SLAC-PUB-8357
cs.CY/yymmnnn
February 2000

Physicists Thriving with Paperless Publishing

Heath B. O'Connell[1]
Stanford Linear Accelerator Center
Stanford CA 94309
hoc@slac.stanford.edu
http://www.slac.stanford.edu/grp/lib/people/hoconnell.html

The Stanford Linear Accelerator Center (SLAC) and Deutsches Elektronen Synchrotron (DESY) libraries have been comprehensively cataloguing the High Energy Particle Physics (HEP) literature since 1974 and the core database, SPIRES-HEP, now indexes over 400,000 research articles, with almost 50% linked to fulltext electronic versions. This database motivated the creation of the first web implementation in the United States (and the second in the world). With this database and the invention of the Los Alamos E-print archives in 1991, the HEP community pioneered the trend to "paperless publishing" and the trend to paperless access, in other words, the "virtual library." We examine the impact this has had both on the way scientists research and on paper-based publishing. With the E-print distribution having evolved from an established tradition of sending out hard-copy pre-prints, the standard of work archived at Los Alamos is very high (70% of papers are eventually published in journals and another 20% are conference proceedings). To allow authors to "thrive" the SPIRES-HEP collaboration has been ensuring that as much information as possible is included with each bibliographic entry for a paper. Such meta-data can include, tables of the experimental data that researchers can easily use to perform their own analyses as well as detailed descriptions of the experiment, citation tracking and links to full-text documents.

[1]Work supported by the US Department of Energy contract DE-AC03-76SF00515.
Invited talk at the AAAS Annual Meeting and Science Innovation Exposition, February 17-22, 2000 in Washington, DC

1 Introduction

The world wide web is now ubiquitous. At a time when every advertisement on television seems to end with a URL it is of some interest to note that the initial interest in the web was generated by its use in the SLAC Library to aid the research of High Energy/Particle physicists. In this talk we explore the role technology has played in the organisation and dissemination of information in High Energy Physics (HEP), where paper is finally a medium of last resort, and discuss why HEP provided the ideal conditions for the rapid adoption of new technology.

2 SPIRES and the Internet 1969–1990

Physicists have always been willing to communicate their results before publication in the journals. They, or perhaps their departments, would send out *preprints* of their work to institutions or other researchers they personally knew. Naturally because of the cost and effort involved in this, large departments and famous scientists had an enjoyed a numerical advantage both in sending and receiving. However for every poor soul in a small, out of the way department lamenting intellectual starvation, a well known researcher would be suffering from an overstuffed mailbox. By the 1960's the sheer number of preprints had somewhat perversely made the communication of research *more* difficult, due to information overload.

In 1968 the Division of Particles and Fields (DPF) of the American Physical Society and the SLAC librarians, Louise Addis, Bob Gex and Rita Taylor decided to bring order to this situation for the entire particle physics community [1]. The SLAC Library was well positioned for this rather large job. Since its foundation in 1962 it had been actively collecting new preprints (and as a world centre for physics it attracted a lot) and publishing a weekly list of them for the SLAC community. In time this authoritative list was being sent to other institutions. Therefore to get your work known all around the world (or at least the title of your work seen) all you would have to do would be send your preprint to SLAC, and similarly you could browse the titles of all the week's preprints by just looking at one list. The SLAC Library was aided in this task by an experimental computer database project the Stanford Physics Information Retrieval System (SPIRES), into which the bibliographic information of the preprints would be added and from which

the list of all preprints entered in the past week could be easily generated. In January 1969 the first Preprints in Particles and Fields (PPF) list was sent out to over a thousand eager subscribers.

A sister list "Anti-preprints" listed the preprints that had since been published in journals. This allowed the original preprints to be discarded, and proved quite popular with the journal editors, who were then able to match references to preprints with the published article. By 1974 the SLAC Library (with invaluable help from its collaborators [2]) was comprehensively cataloguing the majority of preprints in high energy physics, and by extension, the published literature.

This new PPF age, though, was not perfect. Receiving lists of new preprints is all very well, but after several years one is left once again with a stack of paper. Trying to find a particular article or information on a certain subject is very difficult. The Internet allowed a further development.

