Preprints in Scholarly Communications:

Lessons from High Energy Physics

Paige C. Lucas-Stannard

Information Science

Dr. Froelich

Fall 2003

Introduction

Preprints have become an accepted and valued addition to the communication infrastructure of some fields while other fields have stayed wary of the preprint. Those fields that rely on preprints note the speed of delivery and access as well as the continued control of the author to their work as advantages of the preprint model. Lack of peer-review, editorial control, and archiving are some of the disadvantages noted by detractors. A third group, publishers of traditional journal literature, also bemoan what they see as a loss of control on the publishing process (and its revenues). The grandfather of all preprint communities, High Energy Particle Physics (HEP), has not only been successfully using but *relying* on preprints as the major form of information dissemination for more than three decades. The success of HEP preprints can be used as a model for other disciplines. This paper will look at the history of HEP preprints, some of the issues that arise in the preprint environment and present some examples of models that are working to make preprints a viable information source for a wider community.

Defining 'Preprint'

Preprint has been variously defined. Dallman, Draper, and Schwartz (1994) defined a preprint as a "manuscript ready to be submitted to a conference or journal." Traditionally, a preprint was a paper copy of an article mailed to colleagues concurrently with submission to a traditional journal. This, along with personal correspondence and dialogue at conferences, has been a major

component of information exchange in HEP. This type of informal and "pre-publication" discourse helps scientists shape their own research as well as build collaborations that play an important part in the field (Goldschmidt-Clermont, 1965). As technology advanced email and electronic bulletin boards (EBB) became a faster way to let peers know about research findings, in some cases months before they would be printed in a journal. EBB's became databases with advent of File Transfer Protocol (FTP) and the Internet and later the World Wide Web. Preprints, now available in a wide variety of electronic formats are sometimes called *e-prints*. E-prints can be used for documents outside of the traditional preprint. The U.S. Department of Energy (DOE) defined e-prints as "scientific or technical documents circulated electronically to facilitate peer exchange and scientific advancement. Included are pre-publication drafts of journal articles (preprints), scholarly papers, technical communications, or similar documents relaying research results among peer groups (http://www.osti.gov/eprints/)." Thus, today's preprints are e-prints, though all e-prints are not preprints.

Others disagree with the idea of a preprint as a "pre-publication" document. Paul Berman (1994), a lawyer for Covington and Burling, speaking at a panel discussion on intellectual property issues at the American Physical Society E-print Workshop stated that he "has very little doubt that putting something on an EBB…making something widely available to members of the scientific community essentially on a for free or readily accessible basis, almost certainly constitutes publication. There is very little doubt about that." This

concept of a preprint as a published document has important ramifications for future publication in a traditional journal that will be discussed later. In any case, many preprints are eventually published in journals, at which time they are referred to by some as *antiprints*. For the purposes of this paper, preprints will refer to author-archived documents in an electronic format that are available to the public at no cost.

Why HEP?

There are a number of factors that exist in HEP that made it the perfect breeding ground for new information dissemination channels. First, it is important to look at the environment in which HEP physicists are doing their work. According to the NSF survey (2001), 39% of all physicists work for private industry employers while 39% and 23% work for universities or government respectively. The number of HEP physicists that work for private industry is lower than the average due to the highly theoretical nature of the research. This leaves a majority of the work in the public realm (Universities and Government), which has an impact on information access. Creator of the first preprint server at Los Alamos National Laboratory, Paul Ginsparg (2003), pointed out that the findings of publicly funded research should be "freely available as a public good." Ann Okersen, a librarian at the Association of Research Libraries also speaking at the E-print Workshop, mentioned that the copyright act explicitly states, "Works written by federal government employees during working hours become works in

the public domain (section 105)." In disciplines with a smaller percentage of government employees this would not be as important.

Another important difference in HEP research is, although experimentation is expensive, there is not a large commercial application of research results. This differs from some of the biological and chemical sciences where patents are necessary to secure the economic rights of a discovery (Hurd, 2000; Warr, 2003). Patent rights can also make research competitive and secretive. HEP, on the other hand, is largely collaborative with large teams of researchers working on different continents. During the time it takes to set up an experiment, it is important for HEP physicists to know what other work is being done that could parallel their work or render it of no value. Collaboration has always been an important aspect of HEP research and it flourishes in a system with plentiful means of quick communication.

