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

Current and Future Holdings of the Historical Literature in the ADS

Guenther Eichhorn, Alberto Accomazzi, Carolyn S. Grant, Michael J. Kurtz, Stephen S. Murray

Harvard-Smithsonian Center for Astrophysics,
Cambridge, MA 02138, USA
gei@cfa.harvard.edu

Abstract.

The Astrophysics Data System (ADS) has the largest freely available archive of scanned literature in the world. One major part of this archive is to provide access to the full text of the historical astronomical literature. We have already on-line all major journals and many smaller journals back to volume 1. We currently have scanned over 30 journals as well as many conference proceedings.

An important part of this literature consists of observatory publications. The ADS is collaborating with the Wolbach Library at the Harvard-Smithsonian Center for Astrophysics in a project to microfilm these historical observatory publications. We are scanning these microfilms and making the scans available through the ADS. We currently have over 300,000 pages from over 30 observatory publications on-line.

In order to fully utilize these scans we need help in collecting metadata for these publications. We have developed an interactive tool to capture the metadata. If you are willing to help us with this task please contact the first author at gei@cfa.harvard.edu.

The ADS is funded by NASA Grant NCC5-189 and is available at:

`http://ads.harvard.edu`

1. Introduction

The NASA Astrophysics Data System Abstract Service is by now a central facility of bibliographic research in astronomy. In a typical month (May 2002) it is used by ∼60,000 individuals, who make ∼1 million queries, retrieve ∼40 million bibliographic entries, read ∼600,000 abstracts and ∼300,000 articles. The ADS is a key element in the emerging digital information resource for astronomy, which has been dubbed Urania (Boyce 1996). 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 (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 ADS started in 1992 as a distributed system that would allow access to NASA data at various data centers. In 1993 the abstract service was added to the ADS (Kurtz et al. 1993). The ADS was originally based on a custom-built networking software system (Murray et al. 1992). The user interface for the first version of the ADS Abstract Service was built with the proprietary software system that the ADS used at that time. The search engine of this first implementation used a commercial database system. A description of the system at that time is given by Eichhorn (1994).

By early 1994 the World Wide Web (WWW) had matured and was widely accessible through the NCSA Mosaic Web Browser. It now was possible to make the ADS Abstract Service available via a forms interface which was released in February 1994. Within five weeks of the initial WWW release use of the Abstract Service quadrupled (from 400 to 1600 users per month), and it has continued to rise ever since (Eichhorn 1997). At the same time a new search system was designed that was completely custom built. This allowed considerable optimization of the search software as well as easy implementation of advanced special search features. The WWW interface to the ADS is described by Eichhorn et al. 1995a and Eichhorn et al. 1995b.

The second major part of the current ADS is the Article Service. It contains scanned full journal articles for much of the astronomical journal literature going back to volume 1 for most journals. The first full-text images, which were of *Astrophysical Journal Letters* articles, were put on-line in December 1994 (Eichhorn et al. 1994). By the summer of 1995 the scans were current and complete going back ten years.

By now, 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. 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 to be fully on-line as published. 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.

We also are in the process of scanning old observatory reports and defunct journals, to finally have a full historical collection on-line. This work has begun with the scanning of several defunct journals (e.g. Astrophysica Norvegica) and with a collaboration with the Wolbach Library at the Harvard-Smithsonian Center for Astrophysics and the Harvard Preservation Project (Eichhorn et al. 1997; Corbin & Coletti 1995).

Currently there are over 2 million scanned pages on-line in the ADS in ∼280,000 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 (see Grant et al. 2000). The files created are then automatically processed in order to de-skew and center the text in each page, resize images to a standard U.S. Letter size (8.5 x 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.

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 basically in place. Currently the ADS system consists of four semi-autonomous (to the user) abstract services 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 (Eichhorn et al. 2000, Kurtz et al. 2000).

The following article will mainly describe the historical astronomical literature that is included in the ADS, how to access it, and the future plans for adding more of that literature.

2. Data

This section describes the data holdings of historical literature in the abstract and article service of the ADS.

