SLAC–PUB–7636
April 1998

First in the Web, But Where Are the Pieces?*

Jean Marie Deken

Stanford Linear Accelerator Center
Stanford University
Stanford, California 94309 USA

Invited talk presented at the Society of American Archivists Annual Meeting,
8/27/97—8/31/97, Chicago, IL, USA

* Work supported by Department of Energy contract DE–AC03–76SF00515.

A BEGINNING

This is a difficult and intimidating topic to address, attempting to preserve even a part of the World Wide Web, because there is too much material being generated on the World Wide Web for any one person to adequately research even a major portion of it, and because new developments in this medium are occurring at breath-taking speed. But one must begin. The beginning, for me, was in 1996, when I came to the Stanford Linear Accelerator Center, a national basic research laboratory operated by Stanford University under a contract from the United States Department of Energy. SLAC conducts research that probes the structure of matter at the atomic scale with x-rays, and at much smaller scales with electron and positron beams. [1] When I first arrived at SLAC, I was introduced around to the staff as the new archivist, and as people cast about for something to SAY to an archivist, the fact that SLAC was the "first World Wide Web site in the United States" was frequently mentioned (along with mentions of SLAC's past Nobel-prize-winning physics research and current high-energy physics and synchrotron radiation experiments). As I became more familiar with the SLAC archives, I naturally sought out documentation of all of the "milestones" that had been mentioned to me, including SLAC's U.S. primacy in the World Wide Web. My initial forays into the SLAC archives yielded absolutely no documentation of the progress of the World-Wide Web at SLAC. This is not at all surprising, since the Web itself is only 6 years old. [2] There really had not yet been time for the Web to become old enough for anyone to begin to think about documenting the "early days." Also, the Web is a digital phenomenon, existing in an electronic environment. Although SLAC does have an "archive" of electronic data tapes, in addition to a collection of backup tapes, none of these is either in the physical custody nor under the intellectual control of the SLAC Archives and History Office

As 1996 progressed, however, and the Web became an even more widely-dispersed cultural phenomenon, its size and importance at SLAC and in the world at large continued to grow exponentially. There began to be more and more written and said about the World Wide Web in the technical and popular media. At this point, some of the people involved in the early days of the Web — at SLAC and elsewhere — began to get serious about gathering and preserving historical documentation. [3] Also, documenting the history of the Web at SLAC became an issue when there began to be discrepancies in press accounts of the history of the early Web that slighted or ignored SLAC's role. There was one story, in particular, published in a U. S. science laboratory's newsletter that appeared to state that the World Wide Web had made its U.S. debut there.... Reaction at SLAC was swift, and internal e-mails flew back and forth for several days afterward.[4] Why does it matter? It is too early to tell, in my opinion, whether either the Internet or the World Wide Web is the cultural watershed, (or the moral wasteland) that various pundits claim. Yet the World Wide Web does matter to the SLAC Archives and History Office for two very important, and related, reasons. The first reason is that the early Web at SLAC is historically significant: it was the first of its kind on this continent, and it achieved new and important things. The second reason is that the Web at SLAC --in its present and future forms — is a large and changing collection of official documents of the organization, many of which exist in no other form or environment. As of the first week of August, 1997, SLAC had 8,940 administratively-accounted-for web pages, and an estimated 2,000 to 4,000 additional pages that are hard to administratively track because they either reside on the main server in users directories several levels below their top-level pages, or they reside on one of the more than 60 non-main servers at the Center. A very small sampling of the information that SLAC WWW pages convey includes: information for the general public about programs and activities at SLAC; pages which allow physics experiment collaborators to monitor data, arrange work schedules and analyze results; pages that convey information to staff and visiting scientists about seminar and activity schedules, publication procedures, and ongoing experiments; and pages that allow staff and outside users to access databases maintained at SLAC.

So, when SLAC's Archives and History Office begins to approach collecting the documents of our World Wide Web presence, what are we collecting, and how are we to go about the process of collecting it? In this paper, I discuss the effort to archive SLAC's Web in two parts, concentrating on the first task that has been undertaken: the initial effort to identify and gather into the archives evidence and documentation of the early days of the SLAC Web. The second task, which is the effort to collect present and future web pages at SLAC, will also be covered, although in less detail, since it is an effort that is only now beginning to take shape.

