Library and Information Services in Astronomy IV
July 2-5, 2002, Prague, Czech Republic
B. Corbin, E. Bryson, and M. Wolf (eds)

Towards an Automated Retrieval of Publications based on Telescope Observations

Uta Grothkopf & Angelika Treumann

European Southern Observatory, Karl-Schwarzschild-Str. 2, Garching, Germany
esolib@eso.org

Abstract. We analyze papers based on VLT observations regarding (a) the information provided about facilities used and (b) possible retrieval through ADS. The following topics are looked at:

* Where in the papers do authors report on the facilities used and how detailed is this information? Can it be used to measure scientific output of observatories, telescopes, instruments and observing programs?

* Which percentage of relevant papers is retrieved through ADS, how many irrelevant papers are picked up, and why?

* How can automated retrieval of papers be improved and which factors remain problematic?

We conclude that the currently available retrieval options are not sufficiently reliable to abandon manual literature screening. The implementation of dedicated LaTeX-tags in the macros of core astronomy journals and a specific search option through ADS would probably improve the procedure considerably and help to move towards (mainly or entirely) automated retrieval of papers.

1. Introduction

Tracing scientific papers based on telescope observations has become increasingly important. Observatories want to know which articles as well as how many were produced by staff and visiting astronomers for a variety of purposes:

- to evaluate the performance of telescopes and instruments
- to track scientific output from individual observing programmes
- to measure the efficiency of the organization
- for comparison with other observatories and telescopes
- to compile the Annual Report bibliography
- to build Virtual Observatories by linking observing proposals, archived data and published papers

Unlike bibliographic citations, no system is in place that provides similar statistics regarding telescope use. Observatories interested in monitoring papers based on data from their facilities have to create bibliographies manually by screening the scientific literature.

The compilation process is tedious, time-consuming, and more complex than it may seem at first glance. Procedures vary among observatories, as do the selection criteria for inclusion of papers. Comparing statistics across observatories therefore is difficult, if not impossible.

One would assume that the process could be facilitated through automated retrieval procedures, ideally using the NASA ADS Abstract Service. But can automated searches really emulate manual screening? Will results be identical?

The aim of our study was to find out how reliable retrieval via ADS is at present and, if necessary, how it could be improved. In this context, the actual numbers resulting from our example are less important than the general conclusions that can be drawn. We did not intend to give indisputable metrics but wanted to show differences between manual and automated retrieval and provide suggestions for improvement.

2. The ESO Telescope Bibliography

At ESO, a considerable amount of time and effort is spent on creating the telescope bibliography. We are not only interested in the total number of papers based on data from our facilities but also in detailed information regarding telescope sites and instruments used to obtain the data. For the AVO (Astrophysical Virtual Observatory, see `http://www.eso.org/avo/`), it is important to determine the programme IDs under which observations were made so that observing proposals, archived astronomical data, and published papers can be interlinked.

Before literature screening can start, selection criteria for publications have to be defined. This is not done by the library alone but by the Office for Science. As of 2003, the ESO telescope bibliography contains papers which (1) accomplish new research or new interpretation based on ESO data and (2) are published in refereed journals. This excludes papers describing technical aspects of instruments, but includes those that use ESO data as well as data obtained with other telescopes.

Theoretically, authors should acknowledge facilities in the footnotes of publications; however, not all of them actually do so. If a clear indication of ESO telescope use is missing, it must be understood from other parts of the text, for instance the abstract, the section describing the observations, or figure captions. Often programme IDs are not mentioned and have to be identified through the ESO Observing Schedule, sometimes leading to e-mail communication with authors to verify eligibility of papers. Complementary searches are carried out in ADS at the end of the year.

Because of the thorough compilation process, the ESO telescope bibliography is regarded as complete. It serves as the reference database for this study.