It has been suggested that HEP physicists are compulsive communicators (O'Connell, 2002), which is evidenced in their leading the way with each new communications technology. HEP has been at the leading edge of technology through user-driven solutions. For example, Tim Berners-Lee, a HEP physicist at The European Center for Particle Physics (CERN) created the first web page in 1990 and Paul Kunz, a HEP physicist at Stanford Linear Accelerator Center (SLAC) brought Berners-Lee's idea to the United States, creating the World Wide Web (arXiv, 2003; CERN, 2003). The quantity of information also led to early developments in cataloging (and later databasing) preprints. In 1962 SLAC began archiving pre-prints. These pre-prints were cataloged with basic

bibliographic information and a weekly publication was sent to subscribers listing new pre-prints. This set the stage for Paul Ginsparg's first preprint archive in 1991 (arXiv.org, formerly xxx.lanl.gov).

Today's HEP Preprint

There are a variety of preprint archives operating in the HEP field but by far the most important is ArXiv.org. This is the original preprint server and began in 1991 thanks to the confluence of a variety of factors. As mentioned above, the World Wide Web had been introduced the previous year, along with increased bandwidth and wider availability of computers, scientists were ready for quick access. The final technology that made ArXiv work was a simple way for submissions to be made directly from the researcher. This came in the form of a computer program created in 1977 by Donald Knuth and improved upon in the mid-1980s by Leslie Lamport called LaTeX (pronounce lah-tech). This program used a simple markup-style language to typeset a document. It is particularly useful for mathematical equations (see example below) and can be interpreted by any computer with the open source code.

Example of LaTeX:

$$H(s) = \int_0^\infty e^{-st} h(t)\, dt$$

```
<math>
H(s) = ∫<sub>0</sub><sup>∞</sup> e<sup>-st</sup> h(t) dt
</math>
```

in LaTeX:

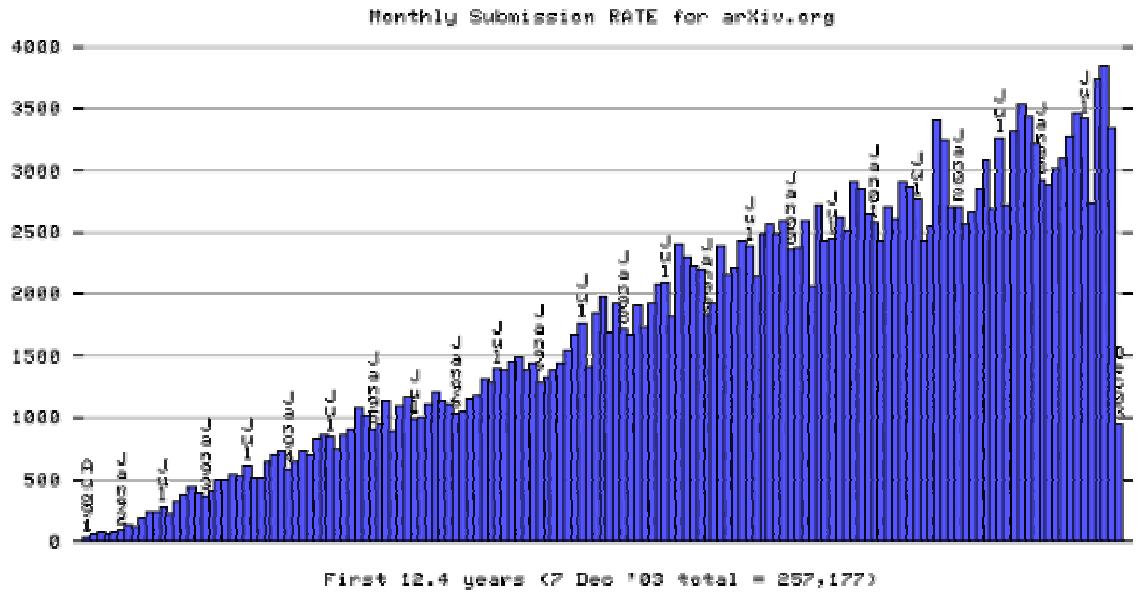
Cartesian closed categories and the price of eggs
Jane Doe
September 1994

Hello world!

