Library and Information Services in Astronomy III
ASP Conference Series, Vol. 153, 1998
U. Grothkopf, H. Andernach, S. Stevens-Rayburn, and M. Gomez (eds.)

Incorporating Electronic Preprints into an Effective Publishing System

R. J. Hanisch, H. E. Payne, J. E. Huizinga, S. Stevens-Rayburn

Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA

E. N. Bouton

National Radio Astronomy Observatory, 520 Edgemont Road, Charlottesville, VA 22903, USA

G. Eichhorn

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

P. B. Boyce

American Astronomical Society, 2000 Florida Avenue NW, Suite 400, Washington, DC 20009, USA

Abstract.

Preprints continue to play an important role in the astronomical literature, both for rapid dissemination of new results and for establishing institutional benchmarks for quality and productivity. Electronic preprint services have become a popular alternative to paper preprint distribution. As centralized services, however, they ignore the importance of preprints in defining an organization's scientific profile. Moreover, the existing electronic preprint services are far from comprehensive in content (relying totally on author contributions), and there is no systematic tracking of the preprint into the refereed literature. We are now implementing a distributed electronic preprint service which will provide a common index to preprint databases located at separate astronomy institutions. Because maintenance of the preprint databases is distributed, our expectation is that the contents will be much more complete than is the case for the existing services. A key element of our approach is to assign unique identifiers to preprints as they are entered into the system, and to use these identifiers to track the preprint into the refereed literature. Once a preprint is published users of the preprint database will be directed to the published version, with the preprint version deleted from the system. We are also developing simple, portable tools for maintaining a local preprint database. Both these tools and the distributed preprint system infrastructure will be extensible to other "gray literature" documents, such as observatory and instrumentation manuals and technical reports.

1. Introduction

The age of electronic publishing is now upon us. The American Astronomical Society has launched the electronic *Astrophysical Journal Letters*, *Astrophysical Journal*, and *Astronomical Journal* (Boyce, 1996; Boyce and Biemesderfer, 1996; Boyce and Dalterio, 1996). The European journal *Astronomy and Astrophysics* is now on-line, and the European publisher Elsevier is now publishing the electronic journal *New Astronomy*. Agreements have been arranged between the publishers of the major astronomical journals to enable readers to follow reference links back and forth among the various journals—a collaboration known as *Urania* (http://www.aas.org/Urania/). Historically, rapid communications have been accomplished via private or institutional distribution of preprints. It is now the time to develop a complementary system of electronic information for the distribution of preprints and technical information, a system that can ultimately be fully integrated with the formal (electronic) publication of refereed papers.

The electronic journals will not eliminate the need for preprints. In the astronomy community peer review continues to play a key role in the formal publication process, but the preparation of papers for publication, including both quality "typesetting" and preparation of the electronic presentation format (including the addition of hypertext links, back- and forward-references), is time consuming. Authors desiring truly rapid dissemination of their research still want to issue preprints, in the sense of making their work available prior to its appearance in the refereed literature. Many manuscripts are already in the standard electronic formats accepted by the journals, and so are easy to publish as electronic preprints. The number of paper preprints being issued by astronomical research organizations has decreased slightly in recent years as authors have turned to the World Wide Web for distribution, but without the indexing facilities that have been provided for paper preprints and the electronic journals, many of these papers are difficult to locate. In order to take full advantage of the WWW, we require a truly distributed electronic document index and search facility, a system that is dynamically updated and simple for information providers to participate in.

The system of paper journals and preprints serves many functions in astronomy (Boyce, 1996):

1. *Information* — Provides a repository for the body of knowledge that comprises modern astronomy.
2. *History* — Provides an archive of the march of the progress of the science, and access to information about objects or phenomena whose importance may be realized only years after the initial observations.
3. *Author evaluation* — Provides a tool to judge competence, research effectiveness, and impact of authors.
4. *Status* — Determines where expertise resides, and who is working on what projects.
5. *News* — Disseminates the latest, but unverified, research results as rapidly as possible.

The first three items require the step of peer review to be effective, but preprints are valuable in fulfilling the last two functions. Our goal is to develop an electronic system in which both preprints and refereed publications can reside and that can fulfill all the functions now provided by the paper-based infrastructure.

