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

DJIN: Detection in Journals of Identifiers and Names

S. Lesteven,[1] C. Bonnin,[1] S. Derrière,[1] P. Dubois,[1] F. Genova,[1] A. Oberto,[1] F. Ochsenbein,[1] S. Borde,[1] G. Chassagnard,[2] M. Brouty,[1] C. Bruneau,[1] C. Brunet,[1] H. Claude,[1] A. Eisele,[1] S. Laloë,[1] M. Neuville,[1] E. Perret,[1] P. Vannier,[1] P. Vonflie,[1] M.-J. Wagner,[1] and F. Woelfel[1]

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

Abstract. We dreamed of it, we developed it, and now we use it.

DJIN is a powerful tool that recognizes astronomical object names in full texts.

DJIN is very efficient and helpful for the SIMBAD team who have been dealing with an ever increasing number of astronomical articles.

DJIN detects most of the astronomical object names quoted in full-text articles, but the team still has to check and validate the names, to deal with new identifiers, to verify cross-identifications and to update SIMBAD with new astronomical data (position, magnitudes, etc.). That is, the team work is concentrated on value-added aspects, the best use of the team's expertise. This was an important consideration in the design of the software.

DJIN provides more than just the recognition of names; it says how many times an astronomical object is cited in a text (whatever its identifier is), where it is cited (title, abstracts, keyword, tables, figures, text, etc.) and keeps track of the relation between the identifiers and articles.

DJIN is fully integrated in the SIMBAD process, and interfaces the updating software used daily by the team. It is also a starting point for new features like linking SIMBAD and NED, and computing the relevance of each paper attached to one object.

DJIN has been fully tested by the whole team to check both the quality of detection and the tool's ergonomics. Team feedback has been critical for the success of this difficult and risky endeavor.

In this paper we describe this tool, and our experience after two years of usage; we discuss also the the significant changes in our daily work that DJIN has triggered.

1. Introduction: the Context

SIMBAD is the reference database for the identification and bibliography of astronomical objects. It contains identifications, "basic data", bibliography, and selected observational measurements for several million astronomical objects (Wenger 2006). It is used world-wide and receives an average of 20,000 requests daily. This is the result of more than 35 years of evolution. First created as the Catalog of Stellar Identifications (CSI) in 1971, it is now in its fourth version, following, during these 35 years, the evolution of hardware, software and computer languages and technical constraints. During these 35 years, these essential guiding principles have remained: maintaining the quality of the content and making the best usage of available hardware and software. The database content is built by an experienced team of information scientists and scientists. The last big improvement came with the development of a new tool to help the information scientists to quasi-automatically retrieve astronomical object names in full-text papers.

In the old days, the information scientists had to create and manually enter the references into SIMBAD. They would read papers and every time they would recognize an object name in the article (title, abstract, table, etc.), they would update the database. This meant that they had to research how each object name should be written in the database, and whether the object is already in the database or not. If not, they would create the new object.

During the last ten years the number of papers entered in SIMBAD has increased significantly and continuously. For instance, *MNRAS* contained 937 references and 9267 pages in 2000. In 2009, it had 1815 references and 19 916 pages. Currently, the CDS team indexes more than 12 000 articles each year. DJIN has been developed to help the team to deal with this huge number of papers.

2. Overview

DJIN is an acronym and stands for "Detection in Journal of Identifiers and Names"; its functionality is illustrated in Fig. 1.

For each article, described by a *bibcode* (a bibliographic code that identifies each paper, Schmitz et al. 1995), DJIN downloads the PDF file of the paper and extracts its contents in a readable format. The program then searches for object names directly in this text by comparing the words with information stored in the *Dictionary of Nomenclature of Celestial Objects* (Lortet 1994). The system retrieves all possible matches and submits to the information scientists a screen of the article where all of the object names are highlighted and linked to the SIMBAD database. It is therefore easy to check whether it is the right object, since all of the details are just one click away. After validation by the information scientist, DJIN generates the commands to update SIMBAD, and executes them.

