SLAC-PUB-8024
April 1999

# The Role of the Library in the Scholar's Electronic Research Environment

Patricia A. Kreitz,
Stanford Linear Accelerator Center, Stanford, CA 94309, USA

**Abstract**
This paper examines how new systems of electronic information storage and delivery allow libraries not only to improve the traditional ways in which they have participated in the scholarly research process, but also to expand that participation well beyond traditional library functions. The effect of these new systems on libraries is most dramatic in the field which invented the Web, high-energy physics. The experiences of high-energy physics libraries over the past four years show the pivotal role libraries can play in creating a comprehensive electronic research environment.

*Contributed to the*
*Digital Libraries and Information Services for the 21st Century Conference,*
*ICPR, Seoul Korea, September 10 – 13, 1996*

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

INTRODUCTION:

The research process by which a scholar solves a problem, invents new ideas, or tools, or creates new ways of explaining the world, involves a complex cluster of thought and action which is, in the majority of scholarly fields, highly information dependent. While libraries have traditionally been the most comprehensive source of information supporting the research process, no single library can fully supply all the scholar's research needs due to limitations on the library's budget and physical space, as well as unavailability of some materials via normal library lending methods. Furthermore, a significant amount of the information a scholar uses exists outside of the traditional area of library collecting strength -- books and journals from the mainstream publishing process. In most cases, scholars' multi-dimensional and multi-threaded information needs are as often satisfied outside of the library's walls as within them.

In 1945 Vannevar Bush imagined the ideal scholar's research tool. His idea was to improve on the traditional limitations of libraries and provide immediate access to all existing information as well as a convenient way to incorporate new information as the scholar encountered or developed it. Dr. Bush envisioned this tool, which he named Memex, as a combination microfilm scanner and reader, holding the vast content of the world's libraries in miniaturized format, and an automated indexing and retrieval system. This was the idealized desktop library -- delivering information into the researcher's hands when needed and permitting both serendipitous browsing and structured retrieval of the billions of pages of information it contained. Bush, himself a physicist, believed such a tool would "extend the powers of the mind" (1) helping scholars discover new ideas, encounter old information in new ways, and communicate and cooperate more effectively with colleagues.

In the past five years, the scholarly world has witnessed a combination of technical inventions and improvements that bring us much closer to Bush's vision. However, unlike Bush's description in which libraries play a minor role -- providing the raw material for this vast miniaturized repository -- libraries are building on their experience and knowledge to create better ways to connect scholars with the variety of information sources available in the electronic research environment. (2)

Innovations such as cross-platform interoperability of software, word processing, printing and viewing systems, digitization combined with optical character recognition systems, the explosion of Internet file transfer and communication, and the ubiquitous and user-friendly World Wide Web, have revolutionized desktop access to information. While the contents of the world's libraries do not yet exist at the fingertips of every scholar, libraries are playing a key role in linking scholars' computers to information that is electronically available, no matter where in the world it resides as well as partnering in new ways to expand the information available electronically. In this electronic virtual environment, the libraries are improving traditional services as well as supporting new areas of the research process.

While there are many critical activities and services a library performs to meet the needs of its user communities, three are particularly affected by the electronic information environment: collecting information; organizing access systems, and providing user assistance and support. A virtual environment allows libraries to improve and extend these functions by providing efficient access to a broader spectrum of information than an individual library can afford to provide alone, by delivering information when, where, and how the scholar needs it, by providing assistance and tools that enable the scholar to work with available information in new ways, and, finally, by linking the scholar to other, non-library parts of the ‘virtual’ research effort.

## HIGH-ENERGY PHYSICS INFORMATION SYSTEMS:

A practical example of how both researchers and libraries have participated in the rapidly changing electronic information environment can be seen in the changes that have taken place in the field of high-energy physics. In the past five years, this field has been at the forefront of inventing new ways of sharing information electronically. Innovation has been driven by practical necessity, by characteristics of the field itself, and by the changing nature of high-energy physicists’ work. This has resulted in profound changes both in the way that work is now being done compared to five years ago and in the nature and scope of information available to the high-energy physics community.