2.1. Tables of Contents

The ADS covers the tables of contents (ToCs) of all major and most smaller journals back to volume 1. Considerable effort has been spent to generate ToCs for the oldest volumes of several journals that did not have printed ToCs. These ToCs provide the capability to search this historical literature by titles and authors.

In addition to journal ToCs we have collected ToCs from many conference proceedings, including all IAU Symposia, IAU Colloquia, all proceedings published by the Astronomical Society of the Pacific (ASP) and the Lunar and Planetary Institute (LPI) , and many other proceedings series.

Altogether, the ToCs in the ADS cover more than 95% of all articles published in the astronomical journal literature. We are currently working on covering more of the proceedings literature. A guess as to the coverage of the proceedings literature would be about 40-50%, possibly more. There are currently (as of July 2002) 814,000 records in the Astronomy database of the ADS.

2.2. Abstracts

The ADS Abstract Service started with a set of abstracts from the NASA STI (Scientific and Technical Information) project. These data covered a good part of the astronomical literature with abstracts from 1975 to 1995, but were by no means complete. For the journal literature, the coverage with abstracts in this range was probably better than 75%.

Since 1995 we have received abstracts from more and more journals in electronic form on a regular basis. By now these electronic abstracts cover over 99% of the articles published in astronomical journals.

The coverage of abstracts for the literature before 1975 however was basically zero. In 2000 we developed a system to perform Optical Character Recognition (OCR) on some of the scanned articles in order to extract the abstracts. This added about 30,000 abstracts to the database for articles published between 1944 and 1998.

Altogether, there are currently (as of July 2002) 388,000 abstracts in the Astronomy database of the ADS.

2.3. Full Text

The following gives a summary of the full text holdings.

Journals The scanned articles provide a very important archive that allows researchers world-wide access to the full articles of the historical literature. We have currently scanned 39 journals back to volume 1, as well as 29 conference series. This covers almost 60% of all articles published in astronomical journals before electronic versions were available. A further 15% are scheduled to be scanned.

Proceedings The second part of the scanned literature are conference proceedings. We have permission to scan several proceedings series, among them the IAU Symposia and the ASP Conference Proceedings series. The series which we have scanned cover a significant part of all conference proceedings. We have scanned ∼35% of the proceedings articles for which we have entries in the ADS, which is probably less than half of the total proceedings literature. We have scheduled for scanning another ∼20%.

Getting permission to scan conference proceedings is time consuming and sometimes impossible, since for any proceedings not published in a series we need to obtain permission from the copyright holder for each volume separately.

Observatory Publications Through all of the 19th century and well into the 20th century many, if not most, of the important articles in astronomy were published in observatory publications. This literature is only available in some of the larger and older libraries, and none of these libraries has all of that literature. The conservation project at the Wolbach Library at the Harvard-Smithsonian Center for Astrophysics and the Harvard Library is microfilming much of this literature. One copy of these microfilms is then scanned to produce bitmaps of the full texts. So far about 600,000 pages have been microfilmed. Of these, about 300,000 pages have been scanned and put in the ADS. The total number of pages from this project will probably be over 900,000.

The problem with these scans is that we do not have any metadata for these publications. We don't even know what image number corresponds to which page number. We have developed a tool that allows our users to enter page numbers and metadata for these scans from microfilms. So far about half of the 500 volumes scanned have been processed and assigned the necessary metadata. If you would be willing to help with this metadata capture program, please contact the ADS at ads@cfa.harvard.edu. Processing the page numbers for one volume takes about 1/2 to 1 hour. Entering the article information for a volume takes half a day to a day. We would welcome any help, even if you can process only a volume or two.

3. User Interface

The ADS services can be accessed through various interfaces (see Eichhorn et al. 2000). Some of these interfaces use WWW based forms; others allow direct access to the database and search system through Application Program Interfaces (APIs). There are several forms that allow access to the historical literature in the ADS. This section describes some of the interfaces that are relevant to the historical literature.

3.1. Main Query Form

The most commonly used interface is the regular search form of the ADS. Apart from searching by author, title, and abstract text, it also allows selecting specific journals for searching with specific date ranges.

