Bull. Astr. Soc. India (2007) **35**, 717–725

Access to the literature and connection to on-line data

Guenther Eichhorn*†, Alberto Accomazzi, Carolyn S. Grant,
Michael J. Kurtz, Donna M. Thompson and Stephen S. Murray
Harvard-Smithsonian Center for Astrophysics, Cambridge, MA 02138, USA

Abstract. The Astrophysics Data System (ADS) provides access to the astronomical literature through the World Wide Web. It is a NASA funded project and access to all the ADS services is free to everybody world-wide. It can be accessed without login through any web browser.

The ADS Abstract Service allows the searching of three databases with abstracts in Astronomy (including Astrophysics, Planetary Sciences, and Solar Physics), Physics/Geosciences, and the arXiv E-prints from Cornell, with a total of over 5 million references. The system also provides access to reference and citation information, links to on-line data, electronic journal articles, and other on-line information.

The ADS Article Service contains the full articles for most of the astronomical literature back to volume 1. It contains the scanned pages of all the major Astronomy journals (Astrophysical Journal, Astronomical Journal, Astronomy & Astrophysics, Monthly Notices of the Royal Astronomical Society, and Solar Physics), as well as most smaller journals back to volume 1.

In order to improve access from different parts of the world, we have set up 12 mirror sites of the ADS in different countries in Europe, Asia, Australia and the Americas.

The ADS is available at: http://ads.harvard.edu

Keywords : digital libraries – data access – literature search system

*e-mail: Guenther.Eichhorn@Springer.com
†New Address : Springer, 101 Philip Drive, Norwell, MA 02061, USA

1. Introduction

The NASA Astrophysics Data System Abstract Service is by now a central facility of bibliographic research in astronomy. In a typical month (February 2007) it was used by more than 165,000 individuals, who made ∼3.4 million queries, retrieved ∼29 million bibliographic entries, read ∼5 million abstracts and ∼130,000 articles, and downloaded ∼1.3 million pages. The ADS is tightly interconnected with the major journals of astronomy, and the major data centres. A detailed description of the ADS has been published in a special issue of Astronomy & Astrophysics Supplements in April, 2000 (Overview: Kurtz et al. 2000; Search Engine and User Interface: Eichhorn et al. 2000; System Architecture: Accomazzi et al. 2000, Data: Grant et al. 2000).

The first major part 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 (Murray et al. 1992). By summer 1993 a connection had been made between the ADS and SIMBAD (Set of Identifications, Measurements and Bibliographies for Astronomical Data, Wenger et al. 2000) at the Centre des Données de Strasbourg (CDS), permitting users to combine natural language subject matter queries with astronomical object name queries (Grant, Kurtz, & Eichhorn 1994).

By early 1994 the World Wide Web was widely accessible through the NCSA Mosaic Web Browser. It now was possible to make the ADS Abstract Service available through a web forms interface; this was released in February 1994. The WWW interface to the ADS is described by Eichhorn et al. (1995).

The second major part of the ADS is the Article Service. It contains scanned full journal articles for most of the astronomical journal literature going back to volume 1 for most journals. The first full article bitmaps, which were of Astrophysical Journal Letters articles, were put on-line in December 1994 (Eichhorn et al. 1994). By now we have scanned all major and many smaller astronomical journals with a total of over 3 million pages.

With time, other interfaces to the abstracts and scanned articles were developed to provide other information providers the means to integrate ADS data into their system (Eichhorn et al. 1996).

2. Data

2.1 Abstracts

The abstracts in the ADS come from many different sources (see Grant et al. 2000). The original set came from the NASA STI database. We now receive basic bibliographic information (title, author, page number) from essentially every journal of astronomy. Most

publishers also send us abstracts, while some who cannot send abstracts, allow us to scan their journals. For these journals we build abstracts through optical character recognition (OCR). Finally we receive abstracts from the editors of conference proceedings, and from individual authors.