Theoretical high-energy physicists are scattered thinly about the globe, at approximately 3,000 university physics departments and laboratories. Most experimental research in this field takes place at less than a dozen particle accelerators located at scientific laboratories around the world. Experiments conducted at these accelerators are lengthy, often taking ten years from proposal through design, construction, operation, and data analysis. No single institution or group can expect to have all the skills required to design and run such complex accelerators or experiments, and no one institution or country can afford to finance these research efforts alone. In fact, teams of high-energy physicists working on a particular experiment may include 500 people from 100 countries. Because of length of time these experiments last, and the world-wide distribution of the participants compared to the concentration of the experimental facilities, the field has evolved, of necessity, into a highly collaborative and international effort with a critical need to share information and conduct work collaboratively over the Internet. The World Wide Web was first invented as a means for these large international collaborations to work together ‘virtually’.

Using the Web, a physicist working in Asia can participate in real-time analysis of experimental data taken overnight from an accelerator running in Hamburg. Performing the type of analysis that formerly could be done only on-site, he analyzes the run data and posts the results in shared Web space within 24 hours of the original data-taking where it is available to colleagues world-wide. Small groups of experimentalists located at different sites can design detector components, discus problems and implement decisions, with virtually no need to meet. Threaded discussion lists permit remote designers to stay abreast of progress in the design and testing of other detector components that must integrate with their sub-system. Research results are written much more rapidly and collaboratively by authors working remotely sharing a ‘master’ copy of the text that tracks each author’s changes and comments. Recently a physicist gave two presentations at a collaboration meeting in Germany from his office at the Stanford Linear Accelerator in California, USA. The Web has changed how physicists work in fundamental ways. Broadening the scope of participation to remote collaborators, not only stretches available funding but also brings more minds to bear on a particular problem.

This field is also heavily compute-dependent. Increasing speed and complexity in both hardware and software are necessary to handle the modeling, control, and analysis systems required both for accelerators and detectors and by theoretical physicists forecasting and analyzing experimental results. Because their field is highly abstract, its language, conclusions, and requirements are fairly incomprehensible even to most other scientists. As a result, high-energy physicists have often written their own software and invented or rapidly adapted new computing systems to meet their needs. Newcomers to the field quickly develop a high level of computer literacy and a great deal of expertise (or at least faith) in using computing systems to do things faster and better.

One of the consequences of this computer dependence and expertise was the invention of a database management system, originally called SPIRES (originally, named the Stanford Physics Information and Retrieval System), which was created to help physicists communicate electronically and became the system used to impose bibliographic control on the field's scholarly literature. In 1985 the SPIRES High-Energy Physics (HEP) database became accessible world-wide through remote access servers. Although Web access to the database has replaced this system, at the peak, it had 662 nodes in 44 countries, representing almost 5,000 non-SLAC remote users of the database. Like many of the early database management systems built to run in a large main-frame environment, this system was hierarchical. However, since its primary purpose was to enable physicist to share information--email, text, data, etc., it incorporated from the initial design, a number of features that have been built into more recent relational and inter-operable database management systems. As the HEP database evolved, physicists in the field came to rely on it increasingly to collect, control, and disseminate their literature. Physicists continue to lend their advice and computer expertise to support its further development. Once Web access to SPIRES-HEP was installed, remote searching skyrocketed, eventually reaching 150,000 remote searches per month.

Coupled with these traditions of collaboration and computer sophistication are additional characteristics that created a fertile seedbed for innovative information sharing. High-energy physics is a fundamental science, with no inherent or intentional practical applications. The "product" is simply knowledge and the most common method of communicating that knowledge is through scientific papers. Theoreticians in this field publish relatively frequently. Their papers almost function as an iterative discussion, a dialog both within their subfield and between them and experimentalists that is severely hampered by the lengthy publication schedules of traditional scholarly journals. For experimentalists, years of labor, hundreds of researchers' effort, and sometimes millions of dollars, culminate in a discovery that may radically alter our understanding of the universe. While neither group can afford to let their writings languish in a publication process that may take two years before a submitted article appears in print, experimentalists are particularly affected by these long delays. An experimental collaboration must know rapidly about their colleagues' work to take advantage of critical technological breakthroughs and to avoid costly mistakes or duplicative work. As a response to this need to share their publications quickly, high-energy physicists began informally distributing advance preprints of articles submitted for publication close to four decades ago.