3. Preparing Automated Searches

For a comparison of manual versus automated searches, we focused on a subset of papers: the 2001 VLT (Very Large Telescope) bibliography; it contains all papers based on VLT observations, published in the year 2001 in refereed journals. Our

intention was to retrieve exactly these papers through ADS. Restricting queries to the publication year 2001 and to refereed journals was simple because ADS provides adequate filters. It was more problematic though to select appropriate search terms capable of identifying all relevant papers and ignoring all irrelevant ones. Various test searches were carried out. Problems encountered during the test phase included:

- Inappropriate search terms: searches for "Very Large Telescope" retrieved a high number of general papers on "large telescopes". We realized that "very" is a stop word and therefore ignored by the system. The ADS Team kindly offered help by implementing case-sensitivity; the capitalized version is now a proper search term.
- Ambiguous acronyms: one of the instruments currently in use is called ISAAC. Searches for this term led to erroneously retrieved articles mentioning the "Isaac Newton Group" and "Isaac Newton Telescope" rather than the VLT instrument.
- Non-standardized instrument names: authors use various names for facilities; for instance, UVES, Ultraviolet-Visual Echelle Spectrograph and UV-Echelle Spectrograph all refer to the same instrument. Our query had to guarantee that all relevant papers were retrieved, regardless of the individual names used for telescopes and instruments.

The test searches revealed that all relevant papers mention either the telescope (VLT, Very Large Telescope), the telescope site (Paranal) or the instrument (available in 2001: UVES, FORS1, FORS2, ISAAC). We concluded that the search string

> VLT or Very Large Telescope or Paranal or
> UVES or FORS or FORS1 or FORS2 or ISAAC not Newton

was suitable for our purpose.

4. Comparison of Manual versus Automated Searches

We were now ready to compare the manually compiled ESO database to the list of papers retrieved via ADS. The comparison resulted in the following figures: The VLT bibliography for the year 2001 contained 100 papers; by nature, all of them are relevant hits. ADS retrieved 106 hits. At first glance, this looked like a close match. However, a detailed analysis revealed interesting facts: only 83 of the 106 papers were relevant; 23 papers were irrelevant and should not have been picked up; 17 relevant papers were not retrieved at all (Fig. 1).

All 23 irrelevant papers picked up by ADS contained search terms in the abstract; however, the context was not the one we intended. Most of them were instrumentation or technical rather than scientific papers; others suggested future projects, using ESO telescopes. Several papers used an acronym defined in our search, but the meaning was completely different. For instance, VLT is the short version of Very Large Telescope but also means Very Low Ti basalt, Very Long Term variation etc. Other reasons, lower in number, were references to research accomplished by others as well as errata.

