The Impact of Electronic Publishing on the Academic Community

Session 1: The present situation and the likely future

Electronic research archives for physics

Paul Ginsparg

Theory Division, Los Alamos, National Laboratory, MS-B285, Los Alamos, NM 87545, U.S.A.


Copyright Information

Abstract

I describe a set of automated archives for electronic communication of research information that has been operational in many fields of physics, and some related and unrelated disciplines, starting from 1991. These archives now serve over 50,000 users worldwide from over 100 countries, and process many millions of electronic transactions per month. In some fields of physics, they have already supplanted traditional research journals as conveyers of both topical and archival research information. Many of the lessons learned from these systems should carry over to other fields of scholarly publication, i.e. those wherein authors are writing not for direct financial remuneration in the form of royalties, but rather primarily to communicate information (for the advancement of knowledge, with attendant benefits to their careers and professional reputations). These archives have in addition proven equally indispensible to researchers in less developed countries.

A major lesson we learn is that the current model of funding publishing companies through research libraries (in turn funded by overheads on research grants) is unlikely to survive in the electronic realm. It is premised on a paper medium that was difficult to produce, difficult to distribute, difficult to archive and difficult to duplicate --- a medium that hence required numerous local redistribution points in the form of research libraries. The electronic medium shares none of these features and thus naturally facilitates large-scale disintermediation, with the resulting communication of research information being both more efficient and more cost effective. A correctly configured fully electronic scholarly journal can be operated at a fraction of the cost of a conventional print journal, and could for example be fully supported by author subsidy (page charges or a related mechanism, as already paid to some journals), ideally allowing free network distribution and maximal benefit both to authors and readers.
Another lesson is that authors are unlikely to accept 'electronic clones' of print journals (i.e. electronic versions identical in content, functionality, methodology and appearance to paper versions), whether transmitted via CD-ROM or via the network. The electronic medium should not be constrained by any former print incarnation and, in particular, easily implemented quality appraisal mechanisms in the electronic realm will be dramatically superior to the binary (i.e. one-time, all-or-nothing) procedure employed by the print medium, which in turn frequently conveys inadequate signal. Moreover, authors and their funding institutions will be empowered to insist upon retaining the right to distribute electronic research documents and attachments in the format produced by the authors. Authoring tools already allow a highly sophisticated end-user format, including automatic network linkages, and will continue to improve.

The essential question at this point is not whether the scientific research literature will migrate to fully electronic dissemination, but rather how quickly this transition will take place now that all of the requisite tools are online. Secondary open questions include determining the most effective means of cost recovery for the disseminators of this information, which agencies will be responsible for insuring the long-term archival integrity, indexing and cross-compatibility for the various research databases, and how peer review will be organized for those disciplines that depend on the value-added it can, in principle, provide.

Finally, I describe some of the major improvements, enhancements in functionality and other expansions projected over the next few years for the existing archives.

**Introduction**

Electronic publishing in science has recently become the focus of an increasing number of workshops and conferences, typically including representatives from professional societies and other scholarly publishing concerns, and members of the library community; but has only a small or vanishing participation from actual researchers. This is ironic since the average scientist provides the lifeblood of scientific publication on a daily basis as reader, author and referee, frequently as editor, and also as organizer of conferences, schools and workshops. Scientists consequently understand research publication from the inside out as few non-researchers ever could, and many have grown frustrated at patronizing attempts to assure them that unthinking preservation of the *status quo* is in their best interest. It is clear that many traditional roles will be shifted by the electronic medium, and new roles will emerge, although precisely which players will acquire the competence to fill which roles, and when, remains to be determined.

In principle, the new electronic medium gives us the opportunity to reconsider many aspects of our current research communication, and researchers should take advantage of this opportunity to map out the ideal research communication medium of the future. It is crucial that the researchers, who play a privileged role in this as both providers and consumers of the information, not only be heard but be given the strongest voice. In particular, we need to dislodge definitively the curiously prevalent notion that the future electronic medium will strictly duplicate, inadequacy for inadequacy, the current print medium. The typical researcher is not interested in 'electronic publication' *per se*, but rather in a 'knowledge network' that better represents the true nature of the research enterprise, unfettered by unnecessary artifacts of its former incarnation as a paper database.
Some history

Rather than relate here the full history of the 'e-print [electronic print] archives' and whatever has occurred since mid-1991, instead I will concentrate only on some highlights that serve to illustrate the major lessons learned to date, and suggest their implications for the future. (For additional background information, see my article at http://xxx.lanl.gov/blurb/ [1], originally adapted from a letter to Physics Today, June 1992. For some of the more recent publicity, see [2,3].