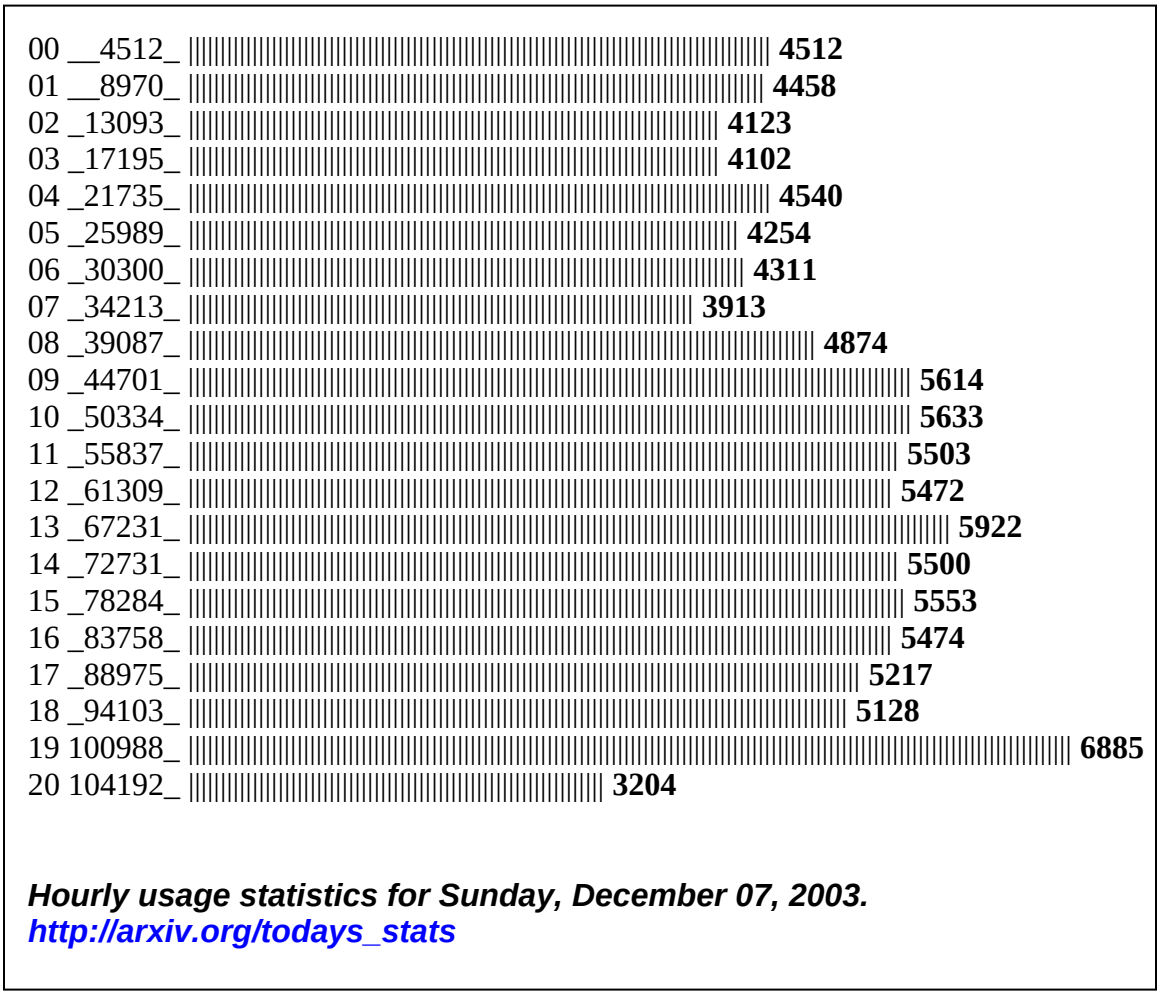
```
\documentclass{article}
\title{Cartesian closed categories and the price of eggs}
\author{Jane Doe}
\date{September 1994}
\begin{document}
  \maketitle
  Hello world!
\end{document}
```

Since LaTeX designates the layout of a document, submitters to ArXiv could email or FTP (and now, more commonly through a web download) their documents and they automatically appear in the database. Documents today are available the same day they are submitted by the scientist, sometimes nine months before the research will appear in a journal.

