find out about all the possible designations and their format, some designations being quite complicated (see example Fig. 2).

2.2 Comments

Another new improvement of SIMBAD, still not completely implemented, is the existence of detailed comments attached to the text of a reference. I shall here restrict myself to the comments used in Paris by the bibliographers, the main other comments concerning the availability at the CDS of tables and abstracts (comments "files" and "flags").

These comments are not added to the text of a reference at the first input described in section 1.2, they come from additional information becoming available later on.

(i) *Existence of an erratum (or addendum or corrigendum)*
When the bibliographer sees such a mention, she just types the code of the original reference and add the comment +erratum vol. NNN, p. NNNN (it is always in the same journal). There is no new reference code for a "paper" which is only a correction.

(ii) *Mention of misprints, misidentifications, truncated designations*
The SIMBAD bibliographers spend several hours, each week, trying to identify objects misquoted by the authors. The simplest case is an inversion of digits, such as HD 21326 instead of HD 23126, the most difficult are truncated designations (which should not be used by the authors and not be accepted by the editors of the journals). To trace down these misprints or misidentifications, we go back to the papers quoted in the bibliography or to the catalogues. So, once more, we urge authors, as recommended by the IAU, to always give a second designation and/or positional information next to their main designation.

(iii) *Mention of objects not recorded in SIMBAD*

If all our searches remain unsuccessful, or if we have doubts about an identifier, we start exchanging messages with the CDS astronomers. If the specialist at CDS can identify the object, we input it in SIMBAD and we add to the reference a comment such as "RX J0558+53 = RX J05580+5353". If the object remains unidentified, we add a comment "object XXX not in SIMBAD". There is still a hope that the author will one day see the comment and give a clue to the mystery!

(iv) *Mention of new acronyms*

This comment, the "dictionary" comment, corresponds to what you find in the info, i.e. the explanations about the acronym taken from that reference.

Ex. Viewing the reference 1992A&A.266.37B, you immediately see that this is the 'basic' or 'reference list' for the acronym [BV92], what the two possible formats are, and the number of objects in the list.

(v) *Internal or work comments*

These comments, which appeared in 1994, are not visible by the users but are very important for the updating team. They concern the status of the papers which cannot be considered as finished and which will be completed sooner or later. This could be:

(a) Waiting for a new acronym to be made available
(b) Too long to be done in Paris, will be sent to Bordeaux Observatory with the necessary explanations
(c) Some tables should be obtained from the authors in electronic form, or be scanned from the paper version
(d) The tables are now available in electronic form, but the objects are not yet incorporated in SIMBAD,
(e) Etc.

These comments allow the SIMBAD team to know what still has to be done for these papers (about 2% of the total number of references). References with no comment attached are supposed to be completely treated, which, of course, is certainly not completely true.

These detailed comments, the use of which is still under discussion within the SIMBAD team, are part of a more general improvement of the system, which I shall now briefly mention.

2.3 Quality Control

Among the 84,000 references and the 1,100,000 objects in SIMBAD, it is impossible not to have errors. Some occasional corrections are made when errors are detected by chance or when a user kindly points them out. But a need for systematic control and corrections appeared with the tremendous development of SIMBAD in the nineties. New software tools were developed at many levels and more astronomers became involved in the "object quality". Software controls can protect against typing errors and are made immediately at the input:

(i) *control of the reference code*

- The year should be less than or equal to the present year!
- The volume should correspond to the year of publication. To allow this control, we have to edit a table of correspondence for each journal, basically using the kardex records of the library. For some journals, however, this is not possible: no one can know in advance which will be the number of the last IAU circulars for 1995 and the first one in 1996!
- The last letter of the code should correspond to the initial of the first author.

(ii) *control in the format of data*

- The format corresponding to a given acronym should be respected E.g. NGC 34567 will be rejected (format of the NGC catalogue: NNNN)
- A missing sign for the declination when inputting the coordinates will lead to a rejection by the system, etc. Other controls allow a posteriori corrections:

a) *daily listing of the references created the preceding day*

b) *daily listing of the references deleted the preceding day* (a much shorter listing, often empty!)

(iii) *daily listing of new authors* (implemented in April 1995, will be used for checking the spelling of authors)

(iv) *weekly list of comments.* This allows to follow the status of the articles and to take action if the list is growing instead of decreasing (as seen in section 2.2., many comments should disappear when the problem is solved)

(v) *weekly list of new and deleted objects with a designation "NAME" and "p"*. This allows to prevent the creation of uncorrect designations and to encourage the transformation of previous bad designations into designations unambiguously recorded in the "info".

(vi) *general checking*

From time to time, about once a year, all the bibliographical references in SIMBAD are checked against the literature to detect misspelling for the authors, typing errors in the titles or pages.

The collaboration between the bibliographers and the CDS astronomers have also led to "manual" improvements, such as the numerous "merges" of objects previously identified under two or more different designations.

It is a very exciting but difficult task to input the bibliography in SIMBAD, mainly because of the ever increasing number of papers and of large surveys. Moreover, we now have to deal with electronic publications and tables on CD-ROM, which are taken into account as well as the paper information. But, not being sure that it will be possible in the future to keep up to date with all publications and with all objects, I would be happy to have, if possible, the reactions of our users. What is, in their opinion, the priority? What will they expect to be improved or, on the other hand, what could be ignored? Maybe a kind of users group will one day be set up and help the SIMBAD team take decisions. Anyway, I would like to remind our users (be they librarians, scientists or information specialists) that all reactions, precisions, criticisms (and compliments!) are always welcome. Let us all work together to improve the tools available to the community.

I would like to thank Suzanne Borde, Françoise Genova and Marc Wenger for their careful reading of this manuscript.