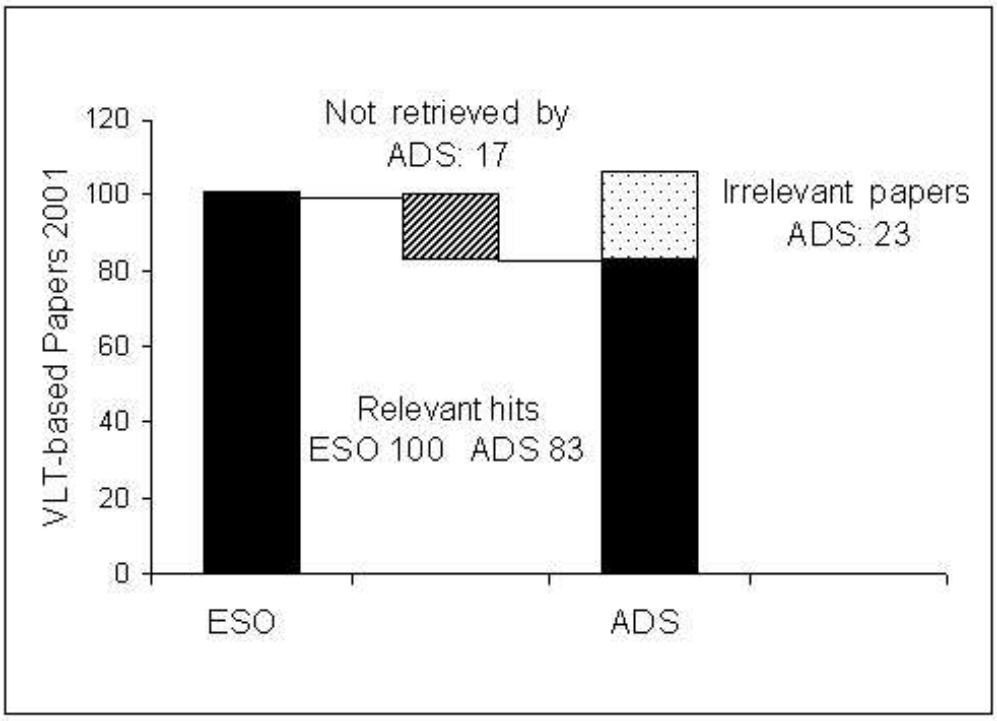

Figure 1. Manual vs. Automated Searches (ESO vs. ADS).

The 17 missed papers had been published in Nature (6), ApJ (5), A&A and MNRAS (2 each) as well as AJ and NewAR (1 each). Most of them could not be retrieved because search terms appeared only in text sections which are not accessed by ADS, most notably the "observations" section. Contrary to common belief, ADS text searches currently cover only title, abstract and footnotes; papers acknowledging facilities elsewhere in the text are not picked up. A few failures had to be attributed to technical glitches.

5. Details currently available through ADS searches

We used the 2001 VLT bibliography again to investigate in which parts of papers authors acknowledge telescope use and how detailed this information is. Four categories were defined, representing

1. the observatory (ESO, European Southern Observatory)
2. the telescope site (VLT, Very Large Telescope, Paranal)
3. the instruments (UVES, FORS1, FORS2, ISAAC)
4. programme IDs

Mentioning of facilities in more than one field resulted in multiple counting. Not surprisingly, facilities were acknowledged most frequently in the footnote. However, only 80% of the authors of our sample did so; to ensure reliable automated retrieval, 100% would be necessary. 30% of the papers contained search terms in the title, 66% in the abstract. With regard to automated retrieval, the

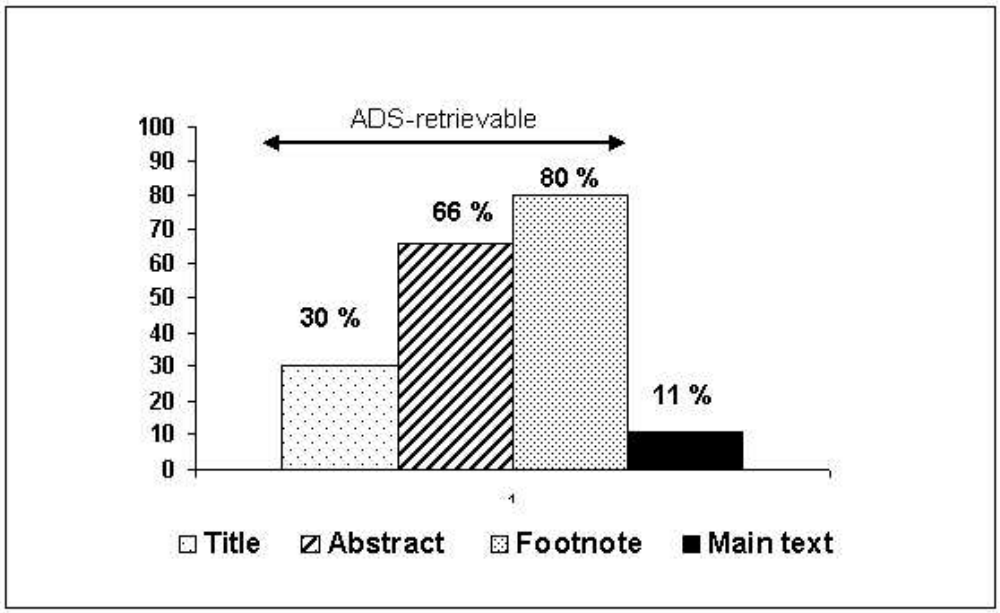

Figure 2. Where do authors acknowledge telescope use?