3.2. Journal/Volume/Page Form

This form allows the user to retrieve records by specifying the journal name, volume, and page of a reference. This can be very useful for retrieving records from article reference lists, since such reference lists provide the information necessary for this form. In this form the journal can either be specified by its ADS abbreviation (as used in the ADS bibliographic codes), or by its full or partial name. If a partial name is specified that is not enough to uniquely identify the journal, the system will return a list of matching journals from which the correct one can be selected to complete the query. The same page also contains a form that allows a query by bibliographic code. This form accepts the wildcard character '?'. The '?' wildcard stands for one character in the code. For partial codes that are shorter than 19 characters, matching is done on the first part of the bibliographic codes. For instance:

1989ApJ...341?...1

will retrieve the articles on page 1 of the *ApJ* (*Astrophysical Journal*) and *ApJ Letters* volume 341, regardless of the author name (the *ApJ Letters* have an 'L' in the 14th place of the bibliographic code).

3.3. Scanned Articles Browse Forms

There are 3 forms that allow the selection of scanned articles, one for journals, one for observatory and society publications, and one for proceedings. Each of these forms has a list of the scanned series for selection and fields to specify the volume and page, as well as whether a regular page, a cover page, or a plate is requested. The form returns the scanned image of the requested page. These pages can be accessed from the ADS Browse page, which is linked from the main ADS page.

3.4. Historical Observatory Publications

The observatory publications that have been scanned from microfilms can be accessed from a separate page. This page can also be accessed from the ADS Browse page. It lists all volumes that have been scanned. Once a volume is processed for page numbering and article metadata, it is also included in the browse page for observatory and society publications (see previous section).

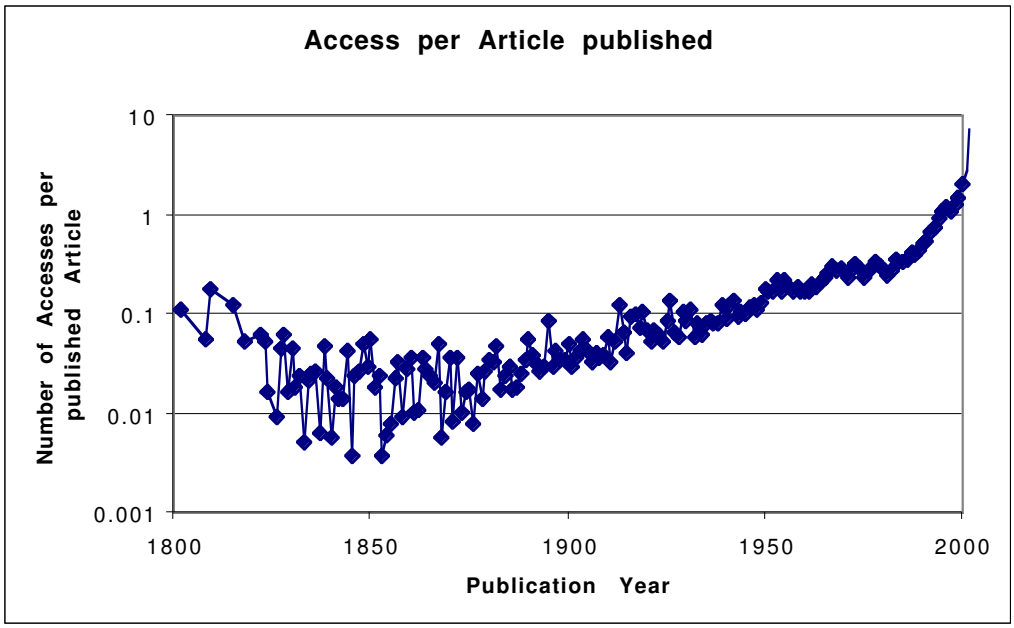

Figure 1. Access to the historical literature as a function of publication year.

4. Use of the Historical Literature

The ADS is used by most astronomers world-wide, as well as by many other interested users. More than 10,000 users use the ADS regularly (at least 10 times per month). Most of the abstracts and articles read are from the more recent literature, but the older literature is also accessed in significant numbers. Figure 1 shows the usage of the literature as a function of the publication year for the month of July, 2002. Usage is defined as a request for an abstract, an electronic article or a scanned article. It is normalized to the number of articles published per year.

