SLAC-PUB-7110
February, 1996

# The Virtual Library in Action: Collaborative International Control of High–Energy Physics Pre–Prints*

P. A. Kreitz, L. Addis, H. Galic, and T. Johnson

Stanford Linear Accelerator Center, Stanford University, Stanford, CA 94309

## Abstract

This paper will discuss how control of the grey literature in high–energy physics pre–prints developed through a collaborative effort of librarians and physicists. It will highlight the critical steps in the development process and describe one model of a rapidly evolving virtual library for high–energy physics information. In conclusion, this paper will extend this physics model to other areas of grey literature management.

## 1. BACKGROUND

The Stanford Linear Accelerator Center (SLAC) Library has been acquiring and cataloging high–energy physics pre–prints for almost thirty years. While the Library maintains a suite of databases for the world wide high–energy physics community's use, its flagship database is called SPIRES–HEP (Stanford Public Information Retrieval System–High Energy Physics). SPIRES–HEP currently contains over 300,000 bibliographic entries for high–energy and related particle physics pre–prints, published journal articles, theses, and technical reports. The database grows by approximately 20,000 new entries each year and for recent years, includes abstracts and links to 27,000 papers, the electronic version of the pre–prints or articles, when available.

This database is tailored to its user community and so bibliographic entries contain information of value to the users. All authors are included with individual links to their associated institutions (the record may be 850 authors for one pre–print). The names of

*Work supported by the Department of Energy contract DE-AC03-76SF00515

Presented at Grey Exploitations in the 21st Century: The Second International Conference on Grey Literature (GL '95), Washington, DC, November 2-3, 1995

experiments and of the collaboration performing the experiment are indexed even when not explicitly part of the title page or the paper. Researchers can trace the effect of a paper on subsequent scholarship because all references to published journal articles or to numbered electronic pre–prints (e–prints) are included to form a 'citation index'. Physicists at the DESY Laboratory in Germany add extensive subject index terms from a controlled vocabulary developed for particle physics. Finally, a code is also added for every conference paper enabling comprehensive searches by conference long before the proceedings are published.

SPIRES–HEP has evolved over time into a collaboration of the particle physics community libraries world–wide. Selected libraries contribute new records to the database, others simply use the database as a local catalog by entering on–line their holdings information for the pre–prints held by their library. A number of sites around the world download significant portions for use at their institutions and others run clone copies of the database to improve response time and decrease network load for the regions they serve. Some of the collaborating institutions include DESY (Germany), Kyoto University (Japan), RAL/Durham (England), Fermilab and Caltech (USA), and CERN (Switzerland).

## 2. CULTURE AND TECHNOLOGY SUPPORT GREY LITERATURE CONTROL

The bibliographic control of high–energy physics pre–print literature has evolved over three decades from a manually produced, weekly print publication that was mailed to libraries and physics departments world wide to an interactive database that provides on–line access and hypertext links to the bibliographic data, abstract, full text, references, and citations within literally hours of the pre–print's first appearance. Some of the key technology enabling this process was invented by physicists themselves, who together with a handful of librarians with unusual vision, repeatedly pushed the database's limits to better fulfill an ideal of comprehensive and universal desktop access to the field's literature. To understand how this process developed, and how other fields can adapt the revolution it has created to their circumstances, one must have some background in the culture and tools of high–energy physics.

High–energy physics is a relatively small community with a strong tradition of international collaboration. Most of the experiments conducted are done on large instruments called detectors located at approximately a dozen accelerator laboratories around the world. Often hundreds of physicists from many countries collaborate to propose, design, build, and run these experiments which, from inception to conclusion, may last a decade. While experimentalists form large teams, theoretical high–energy physicists are scattered thinly about the globe, at approximately 3,000 university physics departments and laboratories. As early networks were established, high–energy physicists quickly recognized their utility to share the work of widely scattered collaborators and to communicate new theoretical insights rapidly amongst colleagues.[1] Researchers in this field began to rely on the precursors of the Internet to share not only electronic mail, but also software programs, data analysis, and early drafts of collaboratively written research papers.[2]

High–energy physicists began to disseminate their research to colleagues in the form of paper 'pre'– prints long before they acquired or developed the tools to disseminate these pre–prints to each other electronically.[3] They shared these research papers in advance of the paper's formal publication in the scholarly literature for several reasons. One of the first purposes was to circulate a draft of the paper amongst many collaborators who were joint authors. A second motive was to circumvent increasingly lengthy journal publication schedules which were causing delay times of up to two years between submission and publication. Theoreticians in this field publish relatively frequently. Their papers sometimes function as iterative discussions – provoking lively dialog both within their subfield and between them and experimentalists. Both at its inception and its conclusion, an experiment would be severely hampered by the long lead times of traditional scholarly journals. For experimentalists contemplating a particular research question, not knowing that another team was already engaged in a similar problem could prove expensively duplicative. And, when years of effort culminate in a publication that may radically alter the fundamental explanations of our universe, experimentalists are understandably reluctant to passively sit out a one to two year publication cycle before their results appear in print.