This desire to share publications quickly prompted a high-energy physicist to invent a system of electronic archives to which authors could submit an electronic version of the full text of their papers and other interested subscribers could receive email notification of the paper's submission. Subscribers to the listserve could then use Internet communication and transfer tools to download a copy of a particular paper, reading it within 24 hours of its original submission (called 'posting') to the electronic archives. In the past four or five years, this invention has replaced the old system of mail distribution of print copies for more than 75% of preprints in this field and has rapidly expanded beyond high-energy physicists to other areas of physics and to other fields outside of physics. FIG. 1 shows how ubiquitous links to full text versions of a preprint have become.

In physics, it has democratized scholarship. (3) Now, a theorist in a remote country working at an institution which has comparatively little resources to publish and distribute that physicist's work, can submit his or her paper to the electronic preprint archives sharing it with the community within hours of its completion. Formerly, differences in the publishing quality of preprints and in the breadth of distribution created a subtle hierarchy with preprints from the (relatively) well-supported institutions being physically more professional looking and being more widely available than those from poorer institutions. Now, the availability of an electronic text has leveled the playing field and a work may be evaluated solely on the merit of the ideas it contains. The electronic archives have had another, broader effect on the publishing industry. Researchers have speculated that this invention has, at the very least, accelerated the traditional publishing industry's migration from print to electronic publications and may one day entirely replace the traditional scholarly journal. (4, 5)

## THE DIGITAL HIGH-ENERGY PHYSICS LIBRARY:

These three critical developments: the Web, SPIRES, and electronic preprint archives, have been key elements in the development of digital high-energy physics libraries. The SLAC library was the first functional Web site in the United States, serving as the front end to the HEP database. (6) This user-friendly and rapid search system proved so popular that within two years of its implementation, the former method of remote searching from registered nodes was discontinued. The following graph shows the dramatic effect these improvements in technology, and thus service, had on the frequency of use of the SPIRES-HEP database. FIG. 2

The early development of the SPIRES database management system permitted high-energy physics libraries world-wide to enter into collaborations similar to those pursued by experimentalists.  Many libraries located in different countries work together to build and improve the suite of databases maintained on SPIRES and used by the high-energy physics community.  Using the example of the flagship database, SPIRES-HEP, some institutions add new records, others simply use the database as a local catalog by entering online their holdings information for the preprints 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).

SPIRES-HEP currently contains over 300,000 bibliographic entries for high-energy and related particle physics preprints, 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 the full online text of the preprints 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 preprint).  The names of experiments and of the collaboration performing the experiment are indexed even when not explicitly part of the bibliographic data or the actual paper.  Researchers can trace the effect of a paper on subsequent scholarship because all references to published journal articles or to numbered electronic preprints (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 by and for the field.  Finally, a code is also added for every conference paper enabling comprehensive searches by conference long before the proceedings are published.

The SPIRES database management system permits flexible and rapid response to changing information needs or new technical innovations.  When the Web was first brought to the SLAC library, systems librarians and physicists quickly established that SPIRES could write out HTML directly, without any intervening programming required.  Thus simple recognition and inputting programs were written to identify the ‘http’ address and automatically add it as an element in the bibliographic record and write it out as a hypertext link when the record is displayed.  SPIRES also permits data from one database to be passed to another file and indexed back to the original information from that file.  With this feature, we are able to automate the checking of incoming references taken from an author’s paper against the SPIRES-HEP database’s holdings.  If found, a hypertext link is added automatically to the SPIRES-HEP record and to the full text of the cited paper, if it is available.