The problem of finding just the right preprints from this ever growing body of research was not an issue for the SLAC community. The SPIRES database allowed searches by date, author, title and a number of fields. Simply by logging in to their computer, SLAC physicists were able to search through thousands of papers, and the paper PPF list would serve as a newspaper, with a similar longevity. With the rise of the Internet anyone in the world could now access this service. A program, QSPIRES[3], developed in 1985 by SLAC database systems developer George Crane [4], allowed physicists to search the SPIRES database using E-mail. One would send off an E-mail query to the database and a swift reply would follow. People were able to find all the papers by a particular author or from a certain institution, or find out how many citations their work had[5]. The PPF and PPA lists could also be sent by E-mail.

3 TEX and the Single Archive

By 1990, lists of new articles were being sent to physicists by E-mail and people were searching for articles in the SPIRES database via E-mail. Everything was modern and electronic, up to the point that you actually wanted to read a paper. In this case, the reader would have to request that an author send a copy in the mail. Thus if you were overseas it could take some weeks before you actually saw the article; clearly unacceptable.

Around this time another technical innovation had sufficiently matured

to obviate this problem. High energy physics articles usually contain a lot of mathematics, which makes them difficult to write with a standard typewriter. In the late 1970's Stanford computer scientist Donald Knuth invented a special typsetting program, TEX (pronounced "tek"), that could display mathematics beautifully [6]. This language soon proved very popular with physicists and mathematicians as it gave them complete control over the production of their documents.[2]

The underlying "tex-file" was simply a normal, (*portable*) text file, where the mathematics was written using certain rules which could then be processed (*on any computer*) into a PostScript (PS) file [7], changing

```
\int_0^1
\frac{e^x}{2\pi} dx
```

to

$$\int_0^1 \frac{e^x}{2\pi} dx.$$

The PS file can then be printed out and read. Therefore to send a paper, all one would have to do is send the tex-file via E-mail and the person at the other end could process it and print it out; another step in the communication process had entered the future. By 1990 the use of TEX among the HEP community was almost universal.

With preprint list distribution, SPIRES database searches, requests for papers, and even the transfer of these papers being done through E-mail, all stages of the process were electronic and pretty much immediate ... except one. SLAC was still receiving the E-prints as "hard-copy". In 1991, Paul Ginsparg of Los Alamos National Laboratory (LANL) decided to do something about this. If authors could send their papers electronically to a central repository, with author-supplied bibliographic data, a list could be sent out of each *day's* additions and the preprints could be obtained directly from this archive [8]. Here, TeX proved a godsend, as it provided a means of storing rather small files that researchers could request and then process into the viewable PS files at their home institutions. In August 1991 the first paper was sent to the Los Alamos archive. Twenty seven papers were sent that month[3] and physicists began receiving daily E-mails of the new papers sent to LANL.

[2]This paper has been written using a derivative of TEX

[3]In August 1999 there were over 800 particle physics papers were sent to the LANL archive

Soon there were over 2000 subscribers to this daily E-mail notification and it gradually began to replace SLAC's PPF list as the HEP community's most immediate source of new research information. Far from competing, though, the two services at LANL and SLAC, complemented each other nicely. SLAC continued to process hard copy preprints that were never sent to LANL and journal articles that had never been preprints, adding such information as their references, experiment numbers, authors' institutions, etc. The electronic preprints, *E-prints*, began to account for a good deal of the HEP literature, and Ginsparg's system allowed for considerable automation and efficiency in SLAC's work. A good deal of useful information could be harvested electronically from the files authors supplied to LANL (though there was still much that had to be done by hand).