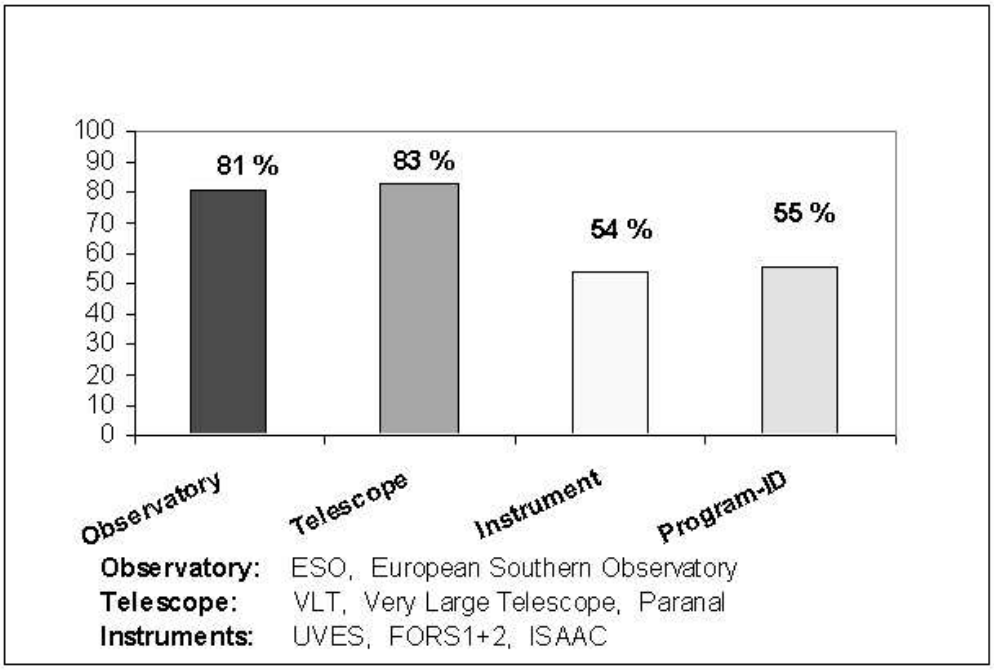

Figure 3. Which details are provided in ADS-searchable fields?

latter group is problematic as search terms often appear in unintended context (Fig. 2).

These findings were confirmed by another analysis of the same set of papers. This time, we looked at the details provided in ADS-searchable fields (title, abstract, footnote). Complete information could not be extracted for any of the four categories. While observatory and telescope site were acknowledged rather frequently (81% and 83% of the papers, respectively), credit to instruments and programme IDs was only given in every other paper (54% and 55%) (Fig. 3).

6. Conclusion

At present, the only reliable way of tracing publications based on telescope observations is through manual literature screening. Database searches in ADS, the most comprehensive bibliographic system in astronomy and astrophysics currently available, neither retrieve all relevant papers nor exclude all irrelevant ones. Bibliographies relying on information derived from these searches would be incomplete and incorrect.

The main reasons for the failure of automated retrieval are the lack of standards regarding acknowledgments of facilities and insufficient application of existing rules. The findings of our study suggest that in order to

- retrieve all relevant papers
- avoid retrieval of irrelevant papers and
- extract reliable statistical information

facilities should be acknowledged in dedicated fields searched by ADS. A possible solution would be the implementation of a special LaTeX-tag for journal macros, agreed upon by major astronomy publishers.

ESO advocates the idea, provided that resulting costs are reasonable. Similar ideas are considered by U.S. American journal editors and the ADS who are discussing so-called "data set identifiers". Specific requirements will have to be defined.

However, several areas remain problematic. To achieve good results, as many publications as possible should be covered. This requires an agreement by all major journals which may be difficult to achieve. Unless editorial offices or certain departments within organizations volunteer to check manuscripts for correct use of the tag, the additional workload and the entire responsibility will be shifted to authors. Guidelines and rules must be developed to ensure proper usage; yet the system must be flexible and extensible. Education will be required to make sure procedures are understood and applied consistently. Possible typos, technical glitches, incomplete or wrong information, etc. cause additional concern.

Before automated retrieval techniques can replace manual compilation of telescope bibliographies, careful preparation will be necessary. Quality control must be in place so that accuracy, completeness, and consistency with previous bibliographies is guaranteed.