THE DIGITAL LIBRARY IN ASTRONOMY: AN EXAMPLE OF A WORLDWIDE COLLABORATION OF INFORMATION PROVIDERS

Guenther Eichhorn[1], Alberto Accomazzi, Carolyn S. Grant, Michael J. Kurtz and Stephen S. Murray
Harvard-Smithsonian Center for Astrophysics, Cambridge, MA 02138
gei@cfa.harvard.edu

The NASA Astrophysics Data System (ADS) is the central part of the Digital Library in Astronomy. The astronomical community in the last decade has built a system of interlinked data centers that is unique among the sciences. It is composed of the closely integrated operations of the ADS, CDS, NED, other data and archive centers, and the astronomical journals. It began functioning in 1993 with the ability to make joint queries to the ADS and the SIMBAD database at the CDS. It now provides the user with a tightly integrated web of information links.

The ADS provides access to over 2.6 million abstracts and over 2 million scanned pages from 276,000 journal articles. One important part of the ADS is the archive of the historical literature. This archive includes all the major journals back to volume 1, as well as about 300,000 pages of scanned historical observatory publications.

The list of references that is returned from a query includes an extensive set of links to many different on-line data sources for each reference. Currently there are over 6 million links in the ADS, about half of these to information sources outside the ADS. This extensive linking to other on-line resources makes the ADS and the Astronomy Digital Library unique in the sciences.

The ADS is funded by NASA and is accessible without restrictions at:
http://ads.harvard.edu

1. Introduction

The **A**strophysics **D**ata **S**ystem (ADS) Abstract Service is a central facility of bibliographic research in astronomy. It is used by over 50,000 users per month. More than 10,000 of these use the ADS regularly (10 times or more per month). The ADS is a key element in the emerging digital information resource for astronomy, which has been dubbed Urania [1]. The ADS is tightly interconnected with the major journals of astronomy, and the major data centers. A detailed description of the ADS has been published in a special issue of Astronomy & Astrophysics Supplements in April 2000 ([2, 3, 4, 5]).

[1] Send correspondence to Guenther Eichhorn. Email: gei@cfa.harvard.edu

The first major component of the ADS is the Abstract Service. It was started in 1993 with a custom-built networking software system to provide access to distributed data ([6]). By the summer of 1993 a connection had been made between the ADS and SIMBAD (**S**et of **I**dentifications, **M**easurements and **B**ibliographies for **A**stronomical **D**ata, [7]) at the **C**entre des **D**onnées de **S**trasbourg (CDS), permitting users to combine natural language subject matter queries with astronomical object name queries ([8]). A description of the system at that time is given in [9]

By early 1994 the **W**orld **W**ide **W**eb (WWW [10]) had matured and was widely accessible through the **N**ational **C**enter for **S**upercomputing **A**pplications (NCSA) Mosaic Web Browser ([11]). It now was possible to make the ADS Abstract Service available via a web forms interface. This interface to the ADS is described in [13] and [14].

The second major part of the ADS is the Article Service. In December 1994 the first bitmaps were put on-line ([15]). The system contains scanned journal articles for most of the astronomical journal literature back to volume 1.

Over time, other interfaces to the abstracts and scanned articles were developed to give other information providers the means to integrate ADS data into their system ([17]).

With the adoption of the WWW user interface and the development of the custom-built search engine, the current version of the ADS Abstract Service was in place. Currently the ADS system consists of four databases covering Astronomy/Planetary Sciences, Instrumentation, Physics, and Astronomy Preprints. Combined there are over 2.6 million abstracts and bibliographic references in the system. The Astronomy Service is by far the most advanced, and accounts for ~85% of all ADS use ([2, 3]).

The following sections will describe the data (section 2), the user interface (section 3), the system of mirror sites (section 4), and the usage statistics of the ADS (section 5).