This evolution from preprints to E-prints had another important facet. The E-prints were assigned a unique number of the form *archive/yymmnnn* (eg 9501251 for the 251st paper in January 1995). In the past "report numbers" had been assigned to preprints by the author's home institution in a less than completely systematic way and with a variety of conventions, which made trying to track citations of them too difficult for the SPIRES database and only citations to published journal articles were recorded. Thus many citations made to an author's paper while it was an unpublished preprint would be lost, to the author's chagrin. The standardised E-print number changed all that, as pretty soon these numbers started appearing in reference lists. After some work, these citations to unpublished works could finally be registered, and it is worth noting that the most cited HEP article in 1998 spent some ten months as an unpublished E-print. The next step was to ensure that the database recognised that the published article and the E-print were essentially the same paper, and thus that the true number of citations was the sum of the citations of both. In some instances this lead to double counting, which needed special programming to eliminate, as it became fashionable to include *both* the E-print number and the journal reference when citing an article!

4 The Mouse That Clicked

With the SPIRES database and the LANL archive you could find a particular paper and be reading it within minutes merely by sending E-mail, after just a little work to process the TEX file and print it out. A lot had changed

in 20 years since the SLAC Library first started sending out a list of all the preprints it had received. However, the guiding philosophy behind computing is that no work a person has to do can ever be "too little," and a new technology was about to make the research process even easier.

In September 1991, SLAC physicist Paul Kunz, was visiting CERN, Geneva, where he was shown the infant World Wide Web, which allowed information on different computers to be accessed in a very user-friendly (and now completely familiar) manner. Realising this would be perfect for searching the SPIRES database he told then librarian and SPIRES database manager, Louise Addis, about it upon his return to SLAC. She was quick to recognise the potential of this [9] and by December SLAC had the first WWW server in the world outside CERN.[4] The E-mail access-system became obsolete as more and more physics departments installed the software necessary to reach the SPIRES WWW interface [10]. A group was formed at SLAC, the WWWizards [11] to provide help with WWW technology [12].

The key feature of the WWW is linking. Therefore when displaying search results in the SPIRES database, a number of things could be linked to each record. One of these is the full-text at Los Alamos which was being stored in a minimal fashion as TEX source. As mentioned before, TEX files need to be processed before they can be printed out (see the discussion on page 4). Initially though, due to either mistakes in the TEX files or different version of TEX being used, some physicists experienced difficulty in processing the files at LANL. In order to enter all the relevant information into the SPIRES database the SLAC Library staff needed needed to print out each paper, and so would process the TEX files sent to Los Alamos into PS files. At first this was done manually, which soon grew to be very time consuming. Luckily the SLAC Library was helped by Paul Mende from Brown University who created an automatic procedure. This PS generating code was ultimately incorporated into the LANL submission process – if your paper couldn't be processed, it wouldn't be accepted. The next step was to link this PS file to the SPIRES record, reducing the work you had to do to read a paper to a few clicks of a mouse. The technical challenge of creating a web browser able to view PS files was surmounted and the SPIRES database began to link directly to PS files generated and stored at the Los Alamos archive. Paper had been eliminated from the process and the virtual library was born [4].

[4]There was no Initial Public Offering.

5 The Many Hands Interpretation

Authors sending their electronic texts to LANL has "pushed" the effort of publication onto the researchers themselves, who benefit from an inexpensive, immediate, wide-scale dissemination of their work. This idea has led many computer scientists to ponder automated systems for indexing and retrieving these full-text papers. How much of the work of collecting information on the literature can be realistically facilitated by the authors? The combination of the Web, the SLAC Library's automated systems and the LANL archives provides an interesting testing ground for this question.

The goal of the LANL archive is to be as automated as possible, so that it can exist without administrative intervention (as opposed to the SPIRES database). To this end it has a number of checks to ensure all submissions meet the entry requirements (one we have already discussed ensures the TEX file is successfully processed into PS). The basic bibliographic information (author, title, etc.) data is supplied by the users in a neatly structured format that can be downloaded into the SPIRES database automatically. This raw information requires only minor attention from the Library staff who, among other things, ensure spelling consistencies, and add in the author affiliations from the INSTITUTIONS database.

