Astronomical Data Analysis Software and Systems IV
ASP Conference Series, Vol. 77, 1995
R. A. Shaw, H. E. Payne, and J. J. E. Hayes, eds.

ADS Abstract Service Enhancements

A. Accomazzi, C. S. Grant, G. Eichhorn, M. J. Kurtz, and S. S. Murray

Smithsonian Astrophysical Observatory, 60 Garden Street, Cambridge, MA 02138

Abstract. The Astrophysics Data System is enhancing functionality and access to the Abstract Service in several areas. A new WAIS server running the freeWAIS search engine has been added to the existing HTTP and ANSA-based servers. Abstract coverage will be expanded to include more of the NASA "RECON" categories as well as abstracts obtained directly from astronomical journals. In addition, we are enhancing the service to include full journal articles in a suitable electronic format.

1. Introduction

The Astrophysics Science Information and Abstract Service (ASIAS) of the Astrophysics Data System (ADS), formerly known as "Abstract Service," has been very successful in providing the researcher and librarian the capability to search the astronomical literature. It currently provides access to over 160,000 astronomical abstracts with a sophisticated search engine.

Use of the service increased dramatically after it was made available on the World Wide Web (WWW), and now averages 30,000 queries and 500,000 retrieved abstracts per month. Its ease of use, flexibility, and data coverage have made it a very well known resource in the astronomical community. In this paper we discuss the current capabilities of the system and its planned enhancements.

2. The Database

The ASIAS provides access to astronomical abstracts since 1975 from the following five categories defined in NASA's Scientific and Technical Information (STI) Database: Astronomy, Astrophysics, Lunar and Planetary Exploration, Solar Physics, and Space Radiation.

The NASA database contains abstracts from over 200 journals, publications, colloquia, symposia, proceedings, and internal NASA reports. They include the majority of astronomical journals as well as several sources from journals loosely related to astronomy. Due to the incomplete coverage of some of these journals by NASA RECON and delays in providing us the abstracts, the database is at the time of this writing about 95% complete and one year behind current publications. As soon as NASA resumes the abstracting, which was interrupted for almost a year, we hope to provide on-line papers which are only a few months behind publication date.

3. The Server

The ASIAS is part of a larger set of programs that can read, format and index the abstracts compiled by NASA RECON and then search them according to queries generated by users on the Internet. The server itself is composed of a module that accepts network requests and a second one that implements the search algorithm. This design has allowed us to keep our efforts focused by having to maintain only one core software system (along with the set of auxiliary files needed to perform a search), instead of having to maintain one for each protocol supported. We will refer to this system as the "ADS search engine" in the rest of the paper.

The abstract server was designed to allow queries on separate "fields" in the documents. For instance, by defining separate fields for abstract text and paper title, a query can specify a term to be searched for only in the title but not in the abstract text. Complex queries can be composed by searching for different terms in some of the fields and then combining the results according to their relevance with respect to the original query. Query results are ranked by a score determining how they match the input query.

3.1. Access Methods and Protocols

As part of the ongoing effort to migrate the functionality and services of the Astrophysics Data System to protocols and software systems in the public domain, we now provide WWW and WAIS access to the ASIAS.

The server itself was originally written as part of a service in an earlier EOS/ADS system based on the ANSA (Advanced Network System Architecture) protocol. WWW access (based on the HTTP protocol) has been available since February 1994. As new HTML features have been implemented in WWW clients, we have been able to provide additional capabilities to our server, which now provides greater functionality and better performance than the ANSA-based one. Users can reach the Abstract Service WWW query form[1] using any browser that supports HTML forms.

The WWW-based server has been implemented by writing a Common Gateway Interface between the HTTP daemon and the server search engine which carries out the query and returns the results. Similarly, the ANSA based server starts a session manager which then runs the same search engine.

Recently we have brought on-line a WAIS server[2] running the freeWAIS-sf[3] search engine. Because of the limitations in the freeWAIS software and the Z39.50-1988 protocol it uses, only a portion of the functionality and flexibility offered by the WWW server are provided to WAIS clients, so whenever possible the WWW server should be used instead. The ADS WAIS server is currently used by the NASA Technical Report Server (NTRS)[4].