2. The Role of Preprints and Technical Documentation in Astronomy

The rapid advances in astronomy and astrophysics are made possible not only by the availability of first-rate research facilities, but also by the fast dissemination of research results within the astronomical community via preprints. Some preprints are mailed directly to researchers in the field. Other preprints are mailed to the astronomical libraries around the world, where a much larger community regularly scans the "new preprints" racks. The printing and distribution of paper preprints is expensive, however, and more and more authors and institutions are turning to electronic methods of making preprints available, most notably on the Web (see Fig. 1). New preprints can then be announced by the authors to other researchers in the field, who can download the preprints through the Web or anonymous ftp. It is not so easy, however, to reach the much larger community of astronomers for whom the preprint may also be of interest. The AstroWeb Consortium (Jackson et al., 1995) provides the URLs of preprint collections[1] world-wide, but a time consuming search through dozens of Web pages to look for possible new preprints is cumbersome in comparison with a weekly visit to the new preprints rack in the departmental library. As a result, the current electronic publishing of preprints is far less effective in reaching a broad readership than the old paper distribution. The advantages and cost savings in making preprints available electronically cannot be fully realized without an efficient world-wide indexing and search facility.

There is also a large and growing body of technical documentation—user's manuals, instrument handbooks, etc.—available via the WWW servers of many astronomical research organizations. An astronomer seeking to find the most current user's manual for the HRCAM at the Canada-France-Hawaii telescope, or the preprints that have been released in the past week related to high-redshift galaxies, is faced with a tedious task of visiting several, or perhaps dozens, of WWW sites to find the relevant documents. Services such as the Astronomical Software and Documentation Service[2] (Hanisch et al., 1994, Payne et al., 1996) and reviews of network-based resources[3] (Andernach, Hanisch, and Murtagh, 1994) aid users in locating relevant Web sites, but have not addressed the problem of access to literature prior to publication in a comprehensive manner. The WWW has made a wealth of information available, but it has not necessarily made it easy to locate information.

---

[1] http://www.stsci.edu/astroweb/cat-preprint.html

[2] http://asds.stsci.edu/

[3] http://www.eso.org/gen-fac/pubs/nra/rrn.html

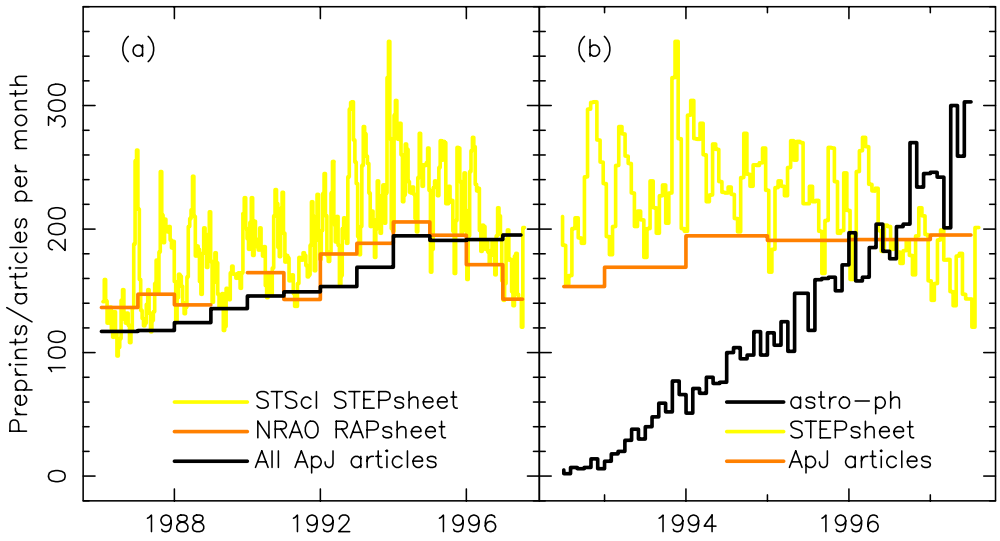

Figure 1. (a) Number of preprints indexed per month in the NRAO RAPsheet and STScI STEPsheet, compared to the number of articles published per month in the *Astrophysical Journal (ApJ)*. In recent years the number of preprints being distributed on paper has been decreasing. (b) Number of preprints appearing in the STEPsheet and the astro-ph preprint database at LANL, compared to the number of articles published in the *ApJ*. Electronic distribution of preprints has grown steadily and is displacing paper distribution.