Far and away the most time consuming part of this is collecting the references of each paper, from which the citation searching is built [5]. As citation results reflect some degree of professional accomplishment, physicists tend to be rather interested them and a good deal of their correspondence with the SLAC Library concerns omissions or mistakes in the reference lists (which is exacerbated by the posting of "revised versions" of E-prints to LANL, some of which contain additional references). Originally, of course, these reference lists were typed into the database, but as another happy spin-off of TEX about 90% of them can now be extracted from the author's file. Unfortunately, this only makes a huge job large, as they still have to be checked by the Library staff, and authors regularly confound the reference extracting program by adopting imaginative new ways of writing the journal-volume-page sequence, or simply making errors in a reference.

Obviously the LANL model suggested there should be some way to place the burden of constructing and checking the reference list on the authors themselves (as they do with bibliographic information), before they send their paper out to the world. Two things really stood in the way of this. First, the original program that extracted the references wasn't really definitive enough

for this sort of user-side checking, as there was no real way of specifying which references the code would extract, and which it wouldn't. The second problem was that with physicists being rather busy, any new system would have to result in less, not more, work for it to be widely adopted.

Once again, TEX helped with the solution. The SPIRES team developed and offered a new service: the database would display records in TEX format, with the exact information needed to construct the reference as an additional tag. The authors then had a very simple process of cut and paste to create their reference lists, and the tag would sit invisibly in the TEX file (not appearing in the final PS file) waiting to be extracted at the SLAC library. The strict structure of the CITATION tag and the simplicity of its extraction allowed for the creation of a checking program that authors could use, for example:

```
\bibitem{O'Connell:1997}
H.~B.~O'Connell,
``Recent developments in $\rho\!-\!omega$ mixing,''
Austral.\ J.\ Phys.\ {\bf 50}, 255 (1997)
[hep-ph/9604375].
%%CITATION = HEP-PH 9604375;%%
```

which becomes:

H. B. O'Connell, "Recent developments in $\rho-\omega$ mixing," Austral. J. Phys. **50**, 255 (1997) [hep-ph/9604375].

So far almost three hundred papers have been written using this system.

A second way we've been experimenting with author–supplied data is asking them to help us with another thing that is of great interest to them: missing papers in our database. Traditionally adding them in was another time consuming exercise that would require Library staff to get the journal from the shelves and type in all the relevant information. Using a web form that authors can fill out themselves allows a paper to be added to the database simply by cutting and pasting on the part of the Library staff.

By inviting our users to help us in maintaining the database and automating as many processes as possible, we have been able to accomplish a lot of additional work without spending significant amounts of extra time. This would not be possible, though, without the eagerness and attention to detail that characterises the Physics community.

6 All the news that's fit to link

The WWW created a unique opportunity for the SPIRES database by providing a system that would conveniently attach to any record all the related information both inside the database and around the world in a very compact manner. Over the past six years we have worked on making this process as efficient as possible both in terms of computer programming and the output display our users see. We shall now discuss how this works.

A single record in the Literature database might have certain *basic* elements such as

- Title
- Author (and author's institution)
- Date
- Publication note

but the article could also be *hyperlinked* to

- the appropriate record in the EXPERIMENTS database
- full text of article at the journal server or the E-print server
- the references of the article
- other articles that cite it
- experimental data in the REACTIONS database,

allowing the researcher to selectively explore any particular related items. Thus an original record only ten lines long becomes a gateway to information stored around the globe.

The construction of these links requires special care and co-operation with outside services. One particular case that followed an evolutionary process was linking to the published version of the document on journal home pages. Optimally, this URL would point to a unique "abstract page," rendered in HTML so that the link to the journal server can be fast. Once there the user can be presented with all the Journal services such as full text in possibly a variety of formats (the most common being PS or the newer Portable Document File (PDF)). From the point of view of the publisher, this means

that a URL has to be found for every article. From our point of view this URL should be calculable from the information we already have about the record, as it then permits us to run in the URL automatically. Articles have always been cited through the journal-volume-page (JVP) convention, so it made sense that a number of major publishers, including the American Physical Society [13] and Elsevier [14] adopted a URL scheme based on these three elements. Setting up such a system does present something of a technical challenge for the journals, but is well worth it in terms of presenting the simplest possible interface to the outside world and providing reliable access to their wares.