The program is written in the Java language, which is also used for SIMBAD; the Java graphic library Swing is used for the Graphical User Interface (GUI).

3. Extraction of the Text: from PDF to TEXT

Each journal has its own way of formatting the PDF document, and the way DJIN extracts the content depends on this format. Some journals are written in one block, others use two-column pages. DJIN reassembles words and paragraphs in the correct order and detects the headers and footers because we don't want to have them in the middle of the text. Tables and figure captions have to be extracted separately for the same reason. The information about object name positions (in the title, subtitle, abstract, table, figure caption, etc.) is also collected at this stage. Furthermore, some special characters (e.g. Greek letters) are only described graphically in the PDF, and the program must compare these with a predefined set of graphical symbols in order to detect them.

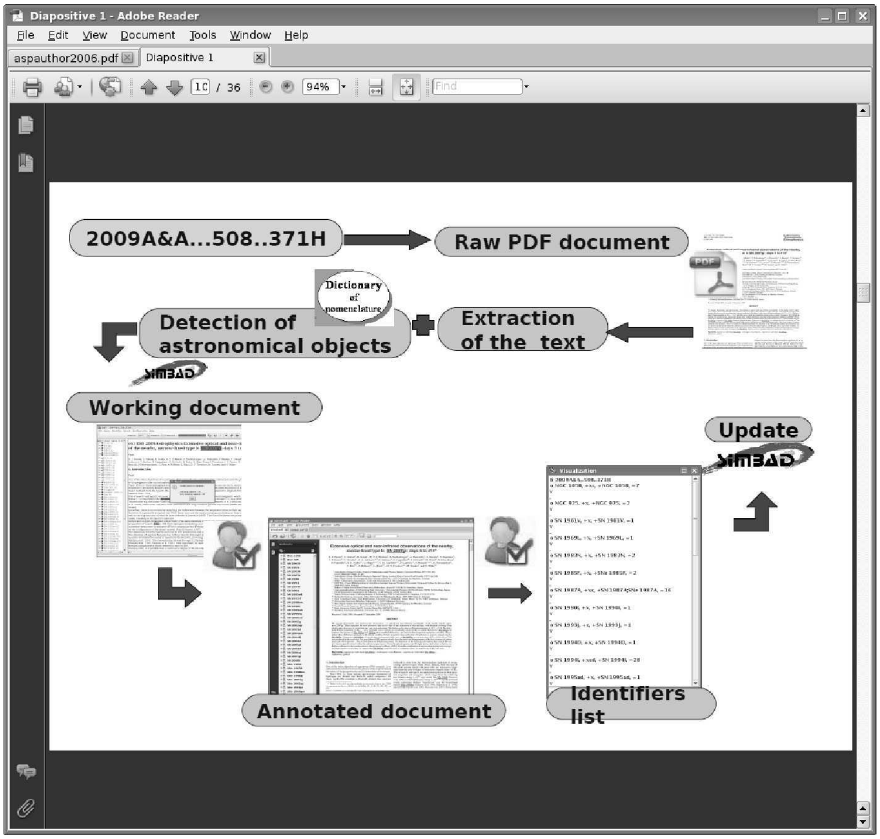

Figure 1. Schema of DJIN functionality

The Java open-source library `PDFBox`[1] is used for the manipulation of the PDF documents; DJIN works also with articles formatted in HTML.

4. Recognition of Object Names

4.1. Identifiers as Written in the Literature

An astronomical object name is not easy to define: it may be short, long, and structured or not; examples are "Orion Nebula", the "Superantennae", "DR21(OH)", "CCDM J00335+4509BC", "NGC 1866", "QSO 0347-3819", "Cl* NGC 2419 SAW V18", "T Tau N", etc.