The first database, hep-th (High Energy Physics --- Theory), was started in August 1991 and was intended for usage by a small subcommunity of less than 200 physicists, then working on a so-called 'matrix model' approach to studying string theory and two-dimensional gravity. (Mermin [4] later described the establishment of these electronic research archives for string theorists as potentially "their greatest contribution to science".) Within a few months, the original hep-th had quickly expanded in its scope to over 1\,000 users, and after a few years had over 4\,000 users. More significantly, there are numerous other physics databases now in operation (see http://xxx.lanl.gov/ physics e-print archives) that currently serve over 50\,000 researchers and typically process millions of electronic transactions per month (see the http://xxx.lanl.gov/cgi-bin/show_weekly_graph weekly stats for an overview of growth in World Wide Web usage alone at xxx.lanl.gov).

At the same time the submission rate has more than doubled in the past three years to an anticipated 20\,000 new submissions during calendar year 1997. The continued stability of the database has moreover led to increased archival usage in all subject areas covered: the vast majority of requests are for papers more than a month old, and over a third of the requests are for papers more than a year old. As of October 1997, the database contained over 64\,000 submissions and its growth rate is accelerating. The physics community is rapidly moving to realize the vision for the future expressed in the "Report of the American Physical Society Task Force on Electronic Information Systems" [5]: "The dominant mode [of dissemination] will be via a single electronic physics library, or Physics Database, which will be the heart of a worldwide Physics Information System."

These systems are entirely automated (including submission process and indexing of titles/authors/abstracts), and allow access via e-mail, anonymous ftp and the World Wide Web. The communication of research results occurs on a dramatically accelerated time-scale and much of the waste of the hard-copy distribution scheme is eliminated. In addition, researchers who might not ordinarily communicate with one another can quickly set up a virtual meeting ground, and ultimately disband if things do not pan out, all with infinitely greater ease and flexibility than is provided by current publication media.

It is important to distinguish the form of communication facilitated by these systems from that of Usenet newsgroups or garden variety 'bulletin board' systems. In 'e-print archives', researchers communicate exclusively via research abstracts that describe material otherwise suitable for conventional publication. This is a very formal mode of communication in which each entry is archived and indexed for retrieval at arbitrarily later times; Usenet newsgroups and bulletin
boards, on the other hand, represent an informal mode of communication, more akin to ordinary conversation, with unindexed entries that typically disappear after a short time.

In effect, the archives do benefit from an automatic form of peer review, since users typically replace their submissions in response to direct feedback, and subsequent revisions frequently benefit as much or more from this feedback as from the conventional referee process. The archives have also recently implemented full 'version control', meaning that any earlier version in a series of revisions can be obtained to aid adjudication of priority disputes or otherwise track changes. The functionality of this unified global raw database offers potential dramatic improvements over the research communication mediated by the artificially partitioned database of the paper journal system. In addition to the efficient two-way transmission capabilities, as well as indexing and automated hyperlink references within papers, the system has a password-protection scheme which allows authors to transfer 'ownership' to any journal (or equivalent third-party overlay) for the purpose of freezing the submission, stamping a 'published' reference or incorporating errata/addenda (all by author/journal negotiation). Some journals (for example Physical Review D) have begun to accept the archive identifier as the electronic submission itself, and conduct their editor/referee interactions as well by means of the version retrieved from the archive.