As of February, 2007 there are ∼1.25 million astronomy references indexed in the ADS, the database is basically complete for journals articles beginning in 1975. In the Physics database there are ∼3.3 million references, and in the arXiv E-print database there are ∼411,000 references. More than 2/3 of all references have abstracts, the others only have titles, authors, and journal information.

2.2 Bitmaps

The ADS has obtained permission to scan, and make freely available on-line, page images of the back issues of all the major journals and most smaller journals in astronomy. All these journals are scanned back to volume 1, page 1.

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 (see Grant et al. 2000). The files created are then automatically processed in order to de-skew and centre the text in each page, resize images to a standard U.S. Letter size (8.5 × 11 inches), and add a copyright notice at the bottom of each page. Adding the copyright notice on each page is important, since the ADS makes it very easy to reprint individual pages. Such individual pages would lose the information on where they came from and who owns the copyright for them. For each original scanned page, two separate image files of different resolutions are generated and stored on disk. The availability of different resolutions allows users the flexibility of downloading either high or medium quality documents, depending on the speed of their internet connection. So far we have scanned ∼3 million pages.

In order to improve access to historical observatory publications, we are collaborating with the Harvard libraries and our library. Our libraries have microfilmed all historical publication that are available here. We are in the process of scanning these microfilms and making the scans available through the ADS. We have currently scanned ∼300,000 pages and expect that to increase to ∼1 million.

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 (see Eichhorn et al. 2000). There are ∼15.5 million hyperlinks in the ADS, of which ∼30% are to sources external to the ADS project.

The largest number of external links are to SIMBAD, the NASA Extragalactic Data-

base (NED), the Space Telescope Science Institute (STScI), and the electronic journals. A rapidly growing number, although still small in comparison to the others, are to data tables created by the journals and maintained by the CDS. These links are an extremely important aspect of the ADS.

Table 1 shows the links that we currently provide when available. A more detailed description of resources in the ADS that these links point to is provided in Grant et al. (2000).

For astronomers it is important to access data that were used in an article, as well as the data published in the article. The ADS is provided with the information on what data are correlated with articles on a regular basis by several data centres and includes links from the articles to the on-line data.

There are two types of data that the ADS links to. One type are data in data services that aggregate and process information, the other are original data collected by telescopes or spacecraft.

Aggregated data are collected by several archives. The most important of these archive services are:
SIMBAD: Information about astronomical objects (Wenger et al. 2000)
Vizier: Data catalogs and data tables (Ochsenbein et al. 2000)
NASA Extragalactic Database: extragalactic objects (Madore et al. 1992).

The distribution of the links to the aggregating data archives is by now fully automated. There are on the order of 220,000 links to these services in the ADS.

The archives that hold original data also provide some information on what data were used in articles. This information has so far been collected by hand by the different data centres that hold the data.

Since collecting this information is a very time consuming task, the Astrophysics Datacentre Executive Committee (ADEC) initiated efforts to improve the linking between journal articles and on-line data. The data centres, the ADS and the UChP developed a system that allows authors to specify which data sets they have used for their article. The system was designed to allow for automatically processing and verifying data set identifiers specified by authors. Following is a brief description of this system.

2.3.1 *Data Set Identifiers*

The basis for this system is the identification of data sets. The data centres assign unique identifiers to each set of data. It is up to the data centre to decide what they call a data set. It could be one spectrum, or one exposure, or it could be a set of exposures of the

Table 1. Link types in the ADS database.