A *structured name* is basically made of an *acronym* and a *numbering* which are both a string of alphanumeric characters. The *Dictionary of Nomenclature of Celestial*

[1] http://www.pdfbox.org/

Objects[2] is a reference work which collects all of the designations quoted in the literature. This dictionary, which is maintained on a regular basis by the CDS, provides full references and usages of about 19,500 acronyms.

A *non-structured name* does not follow this schema; typical examples are: variable star names based on constellation names preceded by one or two letters or a digit or a Greek letter ... and sometimes followed by a duplicity letter (examples: "α UMi", "V343 Hya", "RX And"); or common names such as "Aldebaran" or "the Crab". SIMBAD provides a list of 9000 such names. Identifiers can furthermore include graphical symbols (examples: "α^2 CVn", "SgrA*", etc.). The detection of identifiers in lists separated by commas ("IC 342, 99, 537"), or in tables (column headers may contain a part of the identifier) represents another challenge.

The extraction of all of these kinds of names in the full text is not straightforward. The way these objects are written is heterogeneous and varies from one paper to another, and can even vary within a given paper.

4.2. Matching Names

DJIN is based on the *Dictionary of Nomenclature of Celestial Objects* and on the different lists described above. The information contained in the Dictionary (*acronym* and *numbering*) cannot be directly used by computers. Therefore, we need to translate these into regular expressions; similar transformations are applied for the information coming from the other lists. The result is a list containing more than 50,000 regular expressions that have to be cut into elementary pieces and put into a search tree in order to be matched efficiently against the extracted texts.

In SIMBAD, an object name identifies in principle only one astronomical object. In the literature, acronyms have different synonyms and the usage provides inconsistencies. That is why object names found in the documents often correspond to several astronomical objects: for example **Perseus** may represent the SIMBAD objects PERSEUS REGION, or PERSEUS COMPLEX, or PERSEUS SUPERBUBBLE — 17 different entries exist in SIMBAD for "Perseus", depending on the context.

DJIN retrieves all possible matches, queries SIMBAD for each candidate object name to check its existence, and if it exists retrieves its description. The information scientists have then to find the correct object among the candidates proposed. DJIN generates an annotated PDF file with all of the details and direct links to SIMBAD, but in order to ensure the correctness and completeness of the recognition, a validation by an expert remains essential.

DJIN offers additional features such as searching for object names in tables or looking for incomplete object names. Object names that are not recognized or partially recognized can easily be added to the list of retrieved object names. Similarly, new identifiers can easily be added to the list.

One of the difficulties at this stage is the presence of false detections. For our application, we opted for completeness, which increases the rate of false detections. To reduce these false detections, we use the open-source learning engine Weka[3] to help discriminate between the right detections and the false ones. The program thus behaves

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

[3] http://www.cs.waikato.ac.nz/ml/weka/

like a spam detector with a trash to collect the rejected names and a set of learning data loaded at initialization.

5. Updating SIMBAD

The result of the two steps described above is a list of valid identifiers. For each object name on this list, the information scientists update SIMBAD with new astronomical data (object type, magnitudes, etc.) found in the article. If, in the article, one object name is cited with different identifiers, DJIN gathers them under the same object. For each object name, DJIN gathers additional information like the positions of the object names in the paper (in the title, subtitle, abstract, table, figure caption, etc.), their number of occurrences, and the name used by the authors in the paper to designate the object. The SIMBAD updating commands are then generated, and sent to SIMBAD. These commands can be simulated before their execution. Finally, the result is displayed so that the information scientist can check the final result.

6. Two Years of Usage

DJIN has been used daily by the team since January 2008. Despite a very deep change it introduced in the way we do our work, DJIN turns out to be very well adapted to the SIMBAD updating environment. It is a visual tool, easy to use, eliminates tedious work and lets us focus on value-added aspects.

