EPJ Web of Conferences **186**, 12004 (2018) https://doi.org/10.1051/epjconf/201818612004
LISA VIII

Bibliographical references: From publishers to SIMBAD

Thomas Delacour[1,⋆], *Soizick* Lesteven[1], *Gilles* Landais[1], *Aline* Eisele[1], *Magali* Neuville[1], *Evelyne* Son[1], and *Philippe* Vonflie[1]

[1] *Centre de données astronomique de Strasbourg (CDS)*

Abstract. The SIMBAD astronomical database hosted by the CDS provides basic data, cross-identifications, bibliography and measurements for astronomical objects outside the solar system.
The CDS receives the bibliographic meta-data of the articles published in the main astronomical journals directly from the publishers. How we receive the data and their format vary from one publisher to the next. These data are first extracted and stored in files with a standardised format. Then, to avoid errors or misprints, we perform different tests on these data:

- Author names are compared to a reference list maintained at CDS, and the keywords are compared with the AAS list
- Astronomical objects are verified by checking their name in the SIMBAD database
- A completion test checks that all of articles of a journal volume are present

The next step identifies whether an astronomical object appears inside a title, a keyword or an abstract, and if so, we add a link to the object in SIMBAD. Once all of the verifications and corrections have been made we add the meta-data into SIMBAD. We also add other information such as the number of different astronomical objects studied in the paper, the presence tables and their links to VizieR, any new acronyms, as well as some comments. New developments are in progress to automatically extract the data from the tables in the articles (that have not been processed by, or provided to VizieR) .
In addition, each night automatic checks are executed to list the new data and to test the coherence of these data in SIMBAD.

Introduction: The context

The objective of the Centre de Données astronomiques de Strasbourg [1] is to collect, homogenize, distribute and preserve astronomical information published in astronomical journals for the usage of the whole community. The articles published in the astronomical literature represent the thread of our activity. Before looking for data inside the paper, we have to create the bibliographical entries for the SIMBAD and VizieR databases.

By special agreements with the main journals, the CDS receives meta-data of articles in electronic form. How we receive the data and their format vary from one publisher to the next, and also vary over time. The first operation is the extraction of the required information, and the second step is to store

⋆ e-mail: thomas.delacour@astro.unistra.fr ORCID: 0000-0002-4941-7478

this data in a unique and specific format. Receiving all of these data allows us to be more exhaustive, more efficient and assure a better quality service. For a small number of journals documentalists still add bibliographic data manually.

1 The bibliographical data

In astronomy, each article is designed by the bibcode [2]. We associate this bibcode to the article DOI, the copyright, the numbers of pages, the complete list of authors with their affiliations, the title, the abstract and the keywords. All this information allows us to advertise and locate the papers.

1.1 Extraction/storage

We began our collaborations in 1994 with the A&A journal, which was quickly followed up with the AAS journals [3]. In 25 years, the data provided and their formats have evolved greatly. To take into account all these changes, we continually upgrade our software so that they are always adapted to the received data. Currently, we receive data mainly in XML format, sometimes in ASCII. The useful information is extracted from the files and is stored in parfiles.
The parfile format is used to store the data because it can be easily edited by a human and parsed by a program. A parfile consists on a series of blocks, each of them representing an article, separated by a new line. All the lines inside a block start with the symbol % followed by a letter which indicate what the line contains (Abstract, Title, Author...).

```
%R 2017A&A...600A..57C
%F A+A.ori/m2017-041/v600/aa29522-16.xml
%J-57
%M 15
%DOI 10.1051/0004-6361/201629522
%c {copyright} ESO, 2017
%A Chiaberge, M.
%A Ely, J.C.
%A Meyer, E.T.

...
%K galaxies: active
%K quasars: individual: 3C 186
%K galaxies: jets
%K gravitational waves
```

Figure 1. Parfile format

For each journal, volume per volume or issue per issue, all of the information is stored in parfiles.

1.2 Checks and SIMBAD update

To ensure the quality of the data and to avoid errors that can occur during the data extraction, we carry out a multitude of checks. For each issue or volume we process, we check the completeness of it to be sure that no articles are missing, and when we have an erratum we add this information to the related article. We also maintain a correlation list between the bibcodes and the DOIs. All author names are compared to the list of all authors already existing in our databases. If an author is not found, it must

be verified to check if it is a new one or a misprint. In the same way we check the keywords with the list a keywords from the AAS, and when a keyword does not match any in the list the program will display a message. When an individual object is given in a keyword we first validate the identifier and then add a link to SIMBAD. Once all the data have been checked and corrected we add them into SIMBAD. It is important to point out that no corrections are made to the published data. We only correct errors made during our transcriptions. When there is an error on an object, we add a macro latex that will allow us to link the misnamed object directly to the right object in SIMBAD. In the other cases, we add a note to the data in SIMBAD. The journal editors are contacted, if needed, to signal an error.