While the high-energy physics community did have a pre-existing hard-copy preprint habit that had already largely supplanted journals as our primary communication medium, this is not a necessary initial condition for acceptance of an electronic preprint archive, as evidenced by recent growth into other areas of physics and mathematics, and even to computation and linguistics (http://xxx.lanl.gov/cmp-lg/). The economics for all this remains favourable, with a gigabyte of hard-disk storage currently averaging under $250 (i.e. roughly 25,000 papers including figures can be stored for an average of less than 1 cent apiece). Finally, politically correct elements typically fret over leaving the third world in the dust --- but the reality is that less developed countries are already better off than they were before: researchers in eastern Europe, South America and the Far East frequently report how lost they would be without these electronic communication systems, and how they can finally participate in the ongoing research loop. It will always remain easier and less expensive to get a computer connected to the Internet than to build, stock and maintain conventional libraries --- the conventional journal system had always been much less fair to the underprivileged.

To summarize, to date we've learned:

(i) The exponential increase in electronic networking usage has opened new possibilities for formal and informal communication of research information.

(ii) For some fields of physics, the online electronic archives immediately became the primary means of communicating ongoing research information, with conventional journals entirely supplanted in this role. Researchers will voluntarily subscribe and make aggressive use of these systems which will continue to grow rapidly. The current levels of technology and network connectivity are adequate to support these systems. (Though we anticipate the need for increases in transcontinental network carrying capacity to catch up with the recent explosion in non-
academic usage --- otherwise scientific usage will require either priority routing on the shared
network or an independent network.)

(iii) For some fields of physics, open (i.e. unrefereed) distribution of research can work well and
has advantages for researchers both in developed and undeveloped countries.

**Scholarly versus trade publication**

Before continuing, we must distinguish at this point between two very different types of
publication, formerly grouped together only due to accidental similarities in their modes of
production and distribution. Understanding this distinction is crucial to the future of scholarly
publishing endeavours. (My comments here have been strongly influenced by e-mail discussions
with Stevan Harnad and correspondents, some of which are available at
ftp://princeton.edu/pub/harnad/Psycoloquy/Subversive.Proposal/. Other relevant discussions of
electronic publishing issues by Harnad, with further references, are available at
http://www.princeton.edu/~harnad/intpub.html or equivalently at
ftp://princeton.edu/pub/harnad/).

In scholarly publication (as known as "esoteric scholarly publication"), we are writing to
communicate research information and to establish our research reputations. We are not writing
in order to make money in the form of royalties based on the size of a paying readership. We
have every desire to see maximal distribution of our work (properly accredited of course), and
would fight any attempt to suppress that distribution. In trade publication, on the other hand,
authors write specifically to sell their articles and books, and have direct financial remuneration
in mind from the outset. It is consequently in their interest as well to maximize distribution, but
at the same time to insure that each reader pays per view; for this the intermediation of a
publishing company to maintain an infrastructure to exact money from paying customers and to
root out bootleg distribution may well remain welcome.

So in scholarly publication, we have a situation wherein authors can joke that they would pay
people to read their articles. (N.B. This potential paucity of readership for any given article must
not be used as an argument that support of basic research is intrinsically wasteful --- it simply
results from the naturally restricted size of a highly specialized community, and does not directly
measure the ultimate utility of the research.) So the essential point is now self-evident: if we the
researchers are not writing with the expectation of making money directly from our efforts, then
there is no earthly reason why anyone else should make money in the process (except for a fair
return on any non-trivial 'value-added' they may provide, or except if, as was formerly the case in
the paper-only era, the true costs of making our documents publicly available are sufficiently
high to require that they be sold for a fee). Now we are ready to consider the current role played
by publishers of physics research information (at least in certain fields).

**The current role of physics journals?**

It is ordinarily claimed that journals play two intellectual roles: (i) to communicate research
information; and (ii) to validate this information for the purpose of job and grant allocation.
As I've explained, the role of journals as communicators of information has long since been supplanted in certain fields of physics, so let's consider their other role. Having queried a number of colleagues concerning the criteria they use in evaluating job applicants and grant proposals, it turns out that the otherwise unqualified number of published papers is too coarse a criterion and plays essentially no role. Researchers are typically familiar with the research in their own field, and must in any event independently evaluate it together with letters of recommendation from trusted sources. Recent activity levels of candidates were mentioned as a criterion, but that too is independent of publication *per se*: 'hot preprints' on a CV can be as important as any publication.

So many of us have long been aware that certain physics journals currently play NO role whatsoever for physicists. Their primary role seems to be to provide a revenue stream to publishers, a revenue stream invisibly siphoned from overhead on research contracts through library systems.