THE EARLY WEB: ASSEMBLING THE FACTS

One of the first tasks in the enterprise of documenting the early WWW has been to gather together what is already recorded about the Web at SLAC, and to assemble it into some kind of order. Through this process, the role of individuals who were instrumental in the adoption and development of the Web at SLAC can be documented, and the basic facts, the "who, what, when, where and how" of the World Wide Web, duly recorded

WHO

As everyone undoubtedly knows by now, the World Wide Web was invented at CERN, the European Laboratory for Particle Physics, by Tim Berners-Lee [5] based, in part, on a notebook program he had written in 1990 called "Enquire-Within-Upon-Everything." [6] What is not widely known at all, however, is that SLAC Physicist Paul Kunz brought word of the World Wide Web's existence to SLAC in September, 1991, when he returned from a meeting at CERN. Kunz had immediately seen the possibilities of the Web for streamlining access to a very popular high-energy physics database maintained, in part, by SLAC's Library. SLAC staff quickly saw the value of Kunz' proposal, and work to bring this new phenomenon to SLAC began. Technical difficulties ground the work to a halt soon afterwards, however, and it wasn't until Thursday, December 12, 1991 that the first WWW server at SLAC was successfully installed.[7] Soon afterward, George Crane provided an interface between the SLAC WWW server and SPIRES-HEP, the high-energy physics database.[8] (See Figure 1.) After December of 1991, though, the cast of characters involved with the Web at SLAC widened and branched out. At first there was a very informal group that met under the sponsorship of the Library, and called themselves the WWW Wizards.[9] Participation in Web activity by other staff also flourished. Individuals, scientific collaborations and departments within SLAC, including the Stanford Synchrotron Radiation Laboratory, which became part of SLAC in 1992, mounted home pages on the Web in what quickly became a vast array of offerings covering the whole range of activities undertaken at the Center

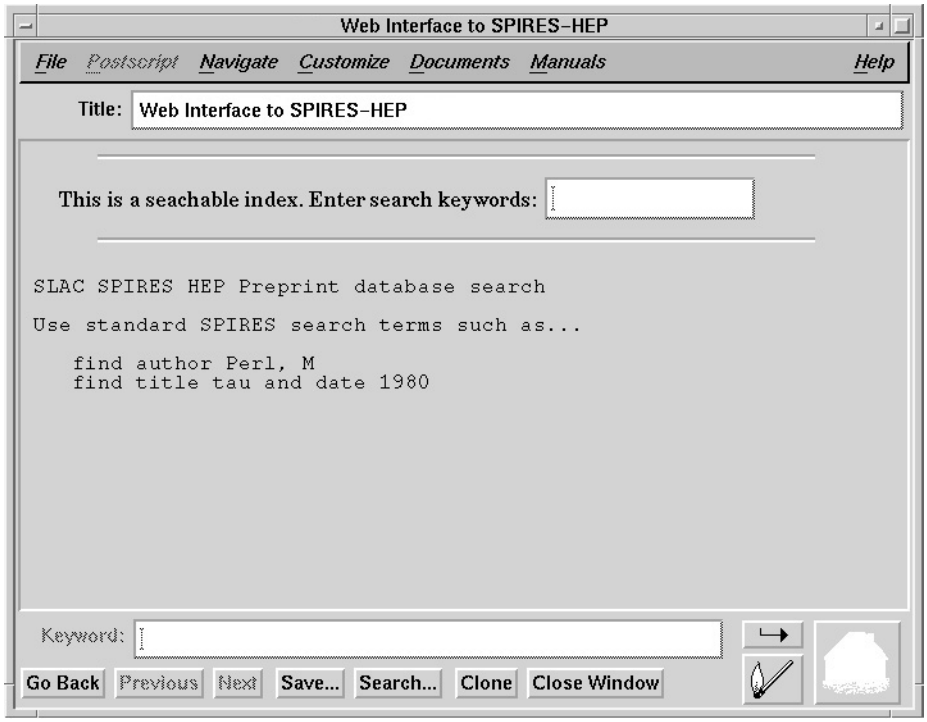

Figure 1

WHEN

The very early days of the World-Wide Web at SLAC date from 1991 to 1994.[10] This is the period that the SLAC Archives and History Office is attempting to reach back and capture, while a separate effort is being made to appraise and arrange for the orderly transfer of any permanently-valuable present and future Web documents. In the beginning, World Wide Web efforts at SLAC had no official status in the organization. This changed some time in 1994, when an ad hoc group called the SLAC WWW Technical Committee was formed to discuss and address technical, "as opposed to stylistic, aesthetic, content or policy" issues. [11] This was followed by the