The Preprint Perplex in an Electronic Age

Ellen N. Bouton

National Radio Astronomy Observatory, Library, Charlottesvile, VA, USA[??]

Sarah Stevens-Rayburn

Space Telescope Science Institute, Library, Baltimore, MD, USA[??]

Abstract

We describe the trends in preprint distribution and the types of services available for access to preprints and preprint abstracts. We discuss development, maintenance, and access to bibliographic listings of preprints such as the RAP and STEPsheets, as well as more recent initiatives that offer full-text preprints via ftp or the World Wide Web. We also look at the merging of these developments in services such as the SLAC database and we discuss prospects, problems, and changes relating to an expected increase in both the electronic availability of preprints and the publication of journals in electronic forms. The focus of these considerations is the role of the astronomy librarian as an information provider in the electronic age.

1 Some History; or, What Got Us Started Down this Road?

In the mid-70s, as librarians at the National Radio Astronomy Observatory, we were responsible for tracking papers by NRAO staff and by visitors using NRAO telescopes in order to produce an annual bibliography and to provide additional information for the compilation of statistics on telescope use. At the same time, staff and visitors kept asking for preprints – by series, by author, by subject, and even by cover color – and, simultaneously, the number of preprints received had begun to exceed our personal memory capacities. In 1978, after consulting with other colleagues concerned with managing preprints, we put

[1] NRAO is operated by Associated Universities, Inc., under cooperative agreement with the National Science Foundation.

[2] ST ScI is operated by the Association of Universities for Research in Astronomy, Inc., for the National Aeronautics and Space Administration.

together a database modelled on the highly successful Preprint/Antipreprint lists produced by the Stanford Linear Accelerator Center.

The database was set up on an IBM mainframe with separate fields for:

- control number indicating year and biweekly period
- institutional abbreviation and preprint number
- author(s)
- title
- citation
- publication status

We got a librarian-friendly computer person to develop a bit of FORTRAN code that could be run with these data to produce separate lists sorted in a variety of ways.

Scientific staff greeted the lists we produced with enthusiasm. Thus was born the RAPsheet (**R**adio **A**stronomy **P**reprints), a biweekly list of new preprints received, and the unRAPsheet, an as-needed list of previously announced papers with their citations. We also made a weekly printout of the entire database sorted by author/title, which was a heavily used reference source in the Library; annual lists of preprints by institution used to identify preprints in series; and a list arranged alphabetically by author of all unpublished papers that we used in scanning the journals in search of citations. Online searching of the database was extremely cumbersome, so we did virtually none in the years the database was on the IBM mainframe.

In mid-1983, SS-R went to the Space Telescope Science Institute, carrying with her a tape of the database, and spent four months setting it up on a VAX and getting the software written to produce the lists, known there as the STEPsheet (**S**pace **TE**lescope **P**reprints). Unlike the IBM, the VAX allowed for online searching, so now not only were there biweekly lists, there was instant access from any computer terminal anywhere in the Institute to the online database.

NRAO was able to provide online access, plus improve database functionality as a result of two changes in 1985: the IBM mainframe was replaced with a VAX and the FORTRAN code was replaced with Inmagic, database management software designed for library and bibliographic functions. Inmagic allowed for the addition of fields (e.g., a code for NRAO telescope used). The VAX has since been replaced with workstations, servers, and PCs.

2 Mechanics

Incoming preprints are added to the database as received, and are then displayed until published. All astronomy journals, the astronomy sections of general science journals (*Nature*, *Science*, etc.), and meeting proceedings are checked against a list of unpublished preprints to find citations for papers in our databases. Other journals are scanned for astronomy and instrumentation papers and ST ScI runs a portion of the unpublished list against the *Current Contents* CD-ROM on a weekly basis, in order to pick up those citations in physics journals to which we don't subscribe. Newly published papers are then pulled from display, and citations, along with any needed corrections to the author/title information, are added to the database.

3 Distribution

Biweekly STEP and RAPsheets are offset by one week. We exchange them by email, allowing each of us to edit (and proofread!) the other's current list, and to append to our own a list of preprints received only at the other's institution. In 1988, at the time of LISA I, the NRAO distribution list for RAPsheets totalled 85 internal and outside addresses, all receiving paper copies. The NRAO library now produces and distributes biweekly RAP and unRAPsheets directly to almost 200 recipients, both individuals and libraries, about two-thirds of these electronically; the paper copies go primarily to those places where email is either unreliable or costly for long messages. ST ScI's distribution is all electronic, to some 100 addresses, several of which are exploders. NRAO also posts the electronic version to 3 usenet groups: a local NRAO one, sci.astro, and sci.astro.research. The full database in flat ASCII format, sorted by author and then by title, is updated biweekly and made available via anonymous ftp from an NRAO server. The subset of the ST ScI database that lists all papers based on observations with HST is also available via anonymous ftp from the ST ScI server.

In 1994, NRAO and ST ScI began producing WAIS-searchable versions of our databases, and have pointers to them on our WWW library home pages.

