EPJ Web of Conferences **186**, 02004 (2018) https://doi.org/10.1051/epjconf/201818602004
LISA VIII

COSIM: The necessary evolution of a cross-identification tool along with data evolution

Catherine Brunet[1,⋆], *Mark* Allen[1], *Marianne* Brouty[1], *Mihaela* Buga[1], *Cécile* Loup[1], *Anaïs* Oberto[1], *Emmanuelle* Perret[1], *Bernd* Vollmer[1], and *Fabienne* Woelfel[1]

[1] *Université de Strasbourg, CNRS, Observatoire astronomique de Strasbourg, UMR 7550, F-67000 Strasbourg, France*

Abstract. SIMBAD is a bibliographic added-value database on astronomical objects, where the data on individual objects are cross-identified as far as possible. The data comes exclusively from what has been published by the scientific community. To treat large tables, the work is done semi-automatically with the help of a customized software. Since 2014, we are using a new one, called COSIM (Comparison of Objects for SIMBAD). It meets the new requirements which is a consequence of the evolution of the available astronomical data. It has increased in number, accuracy and diversity. On the basis of the data presented in a published table, COSIM searches for objects that are already known in SIMBAD, by name or by coordinates. A combination of scores based on the available and comparable parameters, like the main object type, coordinates, velocity and magnitudes, suggests whether the candidate is good for cross-identification or not. As soon as the result of the search is clear, indicating that there is either no matching candidate or only one good candidate, COSIM creates the commands necessary for updating the SIMBAD database. The documentalists can act on the method of calculation of each score, according to the nature of the objects in the table. Thus, with COSIM the documentalists manage to obtain a good cross-identification level with a minimum risk of omitted or false cross-identifications in a relatively short time compared to the treated data number.

1 The SIMBAD database

The SIMBAD database [1] is a service offered by the Strasbourg astronomical Data Center (CDS) [2]. It was created in 1979 and has been maintained since. It is a bibliographic added-value database on astronomical objects outside the solar system, searchable via the Internet.

SIMBAD provides basic data, identifiers, bibliographic references and observational measurements on astronomical objects. All data comes from the scientific literature [6]. SIMBAD now contains about 9 million objects and 24 million identifiers for 330,000 references. This research tool is complementary to VizieR, another database maintained at the CDS, containing long lists of objects and large surveys.

⋆ e-mail: catherine.brunet@astro.unistra.f ORCID: 0000-0001-9392-8817

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

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

SIMBAD is constantly maintained and updated by a multi-skilled team, including astronomers, documentalists (also called information scientists) and computer scientists [1, 2].

2 The documentalists' work

One of the characteristics of SIMBAD is that the information on an astronomical object is gathered together. That means there is a difficulty: the cross-identification. As the authors do not always designate their objects with a standardized name, identifying them is a challenge that the documentalists who update SIMBAD constantly face. For each object, we have to answer the question: which object does this name designate? Is it already in SIMBAD or is it a new one?

For each object name found in the article (in the text, in the abstract, in a figure, in a table...), SIMBAD is searched to see if it is already known. Two searching modes are used:

- first by identifier, if necessary with the help of the Dictionary of Nomenclature [3] made at the CDS [1];
- if the identifier is not found, a search is made around the coordinates (when they are given).

At the beginning of SIMBAD, the number of volumes was reasonable and the documentalists could afford to read the whole article, looking for astronomical objects in the text, the figures and the tables. We thus updated SIMBAD manually till the 90's [3, 4].

Then a software was developed and from then on, long lists of objects have been treated in a semi-automatic manner. From that moment, the documentalist team has been specialized into two branches: one for identifying objects in text, and one for dealing with the long lists [2].

Each of the two team branches has adapted itself to the evolution of the available astronomical data that have increased in number, accuracy and diversity:

- For the processing of texts, a software was created in 2008, *DJIN* [1, 5], in order to assist the detection of object names. The branch became the *DJIN team*.
- For the long lists, a new version of the first software had become necessary. It has been deployed since 2014 and has been called *COSIM* (*Comparaison d'Objets pour SIMBAD*). The branch became the *COSIM team*.

3 The operational COSIM software

The aim of COSIM is to help the documentalists to enter long lists of astronomical objects into SIMBAD. It is a software written in Java language. Starting from a dedicated format table, it will search SIMBAD for good cross-identifications, through a direct link to the database. At the end of the process it will write update commands.