2. Data

2.1. Abstracts

The abstracts in the ADS come from many different sources ([5]). The original set came from the NASA **S**cientific and **T**echnical **I**nformation (STI) database. We now receive basic bibliographic information (title, author, and page number) from essentially every journal of astronomy. Most publishers also send us abstracts. Some publishers, who cannot send abstracts, allow us to scan their journals. For these journals, we build abstracts through **O**ptical **C**haracter **R**ecognition (OCR). Finally, we receive abstracts from the editors of conference proceedings, and from individual authors.

As of July 2002 there were ~807,000 astronomy references indexed in the ADS. The database is complete for all the major and most smaller journals

from Volume 1 with at least tables of contents. In the Physics database there are ~1.25 million references, and in the Instrumentation database there are ~628,000 references. Approximately half of all references have abstracts, the other half only have titles, authors, and journal information.

2.2. Fulltext

The ADS has obtained permission to scan, and make freely available on-line, page images of the back issues of all major journals and most smaller journals in astronomy. We plan to provide for each collaborating journal, in perpetuity, a database of page images (bitmaps) from volume 1 page 1 to the first issue which the journal considers fully on-line. This will provide (along with the indexing and the more recent archives held by the journals) a complete electronic digital library of the major literature in astronomy.

In the future, we will scan old observatory reports and defunct journals, to finally have a full historical collection on-line. This work has already started through collaboration with the Wolbach Library at the Harvard-Smithsonian Center for Astrophysics and the Harvard Preservation Project ([18,19]).

As Of July 2002, there are over 2 million scanned pages on-line in the ADS in 276,000 articles. The bitmaps in the ADS have been scanned at 600 dpi using a high-speed scanner and generating a 1-bit/pixel monochrome image for each page ([5]). The scanned pages can be retrieved in several formats that are generated on demand (e.g. Postscript, PDF, etc).

2.3. Links

The ADS responds to a query with a list of references and a set of hyperlinks showing what data are available for each reference ([3]). There are about 7 million hyperlinks in the ADS, of which almost 50% are to sources external to the ADS project.

The largest numbers of external links are to SIMBAD, the **N**ASA **E**xtragalactic **D**atabase (NED), the **S**pace **T**elescope **Sc**ience **I**nstitute (STScI), and the electronic journals. A rapidly growing number of links point to data tables created by the journals and maintained by the CDS and the **A**stronomical **D**ata **C**enter (ADC) at the Goddard Space Flight Center. These links are an extremely important aspect of the ADS. A more detailed description of resources that these links point to is provided in [5].

2.4. Citations and References

The use of citation histories is an effective tool for academic research ([20]). In 1996 the American Astronomical Society purchased a subset of the Science Citation Index from the **I**nstitute for **S**cientific **I**nformation (ISI), to be used in the ADS; this was updated in 1998. This subset only contains references that were already in the ADS; thus, it is seriously incomplete in referring to arti-

cles in the non-astronomical literature. The citation information from ISI spans literature published from January 1982 to September 1998.

The electronic journals all have machine readable, web accessible, reference pages. The ADS points to these with a hyperlink where possible. Several publishers allow us to use these lists to maintain citation histories.

Additionally we use OCR to create reference and citation lists for the historical literature, after it is scanned ([21]). This process has handled over 10 million references and added over 6 million parsed references to the ADS citation database.

2.5. Data Formats and Data Providers

The ADS receives data from numerous publishers. The data are in various formats, the most common ones being the ADS internal format, SGML, XML, and LaTeX. We can handle most data formats as long as they are consistent and can be automatically parsed. The minimum information that is required for entry into the ADS is the publication name, volume, page, author list, and title for each article. Other valuable data that are used by the ADS are the abstract, author affiliations, object names, keywords, and reference lists.

Any publisher who wants to have records included in the ADS needs to supply us with at least the minimum information mentioned above for each published article. If you want to have your published literature included in the ADS, please contact the first author for more information.

3. User Interface

The ADS services are usually accessed through a forms based interface. Other interfaces are described in [3]. This section describes the forms based user interface and its use, as well as the returned results.