The advent of the electronic preprint archives has significantly accelerated the SLAC library’s internal technical processing.  Once the full text of papers began to be widely available overnight, researchers expected the SPIRES-HEP database to be as current as the electronic archives.  Collaborating with physicists who helped to write the necessary software, SPIRES library programming staff developed systems that integrated this information into our database overnight.  Triggered by the nightly email send to the e-print listserve subscribers, our automated processes now strip out the elements of the bibliographic record and the full abstract from papers submitted to the e-print archive the day before. This program adds a temporary record for the item into SPIRES/HEP with a hypertext link embedded in the record to both to the SLAC-held abstract and the full text of the electronic preprint wherever it is available. All of this is done in the early morning before cataloging staff arrive to begin their day's work. From the ASCII version of each electronic preprint, we strip references, index them and automatically create hypertext links to the full text of the work referenced if it is also contained in SPIRES-HEP. This reference list permits us to build 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 near-instantaneous availability of an author's electronic preprint has had an unpredicted effect on the SPIRES-HEP database. Because the electronic preprint is now available overnight, users assume that everything else about the paper should also be immediately available. Traditional reasons for delay -- printing time, slow physical distribution, the library's internal processing, including the time it takes to check-in, catalog and edit bibliographic records, etc., which were reasonable to users operating in an analog world, are imagined to be trivial by the user operating in a digital environment. Now, library users complain if a paper is not full-text accessible at 9AM on the morning after it was submitted to the e-print archives. In fact, the SLAC Library has developed a reputation for providing bibliographic information and full text links so quickly that when the system appears to fail, a significantly larger proportion of users are vociferously upset and an entirely different level of user communication and intervention must be provided.

Because our automatic systems for handling authors' footnotes is not completely perfect, the references are not always hypertext links when the record first appears that morning in the database. Normally, the electronic preprint archives provides users of their full text preprints clickable links to SPIRES-HEP for these references. We have had to ask their systems manager to add a note explaining that the SLAC library still working on the footnotes for those papers where the automatic citation processing has not worked. The message given by the electronic preprint archives located at Los Alamos National Laboratory, now reads: "This preprint is only 0.08 of a day old. The SLAC Library has not had time to process it fully yet."

The popularity of the Web in high-energy physics has enabled the SLAC library to provide connections between and among information that was only imagined previously. As mentioned above, we can now provide links among the papers presented at a conference and to the additional information about that conference in our CONFERENCES database. Additionally, we can create a link from the bibliographic record for the published conference proceedings back to the papers so researchers can print directly from their terminal a copy of a particular presentation rather than manually copying that section from the printed book. We are partnering with scholarly societies and publishers to provide advance notice if papers have been accepted for publication and links to the electronic full text of those papers on the publishers' servers.

We are also working both to provide new services and to promote new and better information on the Web to support high-energy physics researchers. Collaborating closely with physicists, we have developed the ability for a particular experiment to provide, from the experiment's home page, a comprehensive search of the SPIRES-HEP database for just those publications about the experiment's results. Using these search algorithms, an experiment need not maintain its own bibliographic database but can rely on the currency and comprehensiveness of SPIRES-HEP.

When the electronic preprint archives first began, the preprints submitted were only available to view and print in their native word processing packages--usually a customized version of TeX. Users complained that they spent enormous amounts of time trying to access these documents. The SLAC library, in collaboration with DESY, downloaded the papers each day and converted them to Postscript, a more commonly accessible viewing and print format. After this almost three year demonstration project had proven its value to users, and a number of physicists had helped write the code to automatically convert many of the papers from TeX to Postscript, the system manager of the electronic preprint archives became convinced of its usefulness to the user community and implemented the automatic TeX to Postscript programs at the point at which an author originally submits a preprint. Additional efforts have included working with key physicists to support the American Physical Society's migration to electronic preprints, and collaboration with the Particle Data Group to augment their publication with additional information we can provide and to link the SPIRES-HEP database to the online literature reviews available in the Web version of their Review of Particle Properties.

Because of these innovations, high-energy physics libraries have been able to reduce and streamline internal workflow, reducing data entry in some cases from weeks to hours. We have been able to improve connections between the library's own databases of information as well as improve connectivity and effort among the high-energy physics libraries contributing to SPIRES-HEP and the suite of other databases we maintain for the world-wide community. We have also been able to play an important part encouraging other information providers to take an active role in providing digital information via the Web and to enhance and improve the information being offered to support scholarly efforts in this field. The research environment available to the high-energy physicist has become an integrated, electronic resource linking research with ongoing thought, data with design, analysis with review. FIG. 3

## CONCLUSION:

As shown by the experiences of high-energy physics libraries during the past five years, libraries can play a significant role in making their collections and services available to the scholars' desktops. While this function is critical to future success in the electronic information environment, libraries are also able, because of their background, experience and mission, to play a key role in promoting and organizing other scholarly support information that is becoming Web-accessible. In order to do this successfully, libraries must redefine their mission beyond what may appear to be the goal of digitization of resources to one of manager of the electronic scholarly research environment.