3.1 The place of COSIM in the whole procedure

It is first important to emphasize that before and in parallel with the automatic work, there is a lot of human work, too. That's why we generally speak about semi-automatic work. We present in figure 1 the complete COSIM team procedure. First, in a weekly meeting between astronomers and documentalists, a list of priorities is discussed and it is decided which data are likely to be put into SIMBAD. As a result, the documentalist gets a small notice (step 1), that allows her to extract the useful data for COSIM. It is in this phase generally that the Dictionary of Nomenclature is consulted,

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

to re-write identifiers if necessary. A file is given to COSIM (step 2). COSIM consults SIMBAD to search for cross-identifications and returns an output file (step 3). This file is analyzed by the documentalist. This is an important part, which is sometimes long and difficult. Not only does the documentalist use whatever tool or database she has at her disposal, such as SIMBAD, VizieR, Aladin and others, but she also consults her colleagues (other documentalists, astronomers, sometimes the computer scientist) or authors. Sometimes some preliminary manual updates are necessary. The documentalist adjusts the parameters (step 4), until COSIM writes update commands for all objects of the list. The update in SIMBAD can be then automatically done (step 5).

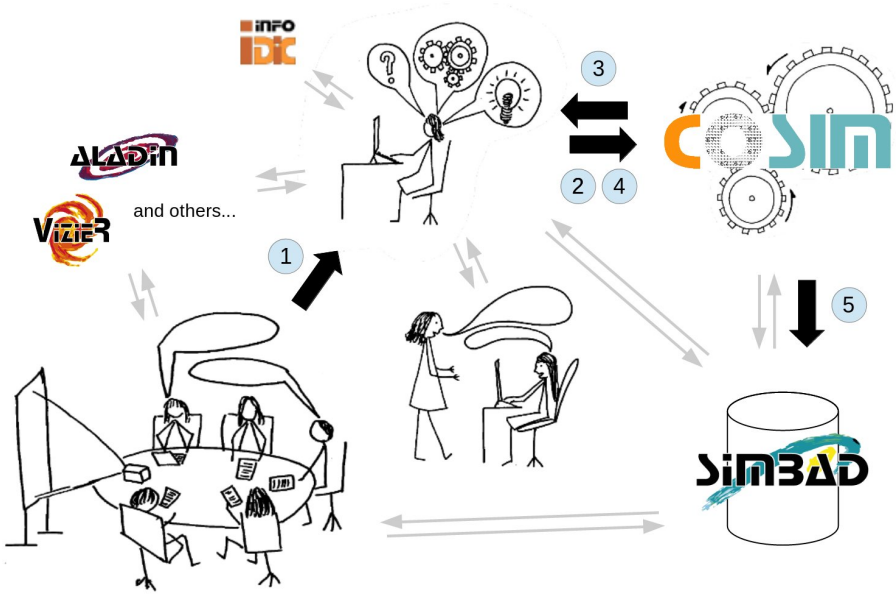

Figure 1. The COSIM team procedure. *Drawings are by Anna Preite Martinez.*

3.2 The COSIM software overview

COSIM searches SIMBAD for cross-identifications by identifier and/or around the coordinates. Then, each candidate found is compared to the entering object. To do this, a score is attributed to each comparable parameter, such as the object type, the coordinates, the velocity, the magnitude. Then all those scores are combined in order to evaluate the candidate (is it good for cross-identification or not?). According to the number of good candidates found, a decision is taken: to create a new object, or to update an existing one. Or, no automatic decision can be taken and the case is to be analyzed by the documentalist.

3.3 The object search in SIMBAD

COSIM searches SIMBAD for the object first by identifier and if not found, it searches around the coordinates. The search radius is automatically adapted to the accuracy of the coordinates in the table. But it can be modified. For example, for a table of galaxy clusters, one will search a larger area.

On the contrary, for deep field observations, one will search a smaller area. One can also decide to force the search around the coordinates, even if the object has been found by an identifier. In certain circumstances this can be useful, for example in order to look for mergers in SIMBAD.

As a result, one has a list of SIMBAD objects, either found by identifier, or that lie in the search area, or both.

3.4 The comparison

3.4.1 The scores attribution and calculation

Each candidate is attributed some scores: one for the object type (OT), one for the coordinates (COO), one for the radial velocity or redshift if any (V), and one for the magnitudes if any (M). In some very particular cases, one or two more scores are introduced: one which tests the presence or absence of an acronym among identifiers (ACRO), and one which tests the presence or absence of a reference among the bibliography (B).