After two years of usage, we may draw a few conclusions. First, we save time by avoiding the papers which do not include any object identifier (theoretical papers), and also with the papers containing only well known objects and simple lists. But, we **still need human checking by somebody who has a very good knowledge of astronomy** to remove the wrong detections and to detect the new identifiers in what is not highlighted in the DJIN result.

Based on a series of articles from several issues of several journals, we can derive the following estimations:

- rate of exact recognition: 75%
- rate of partial recognition: 12%. A partial recognition means that just a part (mainly the format) of the object name is recognized by DJIN (for example “Par-Lup3 4” is recognizes by DJIN, but the complete name is “[CFB2003] Par-Lup3 4”). A significant fraction of the partial recognition is related to new acronyms, meaning that DJIN is able to correctly detect what is not yet in the Dictionary.
- rate of missing recognition: 10%. A large fraction of the missing recognitions is due to the objects listed in the tables or figures. DJIN allows us to easily retrieve these objects in a second run with additional information; but DJIN can’t retrieve object names written in figures.
- rate of false detection: 3%
- rate of noise: 51%. This rate looks quite high, but it is relatively easy to remove this noise with some fine-tuning of the DJIN. On the other hand, the reduction of the noise decreases the rate of exact recognition.

Finally, to the most important question "Do we save time ?" the answer is: "It is uncertain."

7. Other Applications of DJIN

7.1. Identifying the Important Papers for One Astronomical Object

In SIMBAD, the number of bibliographical references attached to one astronomical object has been continuously growing with the accumulation of published literature over the years. Some objects are cited in thousands of papers, and it may be difficult to identify the most relevant papers, i.e. those which contain extensive results about that object.

The location of the object name in a paper (title, abstracts, keyword, tables, figures, text, etc.) is a simple but interesting criterion to identify relevant papers for one specific astronomical object, as is shown in Lesteven (2003). As we showed above, DJIN gives more than just recognition, it specifies how often an astronomical object is cited in a text whatever its identifier is and specifies where the object is cited (title, etc.). This information has been added for two years now, and we are currently working on the older papers to add the indication of citation in titles and keywords. Once this information is recovered, it will be possible to compute the relevance of each paper attached to each object in SIMBAD.

7.2. Links Between SIMBAD and NED

NED (NASA/IPAC Extragalactic Database) contains positions and other basic data for extragalactic objects, as well as bibliographic references from catalogs and other publications. A good fraction of SIMBAD and NED objects represent identical objects even if their identifiers differ. Because NED shares the information about its nomenclature with our Dictionary, it becomes possible to keep track of the relation between one SIMBAD identifier and the corresponding one in NED and vice-versa. It becomes then possible to maintain links between the two databases and furthermore to navigate from one database to the other one.

8. Conclusion

The success of DJIN is the result of the close collaboration of all of the people on the team (information scientists, astronomers and computer scientists). DJIN has been fully tested for two years to assess the quality of the detection, as well as the tool's ergonomics. This feedback has been critical in improving the development and the usage of DJIN.

Even if DJIN has not achieved its original purpose (to save time), it has made the work much more attractive, more complete and more reliable. And if the first trials in using DJIN were not easy — it represents a really big change compared to our previous working habits — after a few weeks of usage, nobody would go back.

Acknowledgments. This study is developed in the frame of the VO-TECH Design Study, funded by the European Commission.

References

Lesteven, S., Dubois, P. 2003, in Proc. Library & Information Services in Astronomy IV, ed. B. Corbin, E. Bryson & M. Wolf, (Washington DC: U.S. Naval Observatory), 243

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

Schmitz, M., Helou, G., Dubois, P., et al. 1995, in Astronomy and Space Science Library, Vol. 203, Information & On-line Data in Astronomy, ed. D. Egret & M.A. Albrecht (Dordrecht: Kluwer), 259.

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

Amelia Laurenceau, Nishtha Anilkumar, Molly White and Juana Maria Sainz Ballesteros during Pune sightseeing (Photo: E. Isaksson)