3. Existing On-Line Preprint Services

There are now several on-line preprint repositories and indexes that have originated primarily in the field of high energy physics. These include the Los Alamos facility[4] developed by Paul Ginsparg (1994, 1996) and mirrored at SISSA in Trieste, Italy[5], and the SPIRES database[6] (Kreitz et al., 1996) maintained at the Stanford Linear Accelerator Center. The American Physical Society (APS) now has a preprint service[7] (Taubes, 1996). Los Alamos, SISSA, and APS are all centralized preprint repositories: authors submit preprints via e-mail or ftp for incorporation into the database. SPIRES is a preprint index, with links to those papers available from the Ginsparg database. Citations and links to published journals are added manually by the SLAC Library staff. Indexes of astronomy preprints include the National Radio Astronomy Observatory "RAPsheet"[8] (Radio Astronomy Preprints, though not limited to radio astronomy) and the

[4]http://xxx.lanl.gov

[5]http://babbage.sissa.it

[6]http://www-spires.slac.stanford.edu/find/hep

[7]http://publish.aps.org/eprint/

[8]http://www.nrao.edu/aoclib/rapsheet.html

Space Telescope Science Institute "STEPsheet"[9] (Space Telescope Exhibited Preprints), which are also maintained manually by their librarians. Recently the International Center for Theoretical Physics (Trieste, Italy) has developed a front-end search facility[10] for both the Los Alamos and its own archive, so that one query will search both databases, and CERN has a front-end[11] that allows users to issue one command to search up to ten different preprint databases. Users are warned, however, that selecting more than a few databases can lead to very long search times. These services are not true distributed databases, in that all they do is format the user's query properly for each preprint server and send off these queries sequentially. There is no integrated index, and searches will fail and relevant documents will be missed if a network host cannot be reached. The preprint lists maintained by NRAO and STScI are simple ASCII text files, with citations added as papers are published. Updated listings of new preprints are distributed electronically to several hundred astronomers and observatories bi-weekly and are posted in the network news. The complete databases going back to 1986 and 1982, respectively, are available on the Web and can be searched with simple text matching commands. Those preprints with URLs are provided with links in the database, but thus far these are a small fraction of the total.

4. A New Approach—A True Distributed Database

We are now developing a distributed index and search facility for astronomical preprints and technical documents. Our fundamental concept is to provide an integrated index for collections of preprints and technical documents located at, and maintained by, astronomical institutions distributed around the Internet. In this approach each organization maintains its own standards for preprints and other documents, and users may consider the organization's standards for release of an official preprint in their evaluation of the material. By leaving the documents themselves under the control of the provider, the user has the best assurance of their correctness and currency. By providing a common index, the user can easily locate documents anywhere on the Web with a single query to an integrated database. By applying the same relevance ranking to all documents, the centralized index assures that the most relevant documents in the entire collection are located, and not just the most relevant documents at each site. And by asking each institution to serve their own preprints, the workload is distributed among a number of institutions, avoiding both a single point failure mode and the need to maintain a large central archive of documents. Only the index needs to be integrated.

The key to this approach is maintaining a dynamic, integrated database of distributed documents. Since the database is not centralized, there can easily be multiple copies, and multiple entry points into the system. The database is integrated so that a search does not require sending a series of queries to a large number of independent Web servers—a process that would easily bog down in a

[9] http://sesame.stsci.edu/lib/stsci-preprint-db.html

[10] http://www.ictp.trieste.it/indexes/preprints.html

[11] http://preprints.cern.ch/search.html

system with a large number of nodes, any of which might be inaccessible on the net, or have slow response time, when the user happened to make a query. It is only when the user begins to examine the results of the database search that the original source documents or preprints are retrieved, via their URLs, from the Web.

Another key element of our project is to develop document data entry tools for preprint provider organizations and tracking tools to follow the preprint, and its revisions, to the final published form. In this regard we will provide links from the original preprint reference to the published literature as indexed by the Astrophysics Data System[12] (Eichhorn, 1994; Eichhorn et al., 1995; Accomazzi et al., 1996). The key to article referencing is the "bibcode," an alphanumeric string that encodes the journal name, volume, page reference, and author, as a unique identifier. In our system, once a preprint has been published the URL at the provider site would be replaced with a link, based on the bibcode, to the actual journal article. Links would also be provided to the article abstract and indexing services of the ADS.

As an added feature to the distributed preprint and document index, we plan to provide a basic notification service. This service will allow users to register an interest profile with the server. On a regular basis—daily, weekly—the database will be searched for new preprints or documents that match the user's profile, and e-mail messages will be sent to the user notifying them of availability of the new articles. This feature is probably the most important in terms of removing the need for paper preprints. Many authors insist on paper preprints, with direct mailings to libraries and colleagues, because such mailings function as a notification service. A notification service that is based on a distributed database of preprints and other documents will be vastly superior, however, in its completeness, timeliness, and directedness.