ArXiv now houses preprints in 12 physics fields, non-linear science, mathematics, computer science, and, starting in September of 2003, quantitative biology. The following graph from http://arxiv.org/show_monthly_submissions shows the increase in the number of submissions to the ArXive in its 12 year history.



Statistics on usage are even more impressive and are available on a daily and hourly basis. The following graph represents the number of users accessing the database for each hour on Sunday, December 07, 2003. Weekday usage is even higher (for an archive of usage see http://arxiv.org/show_stats).



According to O'Connell (2002), 70% of pre-prints submitted to ArXiv are eventually published in journals and another 20% are published in conference proceedings. HEP has long ago reached what is called a critical mass, where as the number of submissions increase, the likelihood of scientists submitting their work also increases – acceptance through consensus. Ninety-five percent of all HEP literature published is available at ArXiv. When articles become antiprints a note is added to the database (author submitted) as well as any revisions. ArXiv

is now housed at Cornell University and partially funded by the National Science Foundation. In addition there are 16 mirror sites in countries around the world.

Emerging Preprints in Other Fields

The success of preprints has not gone unnoticed by researchers in other fields and there are now a wide variety of preprint servers for various fields. For example, CogPrints (http://cogprints.soton.ac.uk) allows self-archiving of papers in psychology, anthropology, philosophy and linguistics. Other fields have recognized the importance of open access and now have archives of journals, for example PubMed Central (http://pubmedcentral.nih.gov/) is the National Library of Medicine's journal archive (although no self-archiving) and GenBank (http://www.ncbi.nlm.nih.gov/Genbank/index.html) is a successful repository for gene data submitted directly by researchers. A number of institutional repositories also allow for preprint publication. For example the University of California's E-Scholarship Repository (http://repositories.cdlib.org/escholarship/), MIT's dBase project (http://www.dspace.org/), and CalTech's Collection of Digital Archives (CODA, http://library.caltech.edu/digital/). There are also a number of initiatives to encourage institutional repositories and archiving (http://www.arl.org/sparc/) and open source software to implement digital archives (http://www.eprints.org/, http://www.dspace.org/).

Problems with Preprints

It is clear that preprints are here to stay in a growing number of fields. However, there are concerns looming even for HEP preprints. The remainder of this paper will look at some of the issues and possible solutions.

Acceptance

Some researchers worry that "publishing" to a preprint archive will never become acceptable in the field. One of the ways that acceptance can be measured is with citation analysis. Very little research has been done to study the citation of preprints in the sciences. Celia Brown (2003) looked at the citation of preprints from the Chemistry Preprint Server (CPS, http://www.chemweb.com/preprint) and found no instances of preprints being cited. The CPS contained only 217 preprints at the time of the study. In another study by Brown (2001) on the citation of e-prints from ArXiv in Physics and Astronomy journals she found more promising results. From 1991 to 1999 there were a total of 35,928 references to ArXiv preprints for a total of 34.1% citation rate. The rate of citation was highest in the HEP fields (phenomenology and theory), condensed matter, and astrophysics. Of the 22,824 preprints in the HEP-phenomenology database, 38.1% were cited in the literature.

Why the disparity? One of the reasons relate to the idea of critical mass. The ArXiv databases have been around for more than ten years and contain a large amount of literature. The sheer amount of documents is hard to ignore as a

resource in research. ArXiv is also indexed by the Chemical Abstract Service which increases exposure. Brown concludes that Chemistry is likely to change over time and eventually reach a level with ArXiv. Other studies of general grey literature indicate a growing acceptance of "alternative forms of publication (Pelzer, 2003)."

Another pressing issue to researchers is receiving proper credit and prestige and, more importantly, tenure and promotion. Cronin, McKenzie, Rubio, and Weaver-Wozniak stated "Scholarly publishing is as much about rewards as texts…a means for allocating credits and auditing accomplishments, as in the case of academic promotion and tenure (Cronin, et al., 1993)." Steve Heller says "You don't get tenure at Harvard by publishing in the Internet Journal of Chemistry (as cited in Warr, 2003)." The established system of merit based on publication in traditional, and often specific, journals deters some researchers from seeking out alternative publication sources. Concerns are often related to peer review, which will be discussed in the next section.