**Potential pitfalls**

So this goes a long way to explaining how it could possibly be that a system whose primary virtue is instant retransmission is able to supplant entirely established journals as a credible information source in certain fields. (Though it is true that e-print archives are technologically somewhat ahead of what established publishers are offering in ease of use and functionality, and are likely to remain so for the foreseeable future.)

With an example of an electronic system that physicists will voluntarily and actively use in hand, it is illuminating to consider how a poor understanding of the properties and potentialities of the electronic medium can lead to badly mistaken implementations. An example of this was an American Physical Society (APS) 'request for proposals' for an online version of *Physical Review Letters* back in autumn 1993. Its superficial problem involved asking that the electronic version be identical in appearance to the printed version --- in other words to clone electronically every unnecessary artifact of the paper version. Its more profound problem is that the entire journal structure and organization needs to be reconsidered in light of the electronic format. In an era of instantaneous communication, why is there still a need for a letters journal with its draconian page limits and atavistic claims of rapid publication? As is well known to potential physicist readers, artificial constraints result in articles too telegraphic to be useful either to experts or to non-experts.

While I have used familiarity with the situation within one small sector of physics publishing to illustrate these points, feedback from researchers in other fields indicates that there is a generic and growing frustration at the slowness of existing publishers to recognize that the needs of researchers can potentially be served in an electronic format in novel and creative ways. The current problem consists both of misguided selection criteria and of misplaced goals: publishers may measure the success of their journals by the number of pages published, whether certain artifactual and unnecessary constraints are met, and whether they're published 'on time' (i.e. with regularity, not with speed). 'Useful', 'readable', 'innovative' are not necessarily primary criteria in this established framework.
Even benign, non-profit organizations and learned societies can easily become addicted to the amenities of scholarly publishing and lose track of their original mandate: thus placing the revenue-generating potential of their established publishing enterprises above the need to furnish creative intellectual services to their constituents. Until recently there were few effective options for physicists or other researchers to break into an intellectually void closed loop involving only publisher and library systems. The resources necessary for production and distribution of conventional printed journals allowed publishers to focus on their mechanics, and avoid any pressure to rethink the intellectual content and quality of their operations.

Problems and possibilities

Why is it that the current implementation of peer review, as employed by paper journals, needs to be entirely rethought in view of new possibilities afforded by electronic publication and dissemination?

A most obvious problem in the current scheme is that as the number of researchers in any given field has grown (both due to global population increase and increased Cold War funding for the sciences), the number of papers published in journals for any given field has vastly exceeded the ability of any one researcher to read and absorb. While perhaps there once was a time when a physicist could pick up a single journal each month and read it from cover to cover to remain abreast of all of physics, this idyllic state of affairs is not even a distant memory for any recent generation of physicists. Nonetheless, this outmoded methodology effectively remains the basis for many aspects of the current implementation of peer review, in physics and in other fields.

Once the mere fact of publication in a journal no longer gives a particularly useful guide, readers are forced to perform the majority of the selection on their own by some set of additional criteria, and their primary need is simply access to the information as quickly as possible. For this reason, a systematic preprint system was set up for high-energy physics institutions in the early 1970s and largely usurped the role of conventional journals as conveyors of topical information. This widespread preference for rapid access over the limited filtering provided by peer review was even more dramatically reinforced with the advent of the electronic preprint (e-print) archives in the early 1990s, which quickly grew to supplant in addition the conventional archival role of journals in many fields.

This is not, however, to argue that peer review cannot in principle provide substantial added value to the reader. One of the foremost problems at present is the large amount of information lost in the conventional peer review process, with the end result only a single one-time all-or-nothing binary decision. Although this may somehow be adequate for the purpose of validating research for job and grant allocation, it clearly provides little benefit to the average reader.