Researchers in this heavily compute–dependent field are, of necessity, extremely computer literate. Because their field is highly specialized and abstract, high–energy physicists have often written their own software and invented new computing programs and tools themselves. Newcomers to the field quickly develop a high level of computer literacy and a great deal of expertise – and faith – in using computing systems to do work faster and better. It is not surprising that these characteristics of computer literacy, networked communication, and international collaboration also created a demand for rapid, comprehensive, and widely accessible control of high–energy physics literature.

## 3. CONTROL OF THE LITERATURE EVOLVES

As some visionary librarians began to acquire, organize and provide access to the pre–print literature, physicists also came to recognize the value of an organized and centralized system of bibliographic control.[4] The pioneering managers of high–energy physics grey literature, Luisella Goldschmidt Clermont from CERN, Dr. Kurt Mellentin from DESY and Louise Addis from SLAC, began to collaborate early on as they collected and organized the pre–print literature. They readily shared their efforts and innovations with each other and looked to evolving computer systems to provide ever increasing levels of information management. The SLAC library received an important boost in this effort in the early 1970's through Louise Addis' early collaboration with a Stanford researcher who began studying high–energy physicists to learn how scientists communicate.[5] As a result of his findings, he spearheaded the development of the SPIRES (originally entitled Stanford *Physics* Information Retrieval System ) database management system. Louise Addis led the implementation at SLAC. Originally created to help physicists communicate electronically, SPIRES evolved into a database management system used to control the pre–print literature as well as many other bibliographic and administrative databases at numerous institutions around the world.

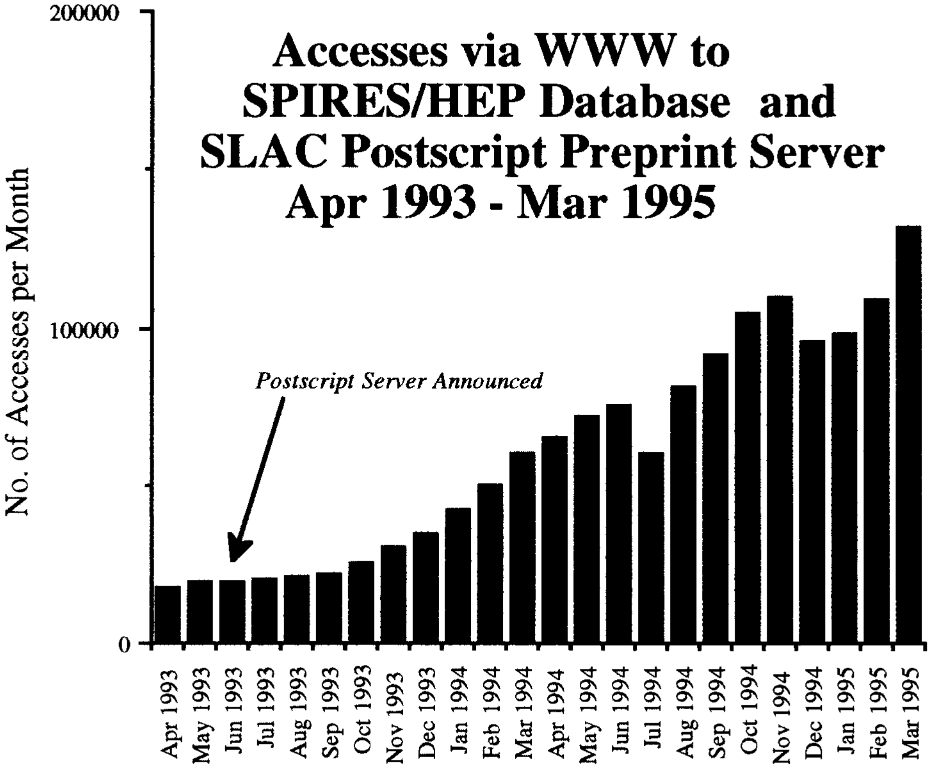
A second important development occurred in 1979 when Donald Knuth from Stanford University invented TeX, a text markup language which quickly became popular within the field of high–energy physics because it provided a high quality mathematical text using simple ASCII characters as input. The spread of TeX and its variants gave the field an electronic *lingua franca* that set the stage for the eventual electronic transmission of pre–prints. A further development occurred in 1985 when the SPIRES–HEP database became accessible world–wide as an electronic database searchable through remote access servers.[6] Eventually, this system was regularly accessed from over 662 nodes in 44 countries representing almost 5,000 non–SLAC remote users of the database. Clones of the SPIRES database management system and the HEP database began to be run in different parts of the world to better serve users in those regions. Certain libraries then began contributing original cataloging of pre–prints to HEP, while other libraries used it as a miniature 'OCLC' to show users the local library's pre–print holdings. There were many other minor technological developments and hundreds of basic improvements in the SPIRES–HEP database during these two decades, brought about through the collaboration of physics librarians at various institutions and by the synergistic interaction of librarians and physicists.