3.1. Search Form

The main query form (figure 1) provides access to the abstract databases.

The query form allows the user to specify search terms in different fields. The input parameters in each query field can be combined in different ways, as can the results obtained from the different fields. The combined results can then be filtered according to various criteria (see [3]).

The database can be queried for author names, astronomical object names, title words, and words in the abstract text. References can be selected according to the publication date.

The database is searched for the specified terms as well as for terms that are synonymous. Successful searches in a free text search system require the ability to not only find words as specified, but also similar words. This starts with simply finding singular and plural forms of a word. It needs to be extended to different words with the same meaning in a particular field of science. In As-

tronomy, for instance, "metallicity" and "abundance" have a similar meaning and both need to be found in a search for either one. The list of synonyms was developed manually by going through the list of words in the database and grouping them according to similar meanings. This synonym list is a very important part of the ADS search system and is constantly being improved ([4]).

The ADS system was designed for expert astronomers. It provides sophisticated options that allow for various search types. The default parameters were selected to yield the best results for normal use. The general philosophy for the design of the search system and the default options was to favor recall over precision. This means that the ADS will usually return more results that may be marginally interesting rather than restrict the results very tightly. This allows the astronomer to make the decision about what is important and what is not.

Figure 1: A query to the ADS Abstract Service requesting a listing of papers on the metallicity of M87 globular clusters. The results from the queries to SIMBAD, NED, and the ADS database are combined and the list shown in figure 2 is returned.

3.2. Display of Search Results

The ADS system returns different information, depending on what the user request was. This section describes the different output formats.

3.2.1. Short Reference Display

The list of references from a query is returned in a tabular format. The returned references are sorted by score first. For equal scores, the references are sorted by publication date with the latest publications displayed first.