Citation counts are a major factor in tenure and promotion (Cronin, et al., 2003) and Lawrence (2001) found a correlation between the number of times an article was cited and its availability online. It would seem that in those fields where preprints are being cited, and these citations are showing up in ISI Citation Index, an effect would be felt on how often an author is cited. Database searches[1] yielded no research that has been done on the role that preprints play in tenure committees in those fields heavily using preprints. Furthermore, with

[1] EBSCO, Library Literature, ERIC, Education Abstracts, INSPEC, and Citation Index were searched for preprint or pre-print or eprint or e-print and tenure or promotion. Some work with alternative measures of citation in tenure review are discussed by Cronin, et al., 2003.

the current model in HEP, where 90% of all preprints are subsequently published in journals, traditional means of assessing merit are still intact. The question for the future is: Will tenure committees relax their rules regarding publication, causing researchers to utilize preprints more or will more and more researchers use preprints causing tenure committees to develop a more holistic view of merit? In this area it is clear that more research needs to be done.

Peer Review

Those wary of preprints often site the lack of peer review and other value-added editorial processes that traditional publishers have normally taken the burden for. The truth is that anyone can submit a document to a preprint server (although, some have to originate from a campus IP address), the worry is: will preprint servers become the "Journal of Not Very Good Science" (Warr, 2003)? The first instance usually cited is the University of Utah's premature announcement concerning Cold Fusion. The preemptive strike by scientists, made in order to beat a competing scientist, led to a flurry of research at other universities and a considerable expenditure of funds, only to find that the theory had little merit[2]. Although this instance did not involve a preprint (the scientists held a press conference) it has become the omen of what can happen without peer review.

Peer review is designed in theory to authenticate a document. If published in a journal it is *assumed* by readers to have correct data and

[2] The researchers at The University of Utah may have jumped the gun; however, current studies are bringing Cold Fusion back into the research arena. See Goldstein, 1994; Davis, 2003.

plagiarism controls (Warr, 2003). However, as Ginsparg (2003) stated, the reality is that publishers guarantee only the basic information: that the author is who they say they are, the article is not "obviously wrong or incomplete", and that it is of interest to the readership. The *British Medical Journal* conducted studies of the peer review process by sending out an article with 8 known errors. The median number of errors detected by 221 reviewers was 2 and none of the reviewers spotted more than 5 (Smith, 1997). The point here is not that the review process is bad, it does improve a paper (Warr, 2003), but most readers overemphasize its role. Ginsparg (NPR, 1996) remarked that discussion group interactions concerning preprints often leads to the reworking of the paper and resubmission with acknowledgements to those who participated in this informal review. Ginsparg (2003) also pointed out that expert readers "don't value the extra level of filtering above their preference for instant availability of material." Even Peter Boyce (n.d.), who tends to be more critical of the lack of peer review, said that "peer pressure from colleagues does seem to keep the quality of the submissions higher than might have been anticipated." Warr (2003) noted that concerns of correctness are stronger in medical fields and this has led the *New England Journal of Medicine* to rule against accepting preprint submissions.

Copyright

Copyright is another issue in the preprint world that has many faces. First is the idea of previous publication. As mentioned, submission to an archive can be considered publication and some journal publishers have restrictions against

submission of previously published work (Pinfield, 2001). Clinical Medicine publishes a list of medical journals that will and will not accept articles that have been previously submitted to a preprint server (http://clinmed.netprints.org/misc/policies.shtml). A UK based organization; The RoMEO Project (Rights MEtadata for Open archiving, http://www.lboro.ac.uk/departments/ls/disresearch/romeo/index.html) also lists allowance of preprint and post-print (antiprint) by publishers. Overall, RoMEO finds that 35.7% of publishers allowed preprint submission, 16.9% allowed both pre- and post-print submissions, and 45.3% do not support self-archiving of articles. Those that do allow self-archiving often have stipulation, as the Institute of Physics states that pre- and post-print submission is allowed given that, "access to such [preprint] servers is not for commercial use and does not depend on payment for access, subscription or membership fees (Institute of Physics Assignment of copyright form, http://www.iop.org/EJ/authors/)."

Harter and Park (2000) studied the impact of prior electronic publication on manuscript policies of publishers. They included as "electronic publication" items that were posted to a listserv, on a personal homepage, in a preprint collection, and published in an electronic proceedings or electronic journal. Their results were interesting and the "yes" for acceptance rates are listed in order below.