The OT score is basically the result of a compatibility table between object types. For example, if the candidate is a star cluster whereas the entering object is a star, the score returned is -1; if the candidate found is a young stellar object candidate whereas the entering object is a young stellar object, the score returned is +1. One has the possibility to modify the results of the compatibility table, at will. For example, one can decide to consider a star and a galaxy as compatible, because the objects are point-like objects in a deep field and no human check has been done.

The COO, V and M scores are all three the result of a formula, that takes into account the uncertainty of both sides, the σ, Δ being the difference between entry and SIMBAD.

$$1.5 - 0.5\frac{\Delta}{\sqrt{\sigma^2_{\text{SIMBAD}} + \sigma^2_{\text{TABLE}}}}$$

The σ is generally given by the authors; if not, we estimate it, according to the instrument used and the wavelength, but also according to the object type. Indeed, for extended objects, we will artificially increase the coordinates sigma.

For each score there are two limit values (see figure 2): a *minimum*, under which the score is *low*, a priori for bad candidates, and a *maximum*, over which the score is *high*, a priori for good candidates. Between both, the score is said to be *medium*. The role of the documentalist is here to adapt those two values, in order to minimize the medium zone. Of course the best situation is when one can let coincide the minimum and the maximum, so that the score is either low or high.

The possibility of determining minima and maxima for each parameter separately is a new thing offered by COSIM. Thanks to this one can, for example, accept large differences in velocity but only for a very low distance.

3.4.2 The combination of scores

Once each comparable parameter has led to a qualified score, COSIM will combine the scores in order to give a status to the candidate.

All the calculated and not ignored scores are taken into account for the candidate evaluation. The OT and COO scores being considered as essential ones, as soon as one or the other is low, the candidate is labeled “bad” (see figure 3, c). It is not the case for the V and M scores: if one of the two is low whereas all the others are high, the candidate remains labeled “undefined” (see figure 3, d). In order to label a candidate “good”, all of its scores must be high if it has been found with a search

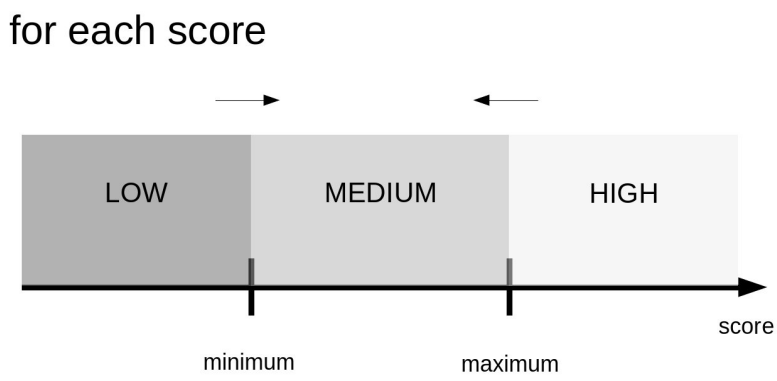

Figure 2. The limit values for each score

around the coordinates (see figure 3, a); or, all of its scores must be at least medium if it has been found by identifier (see figure 3, b). Actually, we are more confident when the candidate is found by identifier, because one can think that the author has already done the cross-identification work.

In order to reduce the number of cases where the software cannot automatically label the candidates either "good" or "bad", the documentalist has at her disposal some options. One can decide to ignore one score (because we estimate there are measure errors, or we estimate that the parameter is not a good comparison criteria for this sort of object). All candidates with this low score can be then nevertheless labeled "good" (see figure 4, a). One can on the contrary decide to attach a great importance to one score, as for the OT and COO scores. This time, all candidates with this low score will be labelled "bad" (see figure 4, b). In some particular cases, where the cross-identifications are very delicate, one can decide to attach even more importance to one score, by asking to cross-identify only if the corresponding parameter is present (see figure 4, c).

3.5 The update decision

According to the number of good and bad candidates found and when an automatic decision can be taken, the software writes update commands. Here is the exhaustive list where the automatic decision can be taken:

- No candidate has been found, neither by identifier nor around the coordinates. COSIM writes update commands to create a new object in SIMBAD.
- No candidate has been found by identifier, and all candidates found around the coordinates have been labeled "bad". COSIM again writes update commands to create a new object in SIMBAD.
- One or several candidates have been found, either by identifier or around the coordinates, but only one has been labeled "good". The others, found around the coordinates, have been labeled "undefined" or "bad", it does not matter. COSIM writes update commands to update the SIMBAD existing object.

At the end of the analysis process, all objects entering the table should lead to one of two decisions (new object created or existing object updated). But before reaching this ideal situation, a lot of various, problematic cases can be encountered.