The figure shows that the number of accesses per published article declines steeply with growing age up to about 50 year old articles. For articles between 50 and 100 years old, the accesses per article declines much more slowly. For articles older than 100 years, the accesses per article are essentially constant.

5. Future Plans

We have planned several additions to the data in the ADS. This section lists these major additions and how we expect them to be useful for ADS users.

5.1. Tables of Contents

We are currently in the process of entering more tables of contents in the ADS. We plan to eventually capture the tables of contents for all conference proceed-

ings and journals. Hopefully this will be done by early 2003. This will allow our users to search the complete astronomical literature by author and title.

5.2. Scanning of Preservation Microfilms

We plan to continue the scanning of the microfilms that are produced by the preservation project at the Center for Astrophysics and the Harvard library. We expect that eventually we will have more than 900,000 pages scanned from these microfilms. A major part of this effort will be the collection of the metadata for these publications to make them really usable. In order to collect the necessary metadata we will need the help from the astronomical and librarian communities.

5.3. Converting Images to Text

We plan to perform Optical Character Recognition (OCR) on all the scanned articles in the ADS. The OCR'd text will be used for several purposes.

Extracting Abstracts We have in the past used OCR to extract about 30,000 abstracts from scanned articles. We will do this for all of the scanned literature in the ADS. This should greatly enhance the abstract search system. Since many older journals and observatory publications did not have abstracts for their articles, there will be a limit as to how much we can do with this project. Even for articles with abstracts it is sometimes not possible for automatic procedures to locate these abstracts. We expect to be able to extract abstracts for most of the 20th century literature, but only for parts of the 19th century articles.

Extracting Reference Lists We have so far extracted about 3 million references from reference lists in our scanned articles. We plan to use the OCR'd text of all scanned articles to extract the reference lists and build a complete citation database for astronomy. As with abstracts, there will be limits as to what the automatic procedures can do, but we expect to extract a large number of references through this procedure.

Full Text Searching One major benefit of the OCR project will be the ability to index the complete text of all the scanned literature. We plan to develop a search system that will allow our users to search the full text. This will give our users capabilities that are currently unavailable. This will be useful for astronomers as well as historians. It will, for instance, let you find articles that use certain names, words, or phrases for the first time in the literature.

Full Text Distribution The OCR'd full text will **NOT** be available on-line. We will not be able to proof read 2 million pages of text, so there will be errors in the text. This is not a problem for search purposes, but it would be a big problem if somebody would use OCR'd data tables uncritically. The text will be made available only on special request when we can make sure that whoever receives the text is fully aware of the potential limitations and knows for what purposes the text is and is not useful.

6. 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. The ADS provides access to a part of the literature that is only available in the largest and oldest libraries in the world. It will allow users to work with this literature who have not been able to so far. With the implementation of the full text searching of the OCR'd text, it will 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

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

References

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

Boyce, P.B. 1995, Bull. AAS 27, 1333

Boyce, P.B. 1996, Bull. AAS 28, 1280

Corbin, B.G. & Coletti, D.J. 1995, Vistas in Astronomy, 39, 161

Eichhorn, G. 1994, Experimental Astronomy, 5, 205

Eichhorn, G., Kurtz, M.J., Accomazzi, A., Grant, C.S.,& Murray, S.S. 1995, Bull. AAS 26, 1371

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

Eichhorn, G. 1997, Ap&SS, 247, 189

Eichhorn, G., Kurtz, M.J., Accomazzi, A.,& Grant, C.S. 1997, Bull. AAS 29, 1262

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

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

Kurtz, M.J., Karakashian, T., Grant, C.S., Eichhorn, G., Murray, S.S., Watson, J.M., Ossorio, P.G., & Stoner, J.L. 1993, ASP Conf. Ser. 52: Astronomical Data Analysis Software and Systems II, 2, 132

Kurtz, M.J., Eichhorn, G., Accomazzi, A., Grant, C.S., & Murray, S.S. 2000, A&AS, 143, 41

Murray, S., Brugel, E., Eichhorn, G., Farris, A., Good, J., Kurtz, M., Nousek, J. & Stoner, J. 1992, Astronomy from Large Databases II, 387