A profound change in both the way scholars communicated and in the speed with which SPIRES–HEP could make bibliographic information available occurred in 1991 when a physicist at Los Alamos National Laboratory, Paul Ginsparg, radically redesigned pre–print distribution by creating an electronic submission system.[7] Physicists could now send electronic copies of their papers written in the TeX formatting language which transmits and displays as ASCII text, to a server which stored them according to their sub–field (theory, phenomenology, etc.), assigned "bulletin board numbers", and then nightly electronically mailed lists of the prior day's pre–prints to listserver subscribers. This system of electronic pre–print (e–print) archives grew in popularity quickly. However, even with a majority of physicists ostensibly writing in the same text formatting package (TeX), the listserver subscribers experienced a significant number of problems displaying and printing the copies of the full text e–prints which they had transferred via file transfer protocol (FTP) to their computers.

The SLAC Library quickly began including the "bulletin board" number in the bibliographic record for the pre–print when it was cataloged in the SPIRES–HEP database. Soon, physicists asked if the Library could provide an easier way to view or print these papers. The Library began providing Postscript versions of each day's e–prints. After the previous day's papers were FTP'd, a TeX processing expert re–formatted each document into a viewable, printable, Postscript version. Important parts of this service were the speed and the quality control. The majority of the papers would be processed within four to six hours of their receipt. Library staff would also test each paper's viewability and printability on common computer platforms. The growing number of papers posted to the e–print archives would have quickly overwhelmed this manual system, but SLAC was soon rescued by the creative programming of Paul Mende, a physicist at Brown University, who successfully automated the TeX/Postscript routines resulting in 80% of the papers processing automatically. Library staff then worked on the 20% of problem papers and on obtaining and processing the figure files received from authors who had not posted them along with the text of their paper to the e–print archives.[8]

The development of the e–print archives coincided with another invention that changed even physicists' Internet use and paved the way for fast, user-friendly access to SPIRES–HEP and the full–text of the e–prints. In late 1991 and early 1992 the World Wide Web was invented by Tim Berners–Lee and his team at the CERN Laboratory in Switzerland.[9] The first working Web site in the United States was installed at the SLAC Library by Paul Kunz, a SLAC physicist who had obtained a copy of the Web software while at CERN, thinking it would provide a vastly improved search system for remote users of the SPIRES–HEP database. The combination of a user–friendly and rapid search system coupled with full–text access to Postscript versions of e–prints within 24 hours of their appearance on the Los Alamos e–print archives, skyrocketed the popularity of the HEP database with its users and proved so useful that within two years of its implementation, the traditional mode of remote searching from registered nodes was discontinued – most physicists were using the Web to search, view and print from SPIRES–HEP. The following graph shows the dramatic effect these improvements in technology had on database use.

**3.1 Graph showing skyrocketing database use after the introduction of World Wide Web access and links to the full electronic texts of pre–prints.**



The SLAC Library is in the middle of this revolution, adapting to changes these two inventions have caused and examining new applications that permit even more radical changes. Integrating these two new technologies into our control of the grey literature changed the nature and speed of our internal workflow. It also expanded the services we provide our users, enabling us to create new information resources and link our databases with other Internet–accessible resources. The result is a vastly expanded virtual library of high–energy physics information.