Libraries must move their goals and vision beyond the provision of digitized electronic texts and a Web page of links to other resources. To create the truly integrated scholar's desktop that Vannevar Bush imagined and most scholars would define as the ideal, libraries must use the skills they have developed in dealing with a wide variety of information providers (scholarly societies, for profit database vendors, trade publishers, etc.) into dealing with both traditional and new information providers in the electronic environment. They must use the skills they have developed through collection building to build a more comprehensive, systematic and useable structure, which connects researchers to the cacophony--or treasure-- of electronically-available information.

Libraries must also build on their expertise in providing users efficient and effective online catalogs and user-friendly access to commercial databases to provide the researcher better interfaces to the multiplicity of databases and 'virtual' information that will continue to burgeon as the Web evolves. As part of this effort, there should be an increased emphasis on the development of a master bibliographic record for each information entity. The MARC format provides a flexible foundation for what will be needed in the future, but librarians need to begin thinking about the kinds of text and version related information proponents of SGML have been advocating. New data elements need to be invented which communicate critical electronically-necessary information to the professional librarian as well as the researcher using the material. (7) Such information might include format, digitization process, provenance, version, copyrights and permissions to distribute electronically. Unlike traditional books, digitized information may need to be supplied with software applications that efficient, platform independent viewing, manipulation, and printing.

A virtual environment may be the fulcrum that truly permits libraries to specialize in areas of strength or expertise and to concentrate their resources and efforts on fewer programs, research fields, and needs. A scholar accessing the library's resources via the Web does not care if the information needed is physically housed on servers the library owns or if the library is merely a connecting link -- assuming the connections are reliable and sufficiently robust -- to the full-text information which may exist elsewhere. Libraries can expand current efforts such as cooperative collection development agreements to apply to cooperative full-text coverage of different research areas. (8) This will allow each library's resources to be used to provide a more comprehensive and better-organized collection in narrower areas.

Libraries have used their professional knowledge and skills to meet the challenges of the electronic information world. We have heard many examples of how they are improving the traditional ways in which they have participated in the scholarly research process. With vision and energy, they will be able to play a pivotal role in creating a comprehensive electronic research environment.

REFERENCES:

1. Bush, V. "As We May Think", Atlantic Monthly, July 1945 (at http://www.isg.sfu.ca/~duchier/misc/vbush/vbush.html reproduced with permission)
2. Jones, P. "Collaboration Station: Why Scholarship is More than Scanning, E-Mail, or even Information Retrieval". Paper presented at ALA/ACRL/STS program "Leaders on the Web Trail". New York: July 7, 1996.
3. Private communication, Dr. Marek Karliner, July 1995 (marek@vm.tau.ac.il).
4. Durran, M. "Terminal Future for Paper Journals?", Science, Feb. 1996 (9:2).
5. Schaffner, A.C. "The Future of Scientific Journals: Lessons from the Past", Information Technology and Libraries, Dec. 1994 (13:4).
6. Personal communication, T. Berners-Lee to P. Kunz, Spring 1996.
7. Hillman, D.I. "Practical Aspects of SGML in Libraries". Paper presented at ALA/ALCTS Preconference: SGML. New York: July 5, 1996.
8. Balas, J. "Online Treasures: Building Virtual Libraries", Computers in Libraries, Feb. 1996 (16:2).

FIGURES:

1. Database Record Showing Number of Links to Full Text of Preprint
2. Web Searches of SPIRES-HEP 1993-1995
3. Using the Web to Link High-Energy Physics in New Ways

Database: **HEP (SPIRES-SLAC)**
Search Command: **FIND AU PESKIN, M**
Result: **81** documents found (25 latest displayed):

1) **THE EXPERIMENTAL INVESTIGATION OF SUPERSYMMETRY BREAKING.**
By Michael E. Peskin (SLAC). SLAC-PUB-7133, Aug 1995. 33pp. To be published in the proceedings of Yukawa International Seminar '95: From the Standard Model to Grand Unified Theories, Kyoto, Japan, 21-25 Aug 1995.
e-Print Archive: **hep-ph/9604339**

References
Abstract and postscript from Los Alamos e-Print server
(and link to Trieste e-Print mirror)
Postscript Version from SLAC Tech-Pub archive
Conference Info