4 Annual Housekeeping

At the end of each year, published preprints received two or more years previously are offloaded into a separate file. Thus, in 1995, the current databases include all unpublished preprints, regardless of receipt date, and all published

papers (with citations) received as preprints in 1993-1995. The back file includes records for all published preprints received in 1986-1992. Note that, in the case of NRAO, both files are indexed and merged for the WAIS version, producing an index to all preprints received from 1986 forward. At ST ScI, back file dates to 1982, but the current couple of years are kept as a separate WAIS database, for those who are **sure** it was a recent paper. At both NRAO and ST ScI, we remove all unpublished preprints more than a few years old from our main display areas, although we do keep them in a separate area.

5 Database Sizes; or, How the Snowball Became an Avalanche

On 20 July 1988, the NRAO database contained 3207 preprints, of which 1602 were unpublished. The ST ScI database contained 3792 total and 1502 unpublished, including 1987A papers added whether preprints were received or not. On 10 April 1995, the NRAO database contained 6315 papers, of which 1604 were unpublished, and the ST ScI one had 7571 papers, with 2150 unpublished. It is interesting to note that of the 1502 unpublished papers in mid-1988, 192 are still without citation information.

6 Obstacles to Perfection; or What We Wish They Wouldn't Do

Problems that have remained with us from the start include:

- Preprints that arrive after the journal version has appeared in print: since many preprints do not say what issue of the journal they are going in, we have no way of knowing the item is no longer a **PRE**print. Further, when staff members give us stacks of preprints from their offices, we must check them all against the journals since we have no idea how long ago a preprint was distributed. This situation has been aided a lot by the availability of the A&A online indexes and the *Current Contents* CD-ROM, but having to pre-search suspect preprints adds to the time involved.
- The same preprint arriving from several different places: we can and do show multiple issuing institutions on a record, but unless we identify the preprint as a duplicate we end up with multiple records.
- Abstracts or poster papers that appear as preprints: we figure that if it's worth postage (or electronic bandwidth), it ought to actually contain some science and not be just "we thought about this and some day we'll write it up." Unfortunately, we have not yet convinced some of the astronomical community of this. Note that we do not include in this category paper abstracts of full length preprints available electronically. As we shall note later on, we believe those to be a viable alternative to full length paper

preprints. What we are opposing here is the "abstract as full paper" foolishness of many conference proceedings.
- Finding citations for the papers that appear in journals or proceedings we do not receive: the larger problem here is the inclusion of non-astronomy papers in an astronomy preprint series. Many of the older supposedly unpublished papers in our databases and on our shelves are really "physica esoterica" that are not particularly relevant to our users and therefore not in the sources we scan.

As you can imagine, the receipt, entering, and tracking of preprints, as well as the distribution of RAP/unRAPsheets and STEPsheets in a variety of forms, is very time-consuming. Keep in mind, however, that we would be doing most of this anyway in order to track NRAO and HST papers; the databases simply assist us in that process. In addition, whether in paper form, the online database form, the WAIS-searchable form, or the usenet form, the preprints and the preprint database are among the most heavily used part of our collections. Scientists recognize that the databases and the lists produced from it reflect the most current state of the most current literature in astronomy.

7 Where Do We Go from Here?

Up until quite recently, we were dealing with preprints that were actual physical objects. They came to us in envelopes in the post, or were hand-carried to us by staff scientists or visitors. Librarians have known for centuries how to handle print materials, and what we have done with preprints since 1978 is create some useful ways of organizing, tracking, and distributing information about these paper items.

In the last year we have seen a variety of preprint forms in addition to the traditional paper:

- preprints in electronic form only,
- electronic abstracts of preprints (i.e., SISSA),
- distribution of a paper preprint also available electronically (some paper copies list a URL or ftp site)
- distribution of a paper abstract listing a URL or ftp site for the full version

When given a choice (some institutions have asked) we have requested that our libraries remain on the distribution list for paper preprints. Staff and especially visitors at our institutions still like to browse through the paper versions and scan them on the shelves, and the paper also allows us to continue entering them in our databases in the traditional way.

We foresee an interim period in which the number of paper preprints will very gradually dwindle (although they probably won't disappear altogether for quite some time) and electronic distribution or access will increase. Indeed, there's already slight evidence of this: the first 16 weeks of 1995 saw an average of 103 new preprints per STEPsheet, versus 126 for the same period last year. It's still too early to tell if this is a timing variance or a real trend, but we're certainly keeping an eye on it.

In order to achieve the same level of accessibility for electronic preprints as the RAP/STEPsheets provide for paper ones, there must be an index to their electronic repositories. Right now, when a preprint (or an abstract) comes in with an ftp address or a URL on it, we enter that in the citation field in the database. At the moment, however, these entries are not hot links in our databases; merely a record of the access point, but we are working on how to make them links.