The first phase of this change enabled the SLAC Library to eliminate some manual effort and speed up internal processes with little increase in labor. Triggered by the nightly electronic mail sent to the e–print listserver subscribers, our automated processes stripped out the elements of the bibliographic record and the full abstract from papers submitted to the e–print archive the day before. This processing program added a temporary record for the item into SPIRES–HEP with a hypertext link embedded in the record to both the SLAC–held abstract and the full text of the e–print. All of this was done in the early morning before cataloging staff arrived to begin their day's work. Using an ASCII version of each e–print, the processing program put a copy of each paper's references (footnotes) into a separate database, from which they were subsequently edited, indexed, and linked to the full text of the work being referred to if it were also contained in SPIRES–HEP. From this reference list is built a citation index which is searchable both forward and backward in time and has hypertext links to the full text of documents when available.

The overnight availability of a Postscript version of most e–prints linked with a temporary bibliographic record, has had the effect of increasing users' expectations for the entire SPIRES–HEP database. Users quickly began to assume that everything else about the e–print should also be immediately available. Traditional reasons for delay – printing, mailing, cataloging, etc. – which were understandable in an analog world, were assumed to be either irrelevant or trivial in the digital landscape. Our successes with automating part of the processing raised users' expectations and led to further developments to increase processing speed and make available more quickly other parts of the bibliographic information. Even with some remarkably creative efforts, not all work can be automated successfully. In certain areas, the Library must still rely on "old fashioned" manual effort which is, of course, slower. Because of these differences in processing speeds, we have had to undertake an entirely different level of user communication and intervention.

During the decades when the Library received print versions of pre–prints and cataloged them manually, most physicists using the system understood that all the information about a particular pre–print was not completely accurate until each week's conclusion. Recently we have had to include a new level of message to early morning searchers who find that certain parts of the bibliographic information, say the footnotes, are not yet available for a particular e–print that we have received through our overnight processing. Composing this new message made us realize how these profoundly changed user expectations have altered our own standards of acceptable turn–around time. Now, when a user selects the hypertext link "*See References*" for a preprint which does not yet have the footnotes entered, a program performs a calculation to determine how many tenths of a day old the bibliographic entry for that preprint is and advises the user:

> **"This preprint is only 0.3 days old and the SLAC**
> **Library has not completed processing it yet."**

SLAC staff such as our SPIRES database manager, Dr. Hrvoje Galic, have also created new tools using our bibliographic data as well as linking other externally–available

Web resources into a more efficient information system for researchers. We have created several new directories, combining information which would have been impossible without the seamless integration that the Web permits. One directory enables a user to find information about HEP experiments and includes a bibliography of publications from that experiment from SPIRES–HEP as well as links to an experiment's own home page if one exists. Another Web page provides access to the high–energy physics home pages of all academic institutions and research organizations world–wide that have such a page available. We also provide a small program which can be implemented by these other Web page owners to create an up–to–date bibliography from SPIRES–HEP of publications about their experiments or by authors affiliated with their institutions.

We have begun to broaden the access we provide into a continuum which further links the stages of the research process in this field. If a "grey" pre–print becomes a journal article and available on a publisher's server, we provide a link in our bibliographic record to that "white" literature. Also, we are discussing with our users ways to include the research information that precedes the e–print, perhaps by linking to a collaboration's detector design technical notes or to experimental data itself.

One measure of how thoroughly our database has become integrated into the field's intellectual work is the amount of complaints we receive when users experience access problems. One evening we made a major change to the program underlying the Web forms interface we were using. The Library's main electronic mail address received over two dozen complaints via electronic mail before staff arrived at eight o'clock the next morning. The complaints continued until we could replicate and then solve the problem which, fortunately, only affected part of our user community. Our SPIRES database manager personally responded to hundreds of electronic mail complaints in the several days it took to resolve the problem.

## 4. THE CHALLENGE

The SLAC Library was uniquely poised to take advantage of the e–print archive and Web revolutions. Yet the effects of these technological inventions have flowed beyond the narrow confines of high–energy physics. The e–print archive system has already been established in other fields of scholarship and the rapid popularization of the Web has changed even the general public's use of electronic information.[10,11] Other managers of grey literature can use these inventions and the SLAC Library's experiences to expand and improve their control and dissemination of grey literature.

Not only can the availability of electronic texts enable grey literature database managers to streamline traditional work and provide information much faster, but with hypertext links, managers can create new features and enhancements that will improve service to their communities and link information in new ways. Futurists speculate that the advent of these inventions may herald an information revolution as profound as that of the printing press.[12] The presence of full–text archives and Web hypertext connections are leveling the playing field not only amongst authors and publishers but amongst categories of information. Because of their extensive experience with non–traditional information, grey literature managers are better prepared to recognize the value of these emerging information landscapes and to organize them in innovative ways.