At that stage, the documentalists in charge of updating SIMBAD begin to link the articles to the SIMBAD astronomical objects [4].

2 Table extractions for VizieR

Published articles may sometimes contain small astronomical data tables which should be integrated to Vizier [5]. Until recently, documentalists used the pdf file of the article provided by journals and copy-pasted the table information directly into a text editor. But this is not a convenient way to do it because it can modify the formatting of the table, some characters are not recognized and it demands a lot of work to get the content as it is in the pdf. The tables can now be processed from xml files provided by the publishers. A program has been written to retrieve files, extract, process and convert the tables into ASCII files. Documentalists can then use those files to create the standard Readme description for Vizier.

To download xml files the program is given a bibcode, get the corresponding doi and use it to connect to a publisher site. Once files are stored, the program will search for the xml then start to parse it. Tables are written in mathml or latex format so it has to be convert to ASCII before documentalists can use them. The program parses two times. The first one is to determine the shape of each table. This avoids problems in cases where the number of rows and columns are not consistent. The second time it will extract data, and mathematical symbols and as such are transformed so they can be easily integrated. The program will try to keep the shape of a table and write rows and columns separated by a space.

Besides data in tables, information about the associated article (authors, title, abstract, references, table description...) are also extracted. Each table is then written separately in ASCII and csv files. Sometimes a table cannot be processed into ASCII correctly due to its format or data, in which case a csv file may be used instead to manipulate data easily. A Readme file containing information about the article is generated along with a summary of different tables and reference articles. After creation of all files documentalists can verify the information and data by comparing them to the original tables and article. Once verifications are done the data are put in Vizier and can be consulted on the website along with tables.

In further updates of SIMBAD, we add other information to the bibliographical data such as the number of different astronomical objects studied in the paper, the presence of one or more tables and the links to VizieR, the creation of a new acronym, and also some comments.

3 Daily and weekly tests

Information processing at CDS is a chain [6]. To monitor the progress of the work among all the team members and assure the quality of the data, we have developed automatic test processes and we send the results to the concerned documentalists. Several tests are launched every day and/or week automatically.

Concerning the data we receive from the publishers, every night a program compares the present bibliographical content with the one from the day before. It checks for new references entered in SIMBAD, the deleted references, the number of references, and new authors names. It checks also the objects names in the keywords and the object names tagged by the authors in some journals (A&A). An email is sent to the CDS documentalists who correct, if necessary, the bibliographical data every morning.

Furthermore, each reference has an associated status that indicates that work has been done on it. The status generally goes with a detailed comment that indicates what has already been done. At the end of the week the program counts the total of every status and makes a comparison with the file containing the status of the last week. It reports to documentalists the number of status added, removed and the modifications of comments.

When data are modified in SIMBAD, tests allow verification of links validity and their correction if needed.

4 Conclusion

Receiving all of these data from the publishers allows us to be more exhaustive and more efficient. Bibliographical data go through a lot of steps from the publishers to SIMBAD. These processes and verifications are essential to maintain the quality and coherence of the database. Data are checked automatically or manually until their integration in SIMBAD which makes mistakes rare, moreover permanent checks ensure the correctness over time. Improvements are in progress to homogenize the database with the new abstract, keywords and DOI added to SIMBAD. Developments for table extraction continues with the objective to process more journals and return better results.

References

[1] F. Genova, D. Egret, O. Bienaymé, et al, A&AS **143**, 1 (2000)

[2] M. Schmitz, G. Helou, C. Lague, et al, Vistas in Astron., **39**, 272 (1995)

[3] F. Ochsenbein, J. Lequeux, Vistas in Astron., **39**, 227 (1995)

[4] 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

[5] G. Landais, T. Boch, M. Brouty, et al., *Management of Catalogs at CDS* 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. 57–60

[6] E. Perret, T. Boch, F. Bonnarel et al., *Working Together at CDS: The Symbiosis Between Astronomers, Documentalists, and IT Specialists* 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. 13–21