The major elements of the distributed electronic preprint service are shown in Fig. 2. A guiding principle of our software development strategy is that all tools are light-weight (easy to install and use, with minimal impact on a participating institution's computing facilities) and fully portable to common library computing platforms (Windows 95, Windows NT, Linux, Unix, etc.).

5. Status

The project is just six months into development, and to date we have focused efforts primarily on the document entry tool (see Fig. 3) and in evaluating various approaches to establishing and maintaining the distributed database. We are experimenting with the STScI and NRAO preprint collections as testbeds, and once we have implemented our initial distributed database we will be working with other libraries and organizations to start broadening the base of participating organizations. The project web page[13] provides additional information on

[12] http://adswww.harvard.edu/

[13] http://doright.stsci.edu/Epreps/

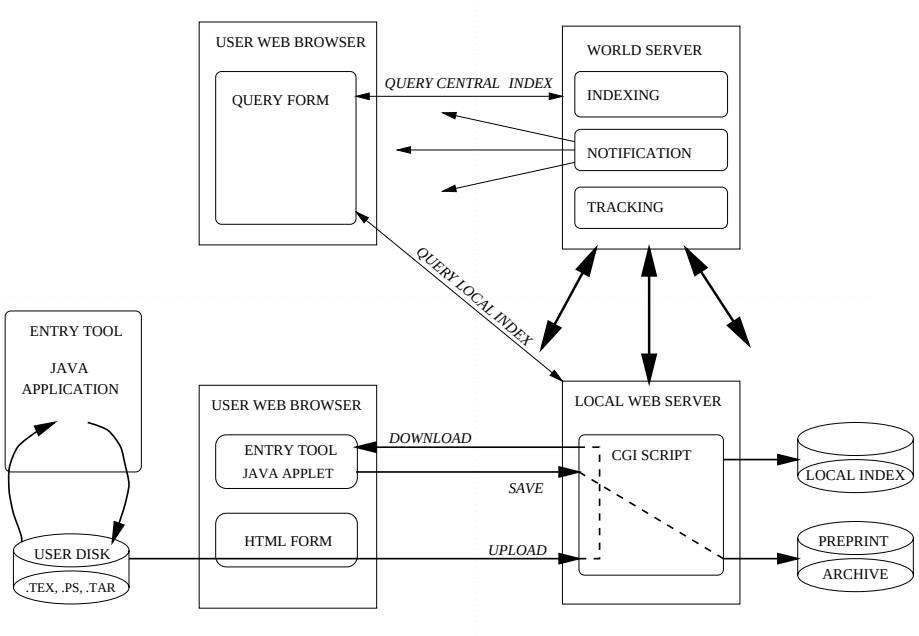

Figure 2. Overall architecture for the distributed electronic preprints system. The end user interacts with either the full index of distributed preprints or a specific local index (e.g., of preprints from just the home institution or any other particular institution participating in the distributed system). Local (institutional) servers maintain copies of the actual preprint documents. Preprints are entered into the system with an entry tool that automatically extracts as much metadata about the document as possible (title, authors, affiliations, abstract) from the source files.

the project, and the STScI preprint collection[14] is an example of how the search and browse facilities will probably look to the end user.

Acknowledgments. This work is supported by NASA's Applied Information Systems Research Program (AISRP) under grant NAG5–3942 to the Space Telescope Science Institute.

References

Accomazzi, A., Grant, C. S., Eichhorn, G., Kurtz, M. J., and Murray, S. S. 1996, in *Astronomical Data Analysis Software and Systems V*, G. J. Jacoby and J. Barnes, ed., ASP Conf. Ser. **101**, 558.

Andernach, H., Hanisch, R.J., and Murtagh, F. 1994, *PASP* **106**, 1190.

Boyce, P. B. 1996, *Computers in Physics* **10**, 216.