A variety of superficial improvements can easily be implemented immediately in the electronic realm. Since there are no financial or physical barriers to widespread dissemination, we can imagine a relatively complete raw archive unfettered by any unnecessary delays in availability. Any type of information could be overlayed on this raw archive and maintained by any third parties. For example, the archive could be effectively partitioned into sectors, graded according to overall importance, quality of research, or other useful criteria, and papers could be shifted
retroactively as dictated by additional information or follow-up research. And rather than face only an undifferentiated bitstream, the average reader could benefit from an interface that recommended a set of 'essential reads' for a given subject from any given time period. There could also be retroactively added descriptive information, "this paper was important since it drew upon a, b, c (hyperlinks to sources) and led to new developments x, y, z (more hyperlinks)" to provide a further guide to the literature. Or the interface could point to a specific paper as having been important, but warn the beginner to go first to a later paper by the same (or other) authors that subsumes, extends or corrects the same results in a more understandable fashion; or say that this paper generated much attention but skip it since the fad played itself out and people returned to more serious pursuits. The literature need not be frozen in time as in the paper medium, but can remain as fluid as the research itself. Even interdisciplinary research (for example if I as a particle physicist wished to peruse the recent literature in biophysics or even biochemistry) can be easily facilitated by an interface that allows rapid identification of papers that provide pedagogic review material or are otherwise likely to be of specific interest to outsiders. Further possibilities such as moderated comment threads attached to specific points in papers together with more exotic features can be added in successive stages as desired. Ideally the standards set by the physics community can serve as a model for the rest of scientific research communication.

Who needs it?

Will the enthusiastic use of the instant communication provided by free access to unreviewed electronic archives ultimately emerge only as an artifact, preferred only in isolated subsets of the scientific community? This is to a certain extent an experimental question, answerable only after all the bits have settled. But it is worthwhile to speculate on features that may characterize those scientific subcommunities most likely to find it practical and efficient in the future to sidestep the conventional peer-review structure for rapid access to new results, while still maintaining some form of electronic peer-review system to provide validation of and guidance to their archival literature. In other words, looking beyond current experience drawn from a well-defined and highly interactive community of voracious readers with a pre-existing hard-copy preprint habit, with a standardized word processor and a generally high degree of computer literacy, with a rational means of assigning intellectual priority (i.e. at the point of dissemination rather than only after peer-review), and with little concern about patentable content, all of which may be regarded as momentary historical accident, is there some more abstract characterization of the required autonomy that allows a circumscribed community to flourish rather than suffocate in its own unreviewed output stream? Again it will be easier to argue these issues in retrospect someday, but at least one noteworthy feature can be identified: in my own research discipline, the author and reader communities (and consequently the referee community as well) essentially coincide. Such a closed peer community may signal a greater intrinsic likelihood for acceptance and utility of free electronic dissemination of unreviewed material.

Research communities comprised of a relatively small number of authors and a much larger number of readers could ultimately settle on a very different model, wherein the institutions that support the research assert copyright privilege, assume the role of publishers, and disseminate material produced in-house for a fee to those institutions that only consume it. Though this would upset proponents of free electronic access to all publicly supported research material, it would at least be a logical system, in which the real risk-takers --- namely the institutions that
support research by way of investment in salary and equipment --- are able to profit from and protect the products of that investment. The current system, which cedes full copyright of high-quality content to low-risk publishers who step in at the last moment and provide at most a comparatively insignificant few hundred dollars of added-value (in most cases even selling it back at high prices to the initial sponsoring institution), has never been particularly sensible.

With the automatic form of peer review and version control described above, the global archives are not at all incompatible with the filtering role historically provided by the journal system. To the contrary, they beckon for learned societies to augment their current roles with new forms of intellectual overlays never before feasible. Professional societies could further speed this development by promoting a shared copyright scheme to their members, explicitly allowing authors (and their institutions) to retain electronic full distribution rights to documents as produced by the authors.

**Cloudy futures**

For the moment, conventional publishers have continued to express their unbridled enthusiasm for open electronic dissemination systems, despite an intrinsic potential for subversion. As long as their bottom line is unaffected, they can afford to be arbitrarily magnanimous in their desire for peaceful coexistence: "After all we have long been in the business of propagating research information, we would never dream of trying to suppress it in any way..."