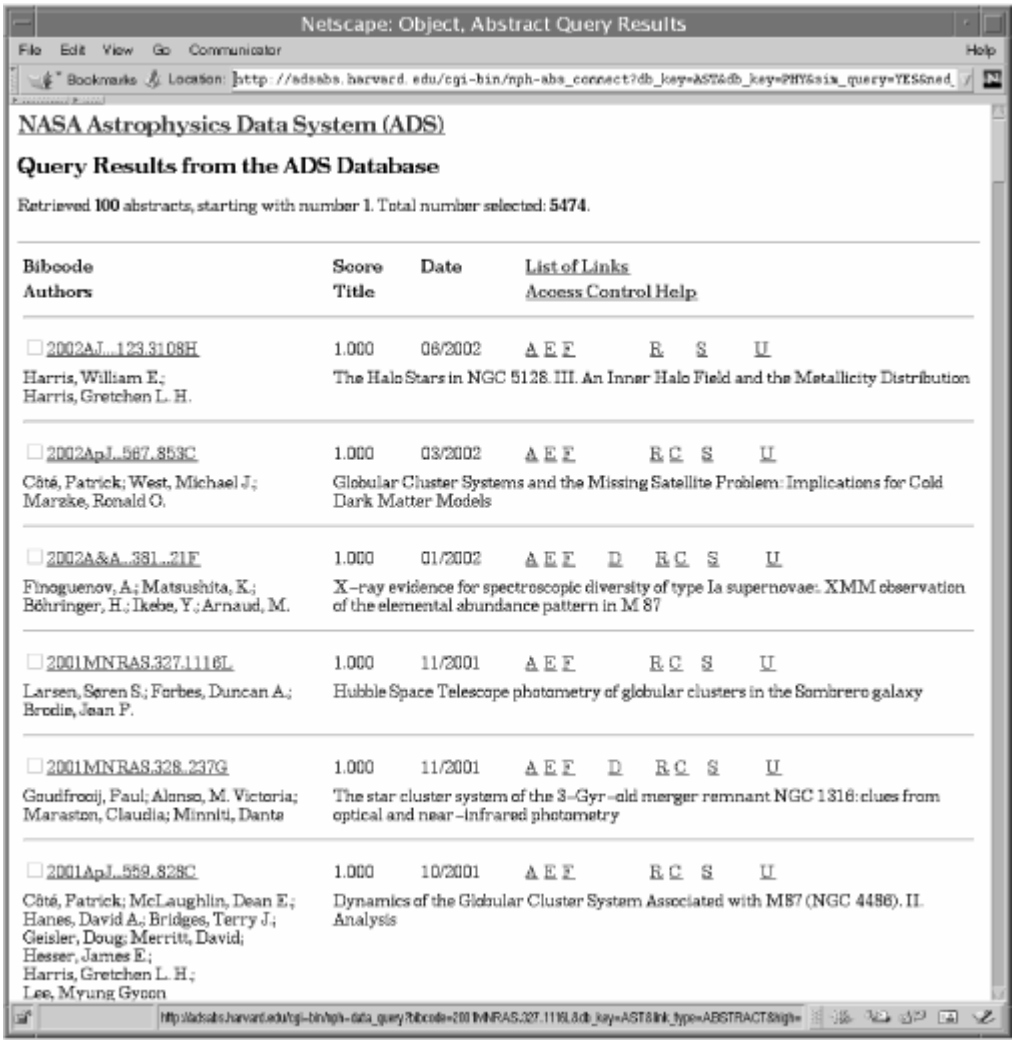

Figure 2: The result list from the query shown in figure 1.

The fields in the results list in figure 2 are as follows:

1. Bibliographic Code: This code identifies the reference uniquely (see [5, 22] for a description). Two important properties of these codes are that they can be generated from a regular journal reference, and that they are human readable and can be understood and interpreted.

2. Score: The score is determined during the search according to how well each reference fits the query.

3. Date: The publication date of the reference is displayed as mm/yyyy.

4. Links: The links are an extremely important aspect of the ADS. They provide access to information correlated with the article (see [2]).

5. Authors: This is the list of authors for the reference. Generally, these lists are complete. For some of the older abstracts that we received from NASA/STI, the author lists were truncated at 5 or 10 authors, but every effort has been made to correct these abbreviated author lists ([5]).

6. Title: The complete title of the reference.

The reference lists are returned as forms. The user can select some or all of the references in that form to be returned in any one of several formats:

i. HTML Format: The HTML (**H**yper**T**ext **M**arkup **L**anguage) format is for screen viewing of the formatted record.

ii. Portable Format: This is the format that the ADS uses for exchanging references (see http://adsabs.harvard.edu/abs_doc/abstract_format.html)

iii. BibTeX Format: This is a standard format that is used to build reference lists for TeX formatted articles.

iv. ASCII Format: This is a straight ASCII text version of the abstract. All formatting is done with spaces, not with tabs.

v. User Specified Format: This allows the user to specify in which format to return the reference. The user can specify a format string in the user preferences. This format string will then be used as the default in future queries.

vi. Other Formats: There are several other formats available that allow the user to easily import the data into other reference managing programs or various LaTeX macro sets.

The user can select whether to return the selected abstracts to the browser, a printer, a local file for storage, or email it to a specified address.

3.2.2. Full Abstract Display

In addition to the information in the short reference list, the full record (figure 3) includes, where available, the journal information, author affiliations, language, objects, keywords, abstract category, comments, origin of the reference, a copyright notice, and the full abstract. It also includes all the links to online data.

The full abstract display includes a form at the bottom that provides the capability to use selected information from the abstract to build a new query to find similar records. This query feedback mechanism is a very powerful means to do exhaustive literature searches and distinguishes the ADS system from most other search systems.

3.2.3. Full Article Display

The article display shows the first page of an article. Below the page image are links to every page of the article individually. This allows the user to directly access any page in the article. Wherever possible, plates that have been printed separately in the back of the journal volume have been bundled with the articles to which they belong for ease of access.