---

[1] http://adsabs.harvard.edu/abstract_service.html

[2] http://adswww.harvard.edu/abs_doc/wais/

[3] http://ls6-www.informatik.uni-dortmund.de/freeWAIS-sf/README-sf

[4] http://techreports.larc.nasa.gov/cgi-bin/NTRS

3.2. The ADS Search Engine

The ADS search engine currently provides great flexibility in searching and scoring documents by allowing users to change, among others, the following settings: the relative field weights to be used when computing the score of a document; what boolean logic is to be used when combining the results for each field; and whether synonym replacement is to take place on selected fields.

A query typically returns a list of documents ranked by relevance. The user can then choose whether to retrieve one or more of the selected abstracts. When a single abstract is retrieved, different portions of the abstract itself can then be selected to perform a "relevance feedback" query by requesting papers similar to the current one.

An additional feature that we recently introduced is providing links to data tables published with the paper, when available. The links are returned as part of the selected documents. Currently the only source of such tables is the Strasbourg Astronomical Data Center (CDS).

3.3. ADS Search Engine vs. freeWAIS

As part of our effort to stay abreast of the technology in the field of distributed database systems and search methodologies, we decided to invest some time using a WAIS-based system to index and search our dataset. We selected freeWAIS-sf among the WAIS packages currently available in the public domain since we found it to be the most advanced; in particular, it is superior to the CNIDR freeWAIS[5] version in that it introduces the concepts of search fields and document structure.

Some time was spent in enhancing the freeWAIS code to run faster by storing some of the frequently accessed data in shared memory, to allow better control of what words would be ignored when indexing, and to support extended headlines (the document identifier strings returned by the server upon completion of a query). These changes have since been incorporated in the freeWAIS code.

As a result, we were able to compare the performance and features of our search engine vs. the freeWAIS-sf one on the same set of data. Our tests have shown that our search engine is from 5 to 20 times faster than the freeWAIS-sf one, while the indexing process needed to create the inverted indexes used by our search engine is about 15 times faster than the freeWAIS-sf version.

This dramatic difference in performance is largely due to the fact that the ADS search engine was designed to work well on the particular dataset at hand, rather than to be a general purpose software package. In particular, the use of small inverted indexes kept in shared memory and the inclusion of publication years in the indexes allows searches to be carried out very quickly.

In addition, many of the search features that are available when using the WWW version of our server could not be implemented because of limitations in the freeWAIS-sf package. For instance, synonym replacement in freeWAIS applies indiscriminately to all of the document's fields, and cannot be selectively turned off at query time since it is built into the indexes. Similar problems exist

[5] http://cnidr.org/cnidr_projects/freewais.html

for the list of "stop words" (words to be ignored when indexing the document), and the minimum length of words to be indexed.

Despite these problems and reservations, we believe that freeWAIS-sf remains the best public-domain general purpose full-text indexing and search engine available today.

4. Planned Enhancements

The emphasis of the ADS in the coming years will be to utilize modern information systems technologies like WWW to provide increased access to its services for a wide variety of users through public domain client software. The protocols that are currently used for search and retrieval are Z39.50 and HTTP. While we do not expect that to change in the next few years, we intend to keep providing a reliable abstract server capable of understanding whatever protocols are in use in the astronomical community on the Internet.

We plan on expanding the abstract database to cover more topics, e.g., Instrumentation and Space Physics. We are also examining adding to the functionality of the Abstract Service by including a citation index that will allow users to browse through the abstracts of references associated with the current abstract.

Development work will include cooperation with the publishers of astronomical literature to provide access to the original author abstracts. We will also work on providing access to the full articles as image bitmaps. As a first step in this direction we plan to bring on-line several years of the Astrophysics Journal Letters as a test case. User response to having full journal articles available and linked with the abstracts will be evaluated. If it proves to be a valuable service, we will work with publishers to digitize more of the old literature and to see whether we can provide access to electronic forms of new articles.

Recently it has become possible to "publish" electronically data tables from a journal article. We have started work on linking these data tables to the abstracts of the articles. Currently we are making use of the on-line data available through the CDS. Our objective is to provide additional links to any data source which is relevant to the documents we have on-line. Obviously this is a long-term goal and will require a lot of work until standards for the access, classification and retrieval of astronomical data available on the network are agreed upon and implemented.

We believe that the efforts described above will extend the scope of the Astrophysics Science Information and Abstract Service and expand it into a wide ranging system with enhanced utility for the astronomical community.

Acknowledgments. We are grateful to Ulrich Pfeifer for his work in enhancing the freeWAIS software and for being receptive to suggestions and proposed changes to the package. This work is funded the NASA Astrophysics program under grant NCCW-0024.