It is not clear that in a couple of years, the distinctions among grey, 'black' and 'white' literature will even be meaningful.[13] Grey literature can now be linked to additional information that was, before this, often only accessible at a particular location or to a particular group. Connections can also be made to the other side of the spectrum, to the full electronic texts of published journal articles, review literature, and books. This technological revolution has already presented an opportunity to provide our users with enhancements that would have been unimaginable, or at least, impossibly costly several years ago.

In fact, institutions and individuals with experience in grey literature collection, organization, and access are uniquely qualified to bring order to the chaos of Internet information resources. Working within the world of grey literature, one must often forge new solutions and operate with little traditional infrastructure or precedent. This is partly because grey literature has been traditionally ignored by established information managers such as libraries and publishers. Grey literature knowledge workers have, of necessity, developed intellectual characteristics that can help them migrate their efforts successfully into the inchoate electronic environment. They bring to this challenge a flexible, questioning approach that recognizes the value of alternative forms of scholarship and information.

Organizers of grey literature have to use unconventional means to learn about and acquire their material which often exists in formats that have been difficult to obtain, handle, and disseminate. They typically maintain a closer connection to their authors and producers than is customary when working with the traditional scholarly work of a field. With these closer connections to their user community, grey literature managers are able to learn more about how information is used in their fields and to find out from their users more about their unmet information needs. Because they can identify and value the many threads of information produced by their user communities, grey literature managers are uniquely able to weave the disparate strands of Internet information in their fields into an intelligible and useful whole.

Using as a model the revolutions experienced by the SPIRES–HEP database, one can see that technological advances, coupled with the unique combination of skills and knowledge developed through working with grey literature combine to present managers of this 'fugitive' information with an unprecedented opportunity to play an new and more central role in managing electronic information. Grey literature managers are facing a challenge that they are uniquely able to meet. They have the opportunity to play a vital role in organizing their field's traditional and non–traditional information – whether black, grey or white – into an organized and coherent "virtual library" for their researchers and users.

## 5. ENDNOTES

1. R. Taylor, "Brave New Internet: Changes in Culture and Technology are Driving the Internet into the Next Millennium," *Internet World* vol. 2 no. 6 (Sept. 1994): 36–42.

2. P. Doty, et al., "Scientific Norms and the Use of Electronic Research Networks," in *ASIS '91: Proceedings of the 54th ASIS Annual Meeting*, Wash. D.C., October 27–31, 1991. Medford, N.J.: Learned Information, 1991: 24–38.

3. L. Addis, "SLAC Library Monitors Underground Physics Press," *SLAC News* (June 1971): 2–3.

4. L. Addis to P. Kreitz, email describing early participants on preprint collaboration, Nov. 1995.

5. Edwin B. Parker, Institute for Communication Research, Stanford University, was the principal investigator for the NSF grant that began SPIRES. See Stanford University, *Design of the Stanford Public Information Retrieval System (SPIRES II),* vol. 1, 2nd ed., revised, July 1973.

6. George Crane, a physicist and SPIRES programmer, has been credited with creating the popular Q–SPIRES remote user access system to SPIRES.

7. P. Ginsparg, "First Steps Towards Electronic Research Communication," *Computers in Physics* 8:4 (Jul/Aug. 1994): 390–396.

8. The SLAC Library created Postscript versions of the e–print papers until Summer 1995 when this effort was successfully migrated to the Los Alamos E–Print Archives. Currently, an author–submitted paper is automatically converted to Postscript upon submission. If the processing fails, the author is automatically notified by email and advised to check his word–processing markup for errors before re–submitting the article to the Archives. Currently the SLAC Library continues to collaborate in this process by receiving, processing and posting the figure files for those authors who are unable to submit them with their e–print texts to the Archives.

9. T. Berners–Lee, et al., "World–Wide Web: The Information Universe," *Electronic Networking* 2 (Spring 1992): 52–58.

10. See URL http://xxx.lanl.gov under "Some related and unrelated servers".

11. C. Bournellis, "Internet '95: 'The Internet's Phenomenal Growth is Mirrored in Startling Statistics,' " *Internet World* v.6, no.11 (Nov. 1995): 47–48, 50, 52.

12. A.C. Schaffner, "The Future of Scholarly Journals: Lessons from the Past," *Information Technology and Libraries* vol.13, no.4 (Dec. 1994): 239–247.

13. W. Winograd & R.N. Zare, Editorial: " 'Wired' Science or Whither the Printed Page?" *Science* 269 (Aug. 4 1995): 615.