We think, in fact, that this idea of paper abstracts with URLs, pioneered almost simultaneously by the Princeton University Observatory and the Institut d'Astrophysique de Paris, is the next logical step in the astronomy preprint perplex. It solves several of the reasons we've heard for resistance to electronic preprints:

- Advertising – this sounds a little crass, but in fact, observatories care **a lot** about their preprints with their unique logos appearing on the shelves of all of the major observatories in the world.
- "Browse-ability" – as we mentioned earlier, being able to browse the paper copies is important, especially for those who spend most of their time in front of a computer monitor. They want to take a break from that, go somewhere else and **touch paper**.
- Illustrations not conducive to electronic transfer could be distributed with the abstract.
- Such abstracts with URLs fulfill one's exchange obligations (at least in our opinion).
- Abstracts provide something for the "Internots", those people and institutions for whom cruising the Internet is not a given.

In the current STEP/RAPsheets, URLs are replaced with the citations when the preprint is published. When we establish those URLs as hot links, the links will still be removed on publication, unless the journal itself is available electronically, in which case the URL will be switched to that of the journal. See for example:

- 94-19 IAS-94/45 BAHCALL, J.N.; KIRHAKOS, S.; SCHNEIDER, D.P. "HST images of nearby luminous quasars", ApJ 435: L11-L14, 1994

We need to keep firmly fixed in our minds in this discussion that we are dealing

with a **PRE**publication phenomenon. The preprint is not necessarily equivalent to the final product and to leave a version different from the published journal out on the 'net is unfair to the authors, the journal publishers, and the users. In viewing the posted preprints at one observatory recently, we were distressed to note that of the 50 listed, 21 of these had been published but the citations had not been added. For 13 of these, the PostScript version was still available; for one, the citation had been added, but the full text of the preprint was still available. At another observatory, there were a total of 21 preprints listed, 13 of which had been published and only one of which had a citation attached. Full text was available for all. The physicists have made a big point of not really caring about the final publication information, but astronomers have consistently said they disagree with this perspective. Yet their institutions are handling electronic preprints exactly as if the official published version had no meaning. We also note that leaving the full text version available after publication may be a violation of copyright law.

The belief out on the 'net seems to be that one can just toss preprints up locally and not worry about them until one needs the disk space. There are currently (mid-April 1995) 74 hits when one searches the AstroWeb listings for "preprint or preprints" and what is there is a hodgepodge of things, from lists to full text. Organization or even similarity between what places decide to mount simply doesn't exist. It's hard to imagine that an astronomer interested in the latest preprints would find the current diversity and lack of organization useful. It is somewhat ironic that when preprints became a popular medium for the exchange of information, a relatively standard format evolved quite quickly, but that with electronic preprints, just the opposite seems to be occurring – observatories and departments are working hard to make sure their preprint page **doesn't** look like anyone else's.

To summarize then, where we foresee going from here, **just for us and the STEP/RAPsheets**, is that we will continue to produce our lists, adding new preprints as received and citations as discovered. In addition, we are exploring ways of coordinating our efforts more closely. For instance, the ST ScI database now includes **all** of the unpublished papers in the NRAO database, so that, at least in theory, we could trade off checking for preprints in specific publications, rather than both of us doing them all. We hope to add hot links for preprints that are available electronically, but likely only to the first one received; that is, if multiple places mount the same preprint, we'll only put in one link. As the lead times between preprint and publication converge, the need for services such as ours will gradually decline, especially if the next part of our vision comes to fruition.

What we would like to see, outside of what we are doing, is the development of a central index that could simultaneously search diverse databases, both ours and others, including any specific institution's electronic preprints, as well as

any journals that are available electronically. The search would return a list of one-line hot links that match the search criteria from which one would choose where to go next. We believe that the hit list should be hierarchically ranked, so that the electronic journal entry, if one exists, would appear first, followed by the RAP/STEPsheet entry (especially if a citation had been added), followed by the local electronic preprint copy if there is one, followed by the local abstract if that's all that existed. The STEP/RAPsheets would, of course, have links to these latter, but would appear higher in the hierarchy because the publication information would be deemed more reliable. Distinguishing among the "class" of hits could be accomplished through color coding of the links: green for published, refereed papers; yellow for RAP/STEPsheet entries; red for preprints. (You or the system designers may certainly choose other colors, but we do like the idea of using a known hierarchy of color codes in this context).

We also believe that both the consistency and reliability of citation information are very important to a successful system. One of the reasons the RAP/STEPsheets have been so useful is that they provide reliable citation information in a consistent format. Further, given the speed at which things happen in the electronic world, it is important that the citations appear on the records in a timely manner. Consistency, reliability, and speed all imply the need for a centralized entry point for the information. The tour of existing astronomy preprint pages mentioned previously indicates that institutions do not follow through quickly (or at all) with correct citation information, and in our experience, the individual astronomer could **never** be trusted to do so.

Finally, we would like to reiterate the point with which we began – we started the RAP/STEPsheets in order to better serve our clientele: we want to connect the researcher with the needed information in the most efficient manner possible. When/if the day arrives that our preprint systems are obsolete, we will be delighted to close down our databases and concentrate our efforts on the next generation of information needs.