Link	Resource	# of Links	Description
A	Abstract	3.5 million	Full abstract of the article. These abstracts come from different sources.
C	Citations	1.9 million	A list of articles that cite the current article. This list is not necessarily complete (see 'R' References).
D	On-line Data	53,000	Links to on-line data at other data centres.
E	Electronic Article	3.0 million	Links to the on-line version of the article. These on-line versions are in HTML format for viewing on-screen, not for printing.[a]
F	Printable Article	2.1 million	Links to on-line articles in PDF or Postscript format for printing.[a]
G	Gif Images	500,000	Links to the images of scanned articles in the ADS Article Service.
H	HEP/SPIRES	385,000	Links to the High Energy Physics digital library SPIRES.
I	Author Comments	700	Links to author supplied additional information (e.g. corrections, additional references, links to data),
L	Library Link	n/a	OpenURL Links to the article (if the user has specified an OpenURL server.
M	Multi-media	1,400	Links to on-line multi-media information.
N	NED Objects	50,000	Access to lists of objects for the current article in the NED database.
O	Associated Articles	44,000	A list of articles that are associated with the current article. These can be errata or other articles in a series.
P	Planetary Data System	2,400	Links to datasets at the Planetary Data System.
R	References	1.8 million	A list of articles referred to in the current article. For older articles these lists are not necessarily complete, they contain only references to articles that are in the ADS database. For some articles that are on-line in electronic form, the 'R' link points to the on-line reference list, and therefore the complete list of references in that article.[a]
S	SIMBAD Objects	170,000	Access to lists of objects for the current article in the SIMBAD database.
T	Table of Contents	600,000	Links to the list of articles in a books or proceedings volume.
U	Also-Read Articles	1.7 million	Links to the list of articles that were read by the same people that read the current article.
X	arXiv e-prints	650,000	Links to the arXiv e-print version of an article.
Z	Custom format	n/a	Link to the abstract formatted according to the user's preferences.

[a]There is generally access control at the site that serves these on-line articles

same object in different wavelengths. This is left completely up to the data centres. In some cases, data sets are defined by the query parameters to a database query. Some

data centres also provide the means for authors to define a collection of data sets that they used in an article and give this collection a unique identifier. The main requirement for data set identifiers is that they have to be unique and permanent. This means that the data centres have to agree to recognize published identifiers in perpetuity. This is extremely important for the long term viability of this system.

The ADEC has agreed on a format for data set identifiers that is compatible with current International Virtual Observatory Alliance (IVOA, Quinn et al. 2004) designs for identifiers:

ADS/FacilityId#PrivateId

"ADS" specifies the ADS as the managing authority for these identifiers. Verification and linking is done through the ADS master verifier and link resolver. "FacilityId" specifies the facility that collected the data, and "PrivateId" is an identifier assigned by the data centre. The ADEC decided that the data centre should not be part of the identifiers, since data sets can potentially move between data centres, which would invalidate identifiers that contain the data centre.

The data centres provide the data set identifiers with each data set that they send to their users. These identifiers should be prominently visible so that authors can easily find them and include them in their manuscripts.

2.3.2 *Identifier Verification*

Once publishers receive manuscripts that contain data set identifiers, they verify that the identifiers are correct. During the verification process they obtain the permanent link for identifiers that are valid. This is done through the ADS Verifier. The verifier can be accessed through SOAP (Simple Object Access Protocol) or through a simple CGI interface. During copy-editing, the publisher sends the data set identifiers cited in the paper to the ADS master verifier for verification. The master verifier contacts the relevant local verifier at the data centre that currently has the data sets for the facility specified in the identifier. It then returns the status of the identifier as returned from the data centre verifier, and the permanent link to the data set if it is a valid identifier.

2.3.3 *Data Set Links*

The permanent links in the on-line journal article to data sets do not point directly to the data centre, but rather to the ADS link resolver. The reason for this is that data sets can move between data centres. The ADS is automatically kept up-to-date about the location of all data sets and forwards data set link requests to the data centre that currently holds the data. The link resolver consults the current data centre profiles to determine which

data centre currently holds the data for the specified FacilityId, and retrieves the current link to the specified data set from the data centre. It then forwards the request to that address. This assures that the links in the on-line journals are permanent and do not have to be changed if data sets move.

2.3.4 *Link Distribution*

In order to fully utilize the linking information, the ADS harvests the correlation between data set identifiers and articles from the participating publishers. This information is then used to link from the ADS records to the on-line data. The ADS also makes these correlations available to the participating data centres. We provide an HTTP interface for harvesting of these correlations by data centres. The data centres use this information to link from their data back to the journal articles.