1. Author's homepage, 77.9%
2. Preprint, 75.2%
3. Listserv, 73.5%
4. Institutional homepage, 66.4%
5. Electronic conference proceedings, 52.2%
6. Electronic Journal, 25.7%

It is interesting to note how high preprint falls on this list. Publishers in the study were more likely to accept preprint submissions than listserv postings and articles on institutional websites. Not surprisingly they found that field of study was the most significant factor in the editorial decision; arts and humanities journals were least likely to accept preprint submissions while physics journals were most likely.

Preprint severs differ in their treatment of copyright. Some limit their database to citations. Others remove full-text, leaving only citation and abstract after publication (Topology Atlas Preprints, http://at.yorku.ca/topology/preprint.htm). CogPrints and PubMed Central divide preprints into two categories: unrefereed preprints and "author authenticated reprints of refereed, accepted papers (Tomaiuolo and Packer, 2000)." And while some limit to preprint postings only, IEEE requires that the final published version appear on the preprint server with proper copyright notice (Ibid.).

Copyright at ArXiv is maintained by the author and ArXiv makes no claims to it or its authenticity. In a recent incident, a submission to ArXiv contained potentially libelous remarks about another scientist. David Mascord, a media-law specialist, said that postings at ArXiv are published documents and subject to libel law. And, although libel claims are difficult to defend in the United States, the ArXiv is subject to libel laws in any country where it can be accessed (Giles, 2003). This is unlikely to happen (the scientists has modified his paper) but it does open the door for future problems in preprint archiving.

Plagiarism

A potential problem with self-archiving is plagiarism. Wills and Wills (1996) hold that "there will surely be occasional instance of [plagiarism] but the benefit of the additional feedback from a body of other interested authors which would not normally be available should more than compensate for such a risk." Once again peer pressure asserts control. In a recent incident (Giles, 2003), users noted that a paper on ArXiv copied parts of the *BaBar Physics Book*. When 6 others instances of plagiarism were found in the author's other works, all 22 articles were removed from the server by the submitter. It turned out that the author had enlisted the help of a colleague to submit the papers. The submitter posted an apology to the ArXiv. A search of Stanford's SPIRES database reveals 33 documents that had a claim of plagiarism (22 of which are from the case mentioned above). All were removed by the submitter or database administration. It is also interesting that 6 of the plagiarized articles were also published in traditional journals. The citations remain for these articles (see below), so the negative effect of peer pressure is once again an inhibiting force in misuse of preprints.

Example of SPIRES entry for plagiarized work:

2) **THEORY OF INCLUSIVE DECAYS OF HEAVY HADRON.**
By Ramy Naboulsi. Apr 2003. 22pp.
Withdrawn from arXiv due to plagiarism of SLAC-R-504.
e-Print Archive: **hep-ph/0304044**

LaTeX(US) | LaTeX(EU) | Harvmac | BibTeX | Citation Search

Conclusion

Journals are a major burden on library budgets. Recently Cornell University announced the cancellation of 200 Elsevier titles in order to accommodate diminishing budgets (http://www.library.cornell.edu/scholarlycomm/problem.html). Scholars are also feeling the weight of traditional publishing. Buck, Flagan, and Coles of the California Institute of Technology (1999) stated, " It is becoming increasingly clear to the scholarly community that we must envision and develop for ourselves a new, affordable model for disseminating and preserving results, that synthesizes digital technology and the ongoing needs of scholars." One of the ways that this is being accomplished in physics, particularly HEP, is with preprint, author-archived databases. This format for scholarly communication can increase speed of delivery, allow for an informal but wider idea of peer review, and significantly increase the availability of research results to a world wide audience for a fraction of the cost of journal publishing. Although preprints still face some problems and are only slowly taking root outside of the physical sciences, the success of preprints in HEP can serve as a model for the rest of the scholarly community.

Bibliography

Berman, P. American Physical Society E-print Workshop Panel on Intellectual Properly held in October, 1994. Retrieved from http://publish.aps.org/EPRINT/KATHD/toc.html on December 7, 2003.

Boyce, P.B. (n.d.). For Better or Worse: Preprint servers are here to stay. Scholarly Communication. Retrieved from http://www.ala.org/cfapps/archive.cfm?path+acrl/scholcomm.html on December 7, 2003.