But ever financially pressed research libraries are poised for triage of their journal subscriptions. And as pointed out by [ftp://princeton.edu/pub/harnad/Psycoloquy/Subversive.Proposal/who-payspiper.14.quinn.peer-review-physics](ftp://princeton.edu/pub/harnad/Psycoloquy/Subversive.Proposal/who-payspiper.14.quinn.peer-review-physics) (Quinn, 1994), there's a potential explicit mechanism to encourage preferential cutting of subscriptions to physics journals. Libraries, faced with difficult choices, may decide that physicists already have an alternate information feed from the raw global electronic database; and physicists may well complain the least (or not at all) when their journals are threatened with cancellation. (Indeed this is already reported to be happening in India and other places with severely limited financial resources --- as argued above, the less-developed countries stand to benefit at least equally from recent technological developments).

The physics and mathematics archives now offer a variety of choices of high-quality output formats (TEX source, hyperdv[1][http://xxx.lanl.gov/hypertext], gzipped hyperPostScript with choice of font resolution or type 1 PS, or pdf) and will be able to support higher level formats as they become available. With this aspect of end-user accommodation thus trivialized, the near-term concerns have shifted to the continued development of a robust global mirroring system, and to better means of handling meta-level indexing information. Additional mirror distribution sites in Australia, Brazil, China, France, Germany, India, Israel, Italy, Japan, Russia, South Korea, Spain, Taiwan and the U.K. have given better response times, especially to international users whose access is increasingly impeded during times of day when their national networks and transcontinental links suffer from the congestion caused by recent increases in non-academic network traffic. In the long term, they also provide a global backup system resistant to localized database corruption and/or loss of network connectivity. The problems of indexing and categorization of information in principle lie within the purview of the library and information science communities, but to date theirs has been a curiously low profile in the electronic realm,
while various amateur brute-force indexing schemes are running dangerously amok. It would be remarkable if centuries of ostensibly relevant experience found little applicability in the network context.

We should also be alert to risks borne by authors who may find themselves prematurely encouraged to abandon "chemicals adsorbed on to sliced processed dead trees" in favour of an electronic-only archival format. There is a certain leap of faith involved here, since every once in a while one does after all get lucky and write a paper that could still attract readership a century from now. The physical format, with a worldwide system of institutional libraries serving as a multiply redundant distributed archive, has proven robust on the time-scale of centuries to anything short of global cataclysm (in which case we'd probably have more pressing concerns). No current electronic format has proven similar longevity --- for the simple reason that all have been in existence for little more than a decade if that. Few claim to know what will be the preferred electronic format a century from now, but I'm willing to go out on a limb and assert that it will be none of TEX, PostScript, pdf, Microsoft Word or any other format currently in existence. On the other hand, this is certainly not a fundamental problem of principle, and perhaps scientists will eventually come to rely on much-needed logistical assistance from future librarians in their role as archivists: just as endangered material on decaying acid paper is currently migrated to microfilm, automated translation to newer and more general electronic formats should always be possible during transition periods, provided there is an acknowledged need to prevent our living research archives from becoming data cemeteries.

One possibility is that some consortium of professional societies and institutional libraries will ultimately acquire the technical competence to provide umbrella sponsorship of the global raw research archive. Those societies that are non-profit publishers as well may continue to organize high-quality peer-reviewed overlays (though perhaps no longer as a means of generating income to subsidize other non-publishing ventures), and certain commercial publishers accustomed to large pre-tax profit margins on their academic publishing activities will probably have to learn to compete in more realistic marketplaces.

In the long term, it is difficult to imagine how the current model of funding publishing companies through research libraries (in turn funded by overhead on research grants) can possibly persist. As argued by Odlyzko (1994; ftp://netlib.att.com/netlib/att/math/odlyzko/tragic.loss.Z), it is premised on a paper medium that was difficult to produce, difficult to distribute, difficult to archive, and difficult to duplicate --- a medium that hence required numerous local redistribution points in the form of research libraries. The electronic medium shares none of these features and thus naturally facilitates large scale disintermediation, with attendant increases in efficiency benefitting both researchers and their sources of funding. As described above, recent developments have exposed the extent to which current publishers have defined themselves in terms of production and distribution, roles which we now regard as trivially automated. But there remains a pressing need for organization of intellectual value-added, which by definition cannot be automated even in principle, and that leaves significant opportunities for any agency willing to listen to what researchers want and need.
References


[Copyright Information](#)