[14] http://www.stsci.edu/science/preprints/

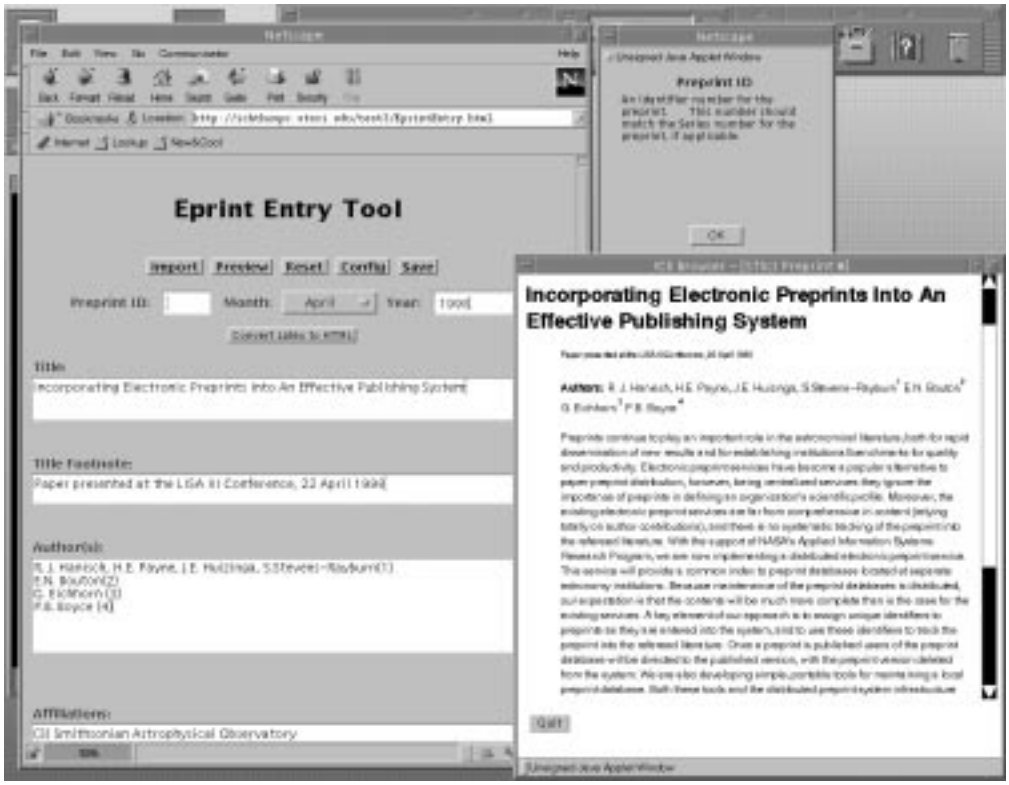

Figure 3. The prototype document entry tool. Metadata about the document (title, authors, abstract, etc.) are automatically extracted from tagged source files (LaTeX, AASTeX) or can be easily loaded into the forms with cut-and-paste. The metadata are written out as an HTML file for subsequent indexing, and for use in browsing the preprint database.

Boyce, P. B., and Biemesderfer, C. 1996, in *Astronomical Data Analysis Software and Systems V*, G.J. Jacoby and J. Barnes, ed., ASP Conf. Ser. **101**, 547.

Boyce, P. B., and Dalterio, H. 1996, *Physics Today* **49 (1)**, 42.

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

Eichhorn, G., Murray, S. S., Kurtz, M. J., Accomazzi, A., and Grant, C. S. 1995, in *Astronomical Data Analysis Software and Systems IV*, R. A. Shaw, H. E. Payne, and J. J. E. Hayes, ed., ASP Conf. Ser. **77**, 28.

Ginsparg, P. 1994, *Computers in Physics* **8**, 390.

Ginsparg, P. 1996, in *Electronic Publishing in Science*, Joint ICSU Press/UNESCO Conference, Paris, February 1996.

Hanisch, R. J., Payne, H. E., and Hayes, J. J. E. 1994, in *Astronomical Data Analysis Software and Systems III*, D. R. Crabtree, R. J. Hanisch, and J. Barnes, ed., ASP Conf. Ser. **61**, 41.

Jackson, R.E., Adorf, H.-M., Egret, D., Heck, A., Koekemoer, A., Murtagh, F., and Wells, D. 1995, in *Astronomical Data Analysis Software and Systems IV*, R.A. Shaw, H.E. Payne, and J.J.E. Hayes, ed., ASP Conf. Ser. **77**, 52.

Kreitz, P.A., Addis, L., Galic, H., and Johnson, T. 1996, SLAC–PUB–7110.

Payne, H. E., Hanisch, R. J., and Warnock, A. 1996, In *Astronomical Data Analysis Software and Systems V*, G. J. Jacoby and J. Barnes, ed., ASP Conf. Ser. **101**, 577.

Taubes, G. 1996, *Science* **273**, 19 July 1996, 304.