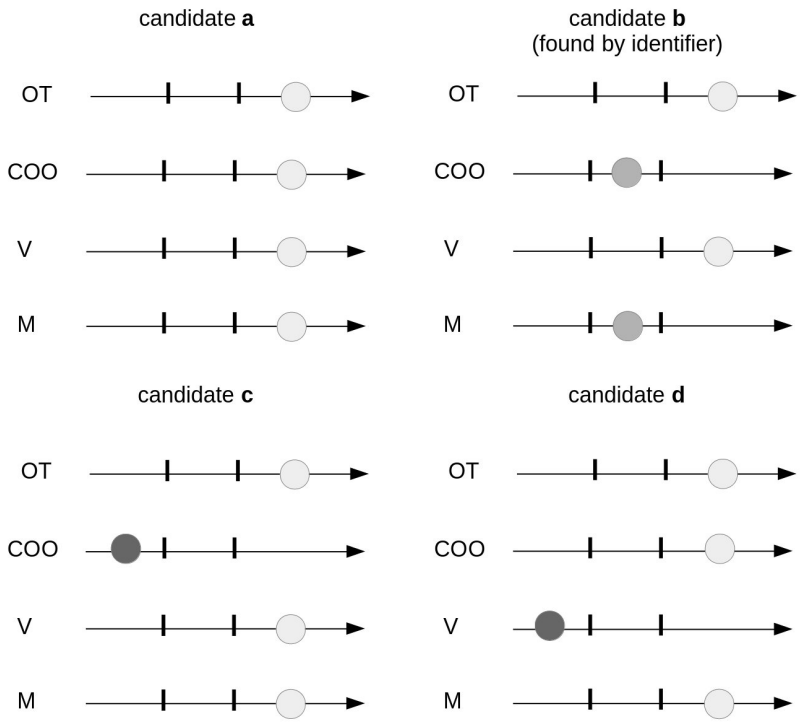

Figure 3. Examples of candidate evaluation by default: candidates which are labeled "good" (a, b); candidate which is labeled "bad" (c); candidate which is labeled "undefined" (d).

Basically, numerous combinations exist where the automatic decision is impossible. Nevertheless, thanks to the good comparison criteria used by COSIM and thanks to the physical character of the data, they do not happen so often and the situation is generally manageable. We have to analyze zero to ten percent of the entering objects manually, knowing that the size of the tables we treat is generally between several hundred to several thousand objects. If too many objects would need an individual analysis, we prefer to miss some cross-identifications and create more new objects.

Among the cases where no automatic decision can be taken, a few ones are notable:

- the “Possible Merger”: two (or more) candidates have been labeled “good”. The software cannot choose and has not been programmed to do the update on several objects. Actually, even if the “possible mergers” put a break on the automatic update decision, they are interesting, because they allow us to “tidy” SIMBAD. Thanks to the great possibility of COSIM to refine the comparison parameters, we can really rely on COSIM for detecting those cases, and only a small portion are wrong. SIMBAD has been improved by a lot of mergers since the use of COSIM.
- the “Conflict”: the author has given several identifiers for one object, and they correspond to different objects in SIMBAD. Here again, the “conflict” puts a break on the software process: COSIM can’t choose. The documentalist will analyze the situation, helped by the status given by COSIM (it can be at the same time a “possible merger”) and do modifications, either in the input file, or directly in SIMBAD.