4) **EXOTIC NONSUPERSYMMETRIC GAUGE DYNAMICS FROM SUPERSYMMETRIC QCD.**
By Ofer Aharony, Jacob Sonnenschein (Tel Aviv U.), Michael E. Peskin (SLAC), Shimon Yankielowicz (CERN). SLAC-PUB-95-6938, Jul 1995. 37pp.
Published in *Phys.Rev.D52:6157-6174,1995*.
e-Print Archive: **hep-th/9507013**

References
Keywords
Citation Search
Abstract and postscript from Los Alamos e-Print server
(and link to Trieste e-Print mirror)
Scanned images of the preprint, from KEK Library

**Fig 1**. Database Record Showing Number of Links to Full Text of Preprint

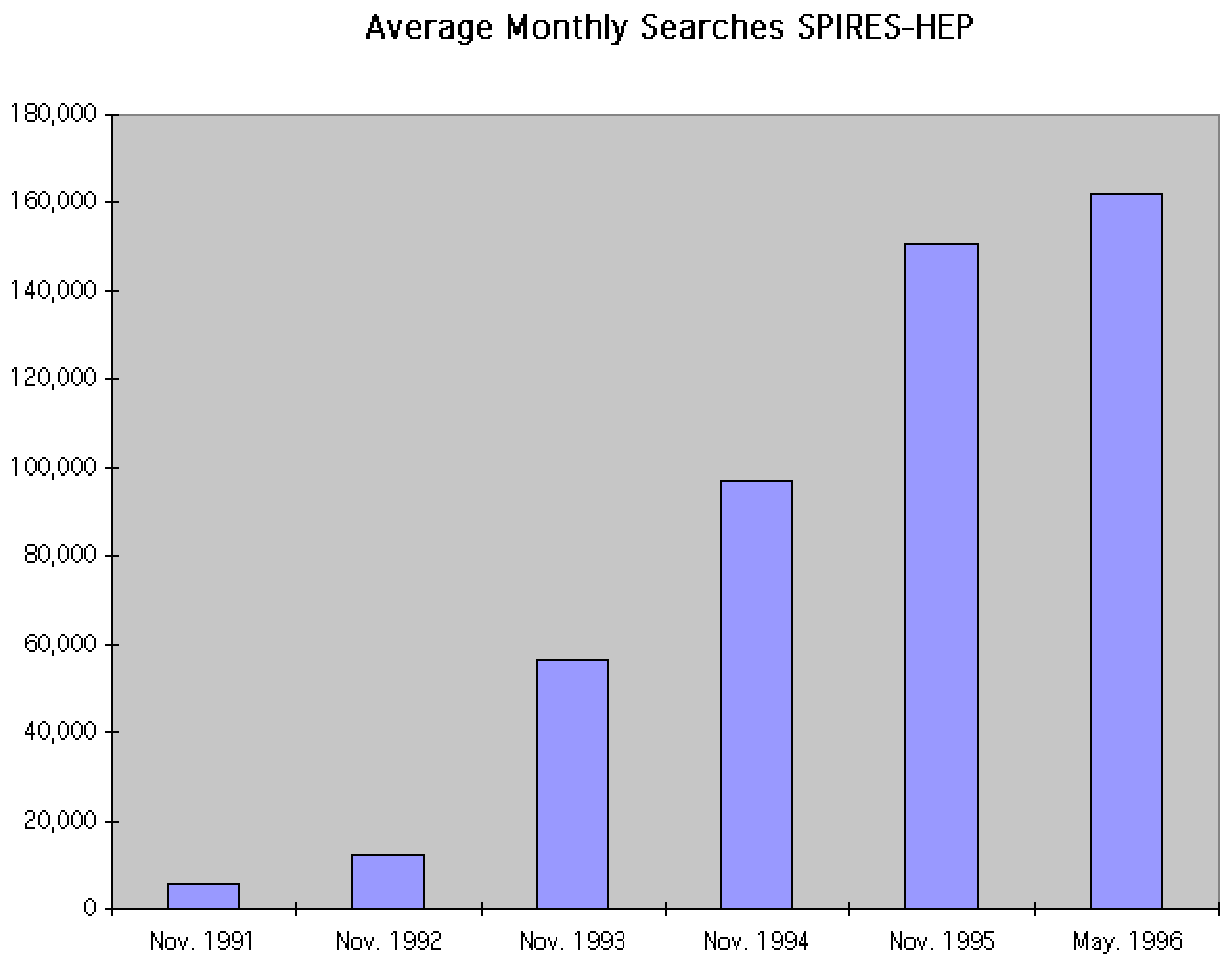


**Fig 2.** Web Searches of SPIRES-HEP 1993-1995

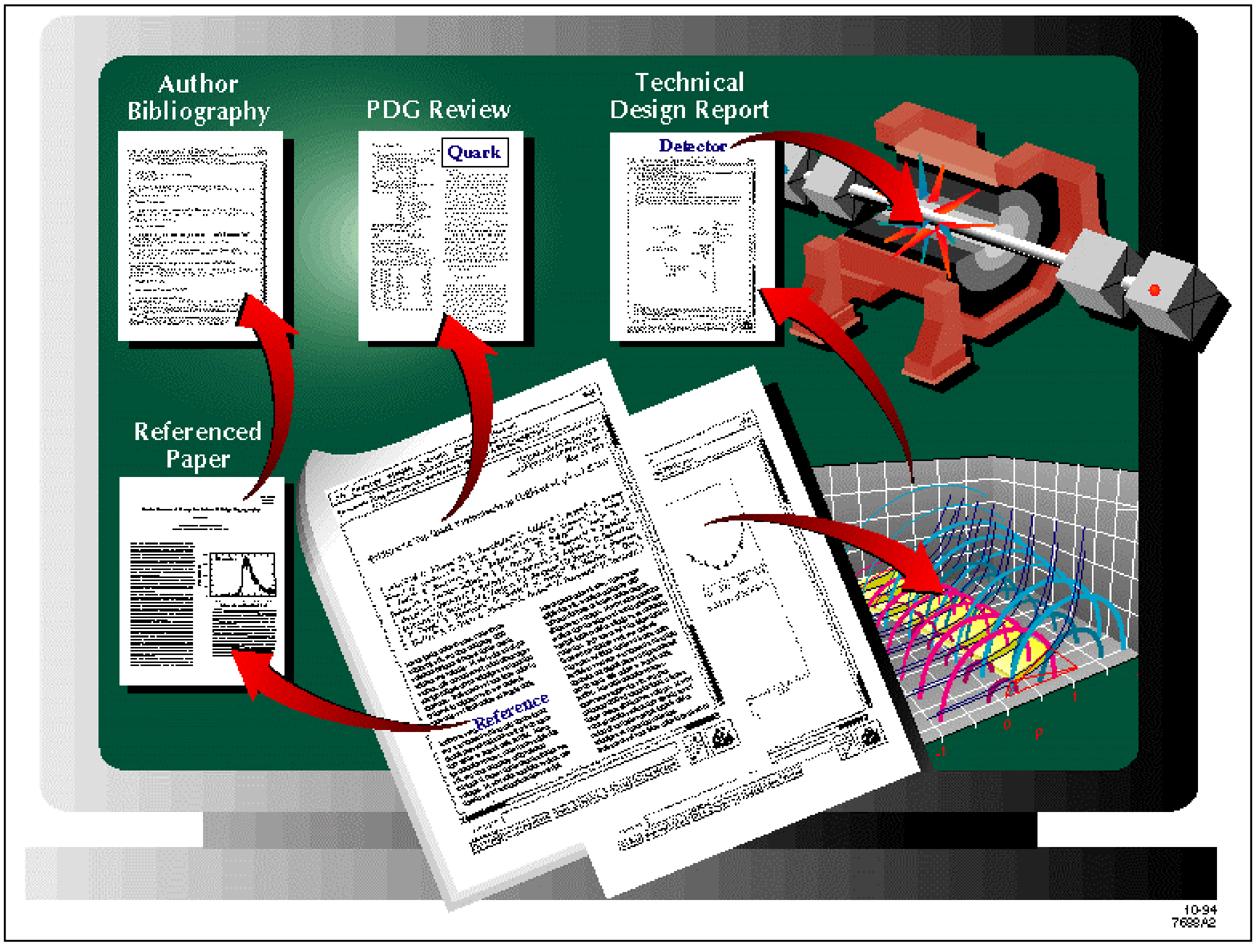


**Fig 3.** Using the Web to Link High-Energy Physics in New Ways