The next part of the displayed document provides access to plates in that volume if the plates for this journal are separate from the articles. Another link retrieves the abstract for this article.

The next part of the page allows the printing of the article or individual pages in a user defined format (default is 600 dpi PDF).

All possible printing options can be accessed through the next link called "More Article Retrieval Options". This page allows the user to select all possible retrieval options. These options are described in detail in [3].

4. Mirror Sites

The ADS is mirrored worldwide at 10 sites. Table 1 shows the current mirror sites and their URLs.

Setting up a mirror site is fairly easy. The hosting institution has to provide a server and an Internet connection. An abstract service mirror site can now run on a Linux PC with 20 Gb of disk space. If you are interested in having a mirror site, please contact the first author for detailed requirements.

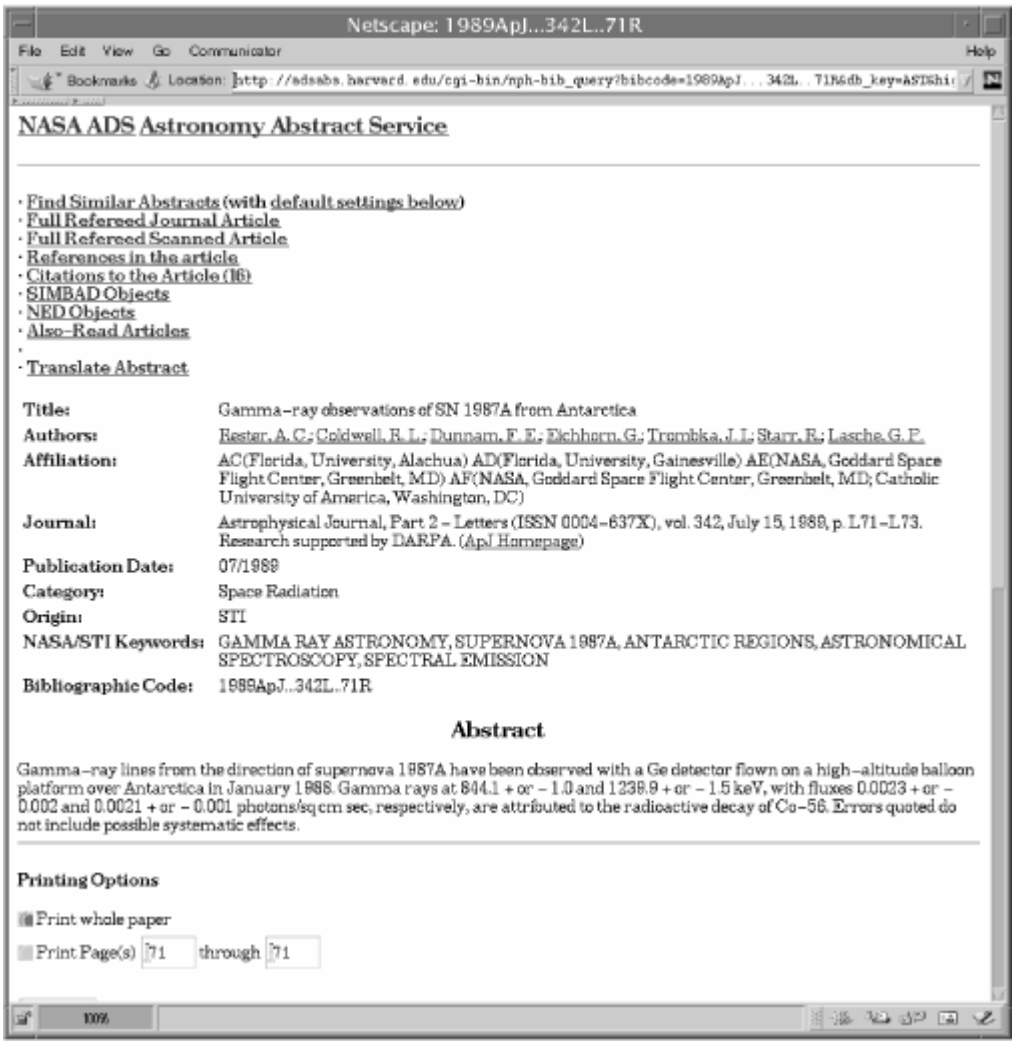

Figure 3: Example of a full abstract.