The journals also link back in to our database, mainly for the references of the paper (though publishers have proven less eager to link to our database record of the actual paper). In an effort which I hope will become more widely adopted, we have worked closely with the American Physical Society's *Physical Review* [15] to share the bibliographic data (including reference lists) and either update an existing record in our database, or add a new record, when the publish new articles.

Where possible we have also tried to link to other literature databases. Some like the CERN Library's database [16] have a significant overlap with our system (there are currently over fifty thousand thousand two-way links between SPIRES and CERN). Others, such as Harvard's *Astrophysics Data System* [17]and the American Mathematical Society's *Mathematical Reviews* (AMSMR) [18] cover a much smaller set of papers in our database, but from a different perspective and connect the paper to other academic disciplines. Here, once again, we need to work with the other databases to ensure maximal efficiency and reliability in this linking. Our links to the Harvard database use a JVP scheme, while those to CERN and AMSMR databases use unique record keys. SPIRES can be linked to either way. There is also the consideration that the link should bring additional information, rather than just repeating what record in the SPIRES database (or else why bother?). Therefore we apply these links selectively.

7 So why HEP?

Why did the High Energy Physics community provide such a fertile ground for this particular aspect of the so-called "Information Revolution"? We have touched on the reasons throughout this talk, but it is useful, perhaps,

to summarise them to understand how this model might be more widely adopted (or why it might *not* be).

The first thing is that it has not been a revolution so much as an evolution. Particle physicists, have always been compulsive communicators. Generally free of commercial or governmental restrictions, the habit of sending out advanced copies of work goes back to the days of carbon paper. They are also have a long standing tradition of international collaboration, so any new advances in communication technology are eagerly adopted, and the high technological literacy facilitates this. This is especially true of the experimental community whose collaborations can have hundreds of members all over the world and whose experiments depend on high speed computer network connections. It is worth noting that the two SLAC physicists most closely involved in spinning the web at SLAC, Paul Kunz and Tony Johnson are experimentalists, though I should also point out that Paul Ginsparg of Los Alamos is a theorist.

In this environment the sheer quantity of existing literature necessitated creating the organisation begun by the SLAC Library in the late 1960's. Through using this PPF list service, physicists became accustomed to the idea of a central "clearing house" for preprints. As the use of E-mail spread through the Physics community, particularly the High Energy sector, the SPIRES E-mail interface was introduced in the mid 80's and before long was being used by thousands of researchers in over forty countries. The TEX typsetting language, favored by those needing to write complex mathematics, allowed one to send simple ASCII files using E-mail, thus giving rise to the electronic distribution of preprints, which was centralised by Ginsparg's E-print archive. The World Wide Web then laid the ground for a highly powerful way to integrate the various facets of research communication.

It should be acknowledged that these are rather special conditions. It is instructive to note that in other scientific communities where TEX is less heavily used, the adoption of the LANL archives has been noticeably slower, though in principle there is nothing to stop them (despite being built with TEX in mind, the archive is not dependent on it, and will accept submissions in a variety of formats). The TEX fluent mathematics community, on the other hand, though quick to adopt the new system, perhaps lacked the detailed computing knowledge to play a leading role.

For different research communities there are other obstacles. Faced with the incredibly rapid and wholehearted adoption of this E-print scheme by high energy physicists, the traditional journal publishers found it best to

not to offer an real resistance. They had nothing to fear, really, as over seventy percent of HEP papers sent to Los Alamos are eventually published in journals and another twenty percent appear in conference proceedings. In other fields, with newer archives, this is still something of a touchy subject, such as in the biological and medical research communities, as can be seen in the Biochemical Society's response [19] to the NIH's proposal for a preprint server for the life sciences [20].

8 Conclusion

In this paper I have described the evolution of paperless publishing in a particular academic community and how special circumstances conspired to make this happen faster and more comprehensively than in any other field. In doing this I have hoped to convey the sense that this was indeed a process of evolution rather than revolution.