Brown, C. (2003). The Role of Electronic Preprints in Chemical Communication: Analysis of Citation, Usage, and Acceptance in the Journal Literature. *Journal of the American Society for Information Science and Technology*, 54(5), 362-371.

Brown, C. (2001). The E-volution of Preprints in the Scholarly Communication of Physicists and Astronomers. *Journal of the American Society for Information Science and Technology*, 52(3), 187-200.

Cronin, B., McKenzie, G., Rubio, L., & Weaver-Wozniak, S. (1993). Accounting for Influence: Acknowledgments in Contemporary Sociology. *Journal of the American Society for Information Science*, 44(7), 406-412.

Dallman, D. Draper, M. & Schwarz, S. (1994). Electronic Pre-publishing for World Wide Access: The case of high energy physics. *Interlending & Document Supply*, 22(2), 3-7.

Daviss, B. (2003). Reasonable Doubt. *New Scientist*, ,177(2388), 36-44.

Department of Energy Eprints. Retrieved from http://www.osti.gov/eprints/ on December 7, 2003.

Giles, J. (2003). Preprint Server Seeks Way to Halt Plagiarists. *Nature*, 426, 7.

Giles, J. (2003). Critical Comments Threaten to Open Libel Floodgate for Physics Archive. *Nature*, 426, 7.

Ginsparg, P. (2003). *Can Peer Review be Better Focused?* Retrieved from http://arxiv.org/blurb/pg02pr.html on December 7, 2003.

Goldschmidt-Clermont, L. (1965). Communications Patterns in High-Energy Physics. Appears in *High Energy Physics Libraries Webzine*. 6. March 2002. Retrieved at http://library.cern.ch/HEPLW/6/papers/1 on December 7, 2003.

Goodstein, D. (1994). Pariah Science. *American Scholar*. 63(4), 527-542.

Harter, S.P. & Park, T.K. (2000). Impact of Prior Electronic Publication of Manuscript Consideration Policies of Scholarly Journals. *Journal of the American Society for Information Science*. 51(10), 940-948.

Hurd, J.M. (2000). The Transformation of Scientific Communication: A Model for 2002. *Journal of the American Society for Information Science*. 51(14):1279-1283.

Lawrence, S. (2001). Online or Invisible? *Nature*, 411(6837), 512.

National Science Foundation (NSF). (2001). *Survey of Doctorate Recepients*. Division of Science Resources Statistics. Retrieved at www.nsf.gov on July 5, 2003.

O'Connell, H. B. (2002). Physicists Thriving with Paperless Publishing. *High Energy Physics Libraries Webzine*. 6. Retrieved at library.cern.ch/HEPLW/6/papers/3 on July 5, 2003.

Okersen, A. American Physical Society E-print Workshop Panel on Intellectual Property held in October, 1994. Retrieved from http://publish.aps.org/EPRINT/KATHD/toc.html on December 7, 2003.

Pelzer, N.L. (2003). Bibliometric Study of Grey Literature in Core Veterinary Medical Journals. *Journal of the Medical Library Association*. 91(4), 434-441.

Pinfield, S. (2001). How Do Physicists Use and E-print Archive? Implications for Institutional E-print Services. *D-Lib Magazine*, 7(12). Retrieved from http://dlib.org/dlib/december01/pinfield/12pinfield.html on December 7, 2003.

Science Journals Online. National Public Radio Interview with Paul Ginsparg and David Voss. Retrieved at http://discover.npr.org/features/feature.jhtml?wfId=1011729 on December 7, 2003.

Smith, R. (2001). *Ethical and Privacy Issues, Particularly in the Biomedical Sciences*. Talk given at UNESCO/ICSU Conference on Electronic Publishing. Retrieved from http://www.warr.com/epubscience.html on December 7, 2003.

Stanford Public Information Retrieval System (SPRIES). (2003). Retrieved from http://www.slac.stanford.edu/spires/hep on July 6, 2003.

Tomaiuolo, N.P. & Packer, J.G. (2000). Preprint Servers: Pushing the envelope of electronic scholarly publishing. Searcher, 8(9), 53-61.

Warr, W. A. (2002). Evaluation of an Experimental Chemistry Preprint Server. *Journal of Chemical Information and Computer Sciences*, 43, 362-373.

Wills, M. & Willis, G. (1996). The Ins and Outs of Electronic Publishing. Journal of Business & Industrial Marketing,11(1), 90-104.