Country	Mirror Site	URL
USA	Harvard-Smithsonian CfA, Cambridge, MA	http://ads.harvard.edu
France	Centre des Données astronomiques de Strasbourg	http://cdsads.u-strasbg.fr
Japan	National Astronomical Observatory, Tokyo	http://ads.nao.ac.jp
Chile	Pontificia Universidad Católica, Santiago	http://ads.astro.puc.cl
Germany	European Southern Observatory, Garching	http://esoads.eso.org
England	University of Nottingham, Nottingham	http://ukads.nottingham.ac.uk
China	Beijing Astronomical Observatory, Beijing	http://baoads.bao.ac.cn
India	Inter-Univ. Centre for Astron. and Astroph., Pune	http://ads.iucaa.ernet.in
Russia	Institute of Astr., Russian Acad. of Scie., Moscow	http://ads.inasan.rssi.ru
Brazil	Observatorio Nacional, Rio de Janeiro	http://ads.on.br

Table 1: Mirror sites of the ADS

5. Use of the System

The ADS is used by a large majority of professional astronomers world-wide on a daily basis, as well as by many other researchers and non-scientists. This section shows some of the access statistics of the ADS.

Figure 4 shows the number of references retrieved per month since the start of the ADS in 1993.

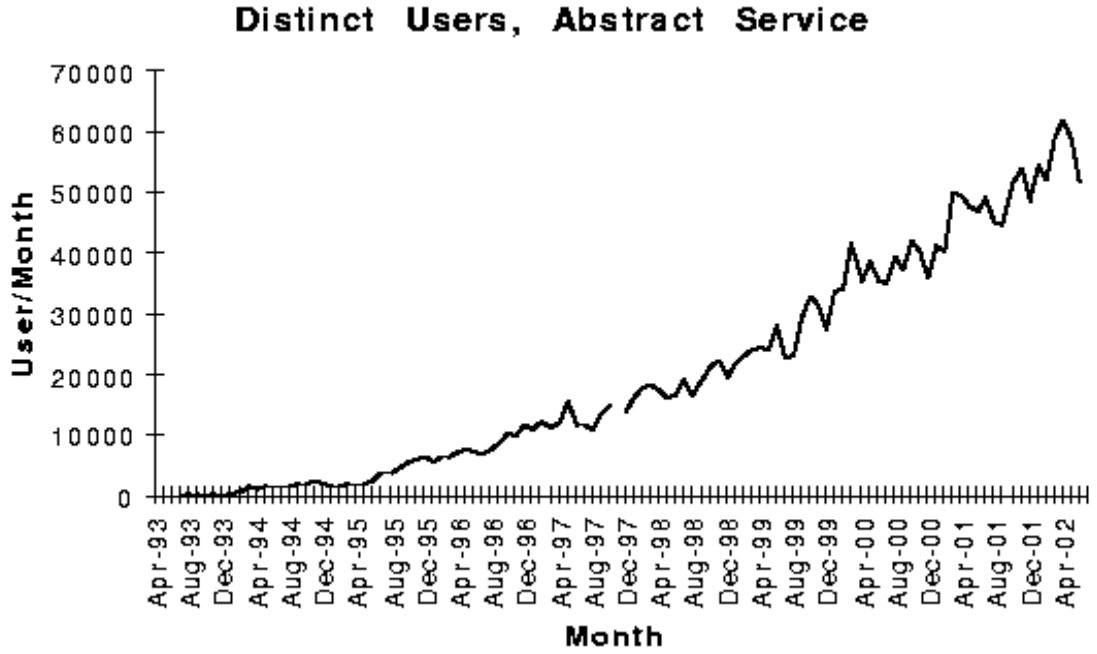

Figure 4: Number of references retrieved per month.

In a typical month the ADS was used by ~60,000 individuals, who made ~1 million queries, retrieved ~40 million bibliographic entries, read ~1 million abstracts and ~300,000 articles. Of the 300,000 full-text articles accessed through the ADS ~50% were via pointers to the electronic journals.

ADS users access and print (either to the screen, or to paper) more actual pages than are printed in the press runs of any single journal in astronomy. In May 2002, 1.4 million scanned pages were downloaded from the ADS archive. About 75% of these were sent directly to a printer, 22% were viewed on the computer screen, and 2% were downloaded into files; viewing thumbnail images make up the rest.

Table 2 shows the usage for the period of 1 Jan 2002 to 30 Jun 2002 for the different regions of the world.

6. Conclusion

The ADS provides access to most of the astronomical literature. The system includes an extensive system of links to relevant on-line information. It has profoundly changed the way astronomers do their research by allowing the scientists to easily locate information in the literature and associated on-line data. We hope that it will continue to facilitate astronomical research in particular in countries that do not have easy access to libraries with astronomical literature. It should also allow new studies of the historical literature that are so far very difficult or impossible. We welcome any questions and suggestions on how to improve the ADS services. Please contact us at **ads@cfa.harvard.edu.**

Region	**Nr. Users**	**Nr. Queries**	**Nr. Records**
North America	105785	2128611	54288848
Europe	57311	1859810	43744492
Asia (including Japan)	17018	480844	12858554
Central and South America	5641	145950	3231706
Pacific	4271	99815	2235661
Middle East	950	23484	694593
Africa	466	12151	311727

Table 2: Usage of the ADS from different regions of the world

Acknowledgment

Funding for this project has been provided by NASA under NASA Grant NCC5-189.

References

1. Boyce, P. 1996, AAS Meeting, 189, 0603
2. Kurtz, M., et al. 2000, A&AS, 143, 41
3. Eichhorn, G., et al 2000, A&AS, 143, 61
4. Accomazzi, A., et al. 2000, A&AS, 143, 85
5. Grant, C. S., et al. 2000, A&AS, 143, 111
6. Murray, S. S., et al. 1992, Astronomy from Large Databases II, 387

7. Egret, D, Wenger, M., & Dubois, P. 1991, Databases & On-line Data in Astronomy, Kluwer Acad. Publ., 79
8. Grant, C. S., Kurtz, M. J., & Eichhorn, G. 1994, AAS Meeting, 184, 2802
9. Eichhorn, G. 1994, Experimental Astronomy, 5, 205
10. World Wide Web Consortium 1999, http://www.w3.org
11. Schatz, B. R., Hardin, J. B. 1994, Science, 265, 895-901
12. Eichhorn, G. 1997, Astrophys. Space Sci., 247, 189
13. Eichhorn, G., et al. 1995, Vistas in Astronomy, 39, 217
14. Eichhorn, G., et al. 1995, ASP Conf. Ser. 77: ADASS IV, 28
15. Eichhorn, G., et al. 1994, AAS Meeting, 185, 4104
16. Boyce, P. B. 1995, AAS Meeting, 187, 3801
17. Eichhorn, G., et al. 1996, ASP Conf. Ser. 101: ADASS V, 569
18. Eichhorn, G., et al. 1997b, AAS Meeting, 191, 3502
19. Corbin, B. G. & Coletti, D. J. 1995, Vistas in Astronomy, 39, 161
20. Garfield, E., 1979, New York: John Wiley
21. Demleitner, M., et al. 1999, AAS Meeting, 195, 8209
22. Schmitz, M., et al. 1995, Info. & On-line Data in Astr., Kluwer, 259