It may be interesting to speculate what the future holds. Will more and more of what used to be called "bibliographic data," become author–supplied "metadata" which annotates each article with various related information? Can this metadata be shared to create cross-database, and cross-discipline, linking [21]? Our experience shows that if you make the procedure simple enough, and offer an advantage to the authors for doing it (such as the ability to check their reference list described on page 8), they *will* use it.

We have begun in the SPIRES-HEP database to exploit Web technology to handle information beyond the traditional bibliographic record or what one could obtain from a print version (such as the full text and reference list). These other facets range from the paper's citation list, which is constantly growing (if the author is fortunate), to the home-page of the experimental collaboration that wrote the paper, to information researchers might really want out of a paper, in a form far more useful than could ever be delivered in paper (such as a computer file of experimental data or 3-D computer images created by the author).

The fluidity of this new, electronic means of publication also has implications for the traditional research process. At what point does an author "give the final answer" and the paper become "set in stone"? The LANL archive, has neatly addressed this issue in an appropriate manner, by archiving each version of an E-print with a date stamp, documenting the growth of the paper. Presently the traditional model holds sway: once the article is published

in a journal, it is a finished piece of work, and any subsequent alterations are to be handled via errata. Other authors may *comment* on the published paper, to which the author may *reply*, but the comment and reply are treated as new articles, and the original is left unchanged. One journal, however, has broken away from the traditional model. *Living Reviews in Relativity* [22] exists solely on the Web, and allows authors to constantly revise their articles. In part this is due to the pedagogical nature of the journal; it seeks to provide reviews that aid learning, rather than publish original research.

Well aware of the difficulties of predicting anything in the Internet world, I have merely tried in concluding to state some of the current trends in electronic publication. I am certain, however, that future holds many exciting new innovations in store.

Acknowledgements

I would like to thank Louise Addis and Pat Kreitz for reading this manuscript and offering helpful comments and suggestions. I have also enjoyed conversations on the topics mentioned here with Richard Dominiak, Mark Doyle, Paul Ginsparg, John Jowett and Paul Kunz.

References

[1] L. Addis, "SLAC Library monitors underground Physics Press," SLAC News (June 1971).

[2] DESY, CERN, University of Durham, KEK, Yukawa Instititute and Fermilab.

[3] H. Galic, "Guide to QSPIRES and the particle databases on SLACVM," SLAC-0393.

[4] P. A. Kreitz, L. Addis and A. S. Johnson, "The Virtual library in action: Collaborative international control of high-energy physics preprints," SLAC-PUB-7110 *Presented at Grey Exploitations in the 21st Century: The Second International Conference on Grey Literature, Washington, D.C., 2-3 Nov, 1995*; L. Addis, H. Galic, P. Kreitz and A. Johnson, "The Virtual library in action," SLAC-PUB-8185 *Presented at 209th American Chemical Society National Meeting, Anaheim, CA, 2-6 Apr 1995.*

[5] In the SPIRES database we use the following terminology: if paper B includes an earlier paper A in its "reference" list, then paper A is "cited by" paper B [http://www.slac.stanford.edu/spires/hep/references.html]

[6] D. E. Knuth, "The TEX book," Addison Wesley 1986 (see also http://www.tug.org).

[7] http://www.postscript.org

[8] http://arXiv.org

[9] T. Berners-Lee, "Weaving the Web," Harper 1999

[10] http://www.slac.stanford.edu/spires/hep

[11] http://www.slac.stanford.edu/find/wizards.html

[12] J. M. Deken, "First in the web, but where are the pieces?" SLAC-PUB-7636.

[13] http://www.aps.org

[14] http://www.elsevier.com/npe

[15] http://prd.aps.org

[16] http://alice.cern.ch

[17] http://adsabs.harvard.edu

[18] http://www.ams.org/mathscinet/search

[19] www.biochemistry.org/ebiomedresp.htm

[20] www.nih.gov/welcome/director/pubmedcentral/pubmedcentral.htm

[21] H. Van de Sompel and P. Hochstenbach, "Reference linking in a hybrid library environment," D-Lib Magazine **5** no. 4 and D-Lib Magazine **5** no. 10 [http://www.dlib.org]

[22] http://www.livingreviews.org