3. Mirror sites

Soon after the inception of the article service in 1995 it became clear that for most ADS users the limiting factor when retrieving data from our computers was bandwidth rather than raw processing power. With the creation of the first mirror site hosted by the CDS in late 1996, users in different parts of the world started being able to select the most convenient database server when using the ADS services, making best use of bandwidth available to them. At the time of this writing, there are 12 mirror sites located on four different continents. Table 2 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. Such an abstract mirror site can now run on a Linux PC with 120 Gb of disk space. A partial article mirror site can run on as little as 200 Gb of disk space. If you are interested in having a mirror site, please contact the ADS@cfa.harward.edu for detailed requirements.

4. Conclusion

The ADS provides free access to most of the astronomical literature. It has profoundly changed the way astronomers do their research. 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 the ADS at: ads@cfa.harvard.edu.

Table 2. ADS Mirror sites.

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
Great Britain	University of Nottingham, Nottingham	http://ukads.nottingham.ac.uk
China	Beijing Astronomical Observatory, Beijing	http://baoads.bao.ac.cn
India	Inter-University 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
South Korea	Korea Astr. and Space Sci. Inst., Daejeon	http://ads.kasi.re.kr
Australia	Australian National University, Canberra	http://ads.grangenet.net
Indonesia	Indonesian Institute of Sciences, Jakarta	http://www.ads.lipi.go.id

5. Acknowledgment

Funding for this project has been provided by NASA under NASA Grant NNG06G-G68G.

References

Accomazzi, A., Eichhorn, G., Grant, C. S., Kurtz, M. J., & Murray, S. S., 2000, A&AS, 143, 85

Eichhorn, G., Kurtz, M. J., Accomazzi, A., Grant, C. S., & Murray, S. S., 1994, American Astronomical Society Meeting, 185, 4104

Eichhorn, G., Accomazzi, A., Grant, C. S., Kurtz, M. J., & Murray, S. S., 1995, Vistas in Astronomy, 39, 217

Eichhorn, G., Murray, S. S., Kurtz, M. J., Accomazzi, A., & Grant, C. S., 1995, ASP Conf. Ser. 77: Astronomical Data Analysis Software and Systems IV, 28

Eichhorn, G., Accomazzi, A., Grant, C. S., Kurtz, M. J., & Murray, S. S., 1996, ASP Conf. Ser. 101: Astronomical Data Analysis Software and Systems V, 569

Eichhorn, G., Kurtz, M.J., Accomazzi, A., Grant, C.S., & Murray, S.S., 2000, A&AS, 143, 61

Grant, C. S., Kurtz, M. J., & Eichhorn, G., 1994, American Astronomical Society Meeting, 184, 2802

Grant, C. S., Eichhorn, G., Accomazzi, A., Kurtz, M. J., & Murray, S. S., 2000, A&AS, 143, 111

Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C. S., & Murray, S. S., 2000, A&AS, 143, 41
Madore, B. F., Helou, G., Corwin, H. G., Jr., Schmitz, M., Wu, X., & Bennett, J., 1992, ASP Conf. Ser. 25: Astronomical Data Analysis Software and Systems I, 47
Murray, S. S., Brugel, E. W., Eichhorn, G., Farris, A., Good, J. C., Kurtz, M. J., Nousek, J. A., & Stoner, J. L., 1992, Astronomy from Large Databases II, 387
Ochsenbein, F., Bauer, P., & Marcout, J., 2000, A&AS, 143, 23
Quinn, P. J., Barnes, D. G., Csabai, I.; Cui, C., Genova, F., Hanisch, R., Kembhavi, A., Kim, S. C., Lawrence, A., Malkov, O., Ohishi, M., Pasian, F., Schade, D., & Voges, W. 2004, Proceedings of the SPIE, 5493, 137
Wenger, M., Ochsenbein, F., Egret, D., Dubois, P., Bonnarel, F., Borde, S., Genova, F., Jasniewicz, G., Laloë, S., Lesteven, S., & Monier, R., 2000, A&AS, 143, 9