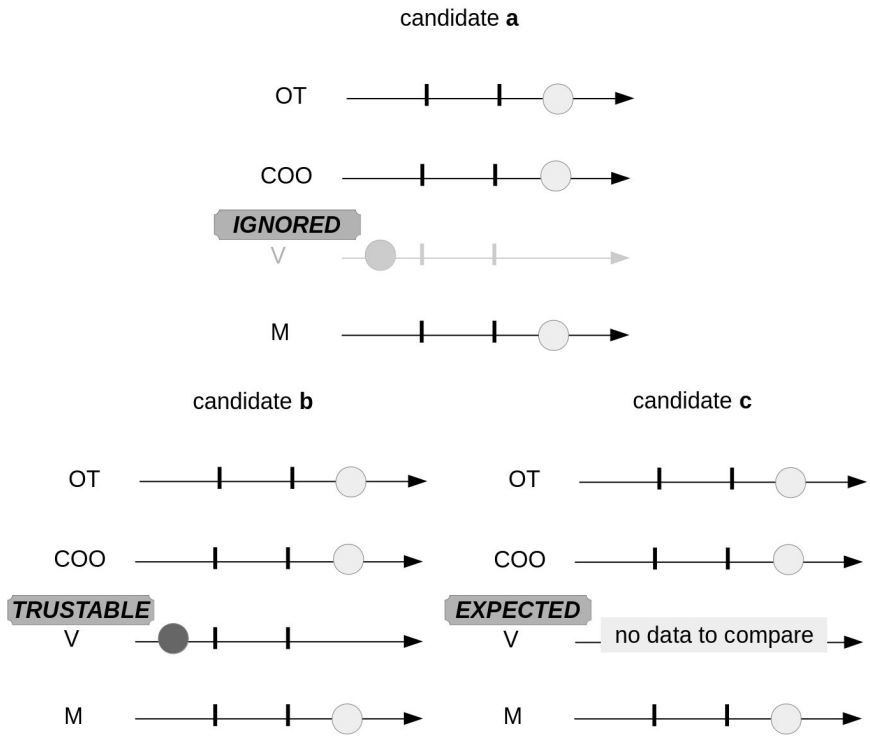

Figure 4. Examples of candidate evaluation with options: candidate which is labeled "good" (a); candidates which are labeled "bad" (b, c).

- the “Already Connected”: the same SIMBAD object is labeled “good” for a second entry in the same table. As we suppose that all entries of the table correspond to different objects without repetition (unless it is clearly written), COSIM is not authorized to update twice the same SIMBAD object.

In all of these three cases and others, the documentalist will analyze the situation and give the appropriate answer so that COSIM will be able to write update commands in the next step:

- either in adapting the criteria of comparison in COSIM, if not already done (modifying the search radius, the sigma in the formulas, the limit values of each score, or adding options);
- or in doing some individual modifications, either in the input file, or directly in SIMBAD.

The latest intervention being time-consuming, the first will be as far as possible preferred.

3.6 Statistics and sorted lists

With a list of several hundred or several thousand objects, it would be impossible and counterproductive to check whether the automatic decision is correct for each object. So to help us in the analysis of the output file, COSIM provides statistics and sorted lists. In one glimpse, we have the number of “Possible Mergers”, the number of “Conflicts”, the number of “Already Connected”, and the number of different final decisions: NEW, UPDATE on a candidate found by identifier, UPDATE on a

candidate found around the coordinates, UNDEFINED cases. And overall, we have information that really matters in the process: a list of the nearest candidates which have been rejected, and a list of the furthest candidates accepted for cross-identification, both lists being ordered by object type. This allows us to check if the level of cross-identification is correct.

4 Conclusion

The COSIM software is an essential part of the SIMBAD operations, with the main advantages being that:

- all the parameters in the comparison are adjustable
- statistics and sorted lists give a precious help
- it is faster than the precedent software

As a result we can rely on these automatic parts of the process, allowing us to use precious time on aspects that need human expertise. Also, by using COSIM we can treat more objects and have confidence in the quality of the cross-identifications. COSIM is thus an important update of CDS procedures that helps us to address the increasing volume of published data. It is also important to understand that while COSIM is a useful tool, the global process must be driven by real people with expertise in astronomy and documentation, which remains essential.

References

[1] M. Brouty, F. Woelfel, C. Bruneau et al., *Information Scientists: Between Editors and Data Centers* in *21st Century Astronomy Librarianship, From New Ideas to Action*, edited by Eva Isaksson, Jill Lagerstrom, András Holl, and Nirupama Bawdekar (2010), Vol. 433 of Astronomical Society of the Pacific Conference Series, p. 195—200

[2] M. Buga, C. Bot, M. Brouty et al., *How Documentalists Update SIMBAD* in *Open Science at the Frontiers of Librarianship*, edited by András Holl, Soizick Lesteven, Dianne Dietrich, and Antonella Gasperini (2015), Vol. 492 of Astronomical Society of the Pacific Conference Series, p. 47–50

[3] S. Laloë, A. Beyneix, S. Borde et al., Bulletin d'Information du CDS **43**, 57 (1993)

[4] S. Laloë, Vistas in Astron. **39**, 179 (1995)

[5] S. Lesteven, C. Bonnin, S. Derrière et al. 2010, *DJIN: Detection in Journals of Identifiers and Names* in *21st Century Astronomy Librarianship, From New Ideas to Action*, edited by Eva Isaksson, Jill Lagerstrom, András Holl, and Nirupama Bawdekar (2010), Vol. 433 of Astronomical Society of the Pacific Conference Series, p. 317–323

[6] M. Wenger, F. Ochsenbein, D. Egret et al., A&AS **143**, 9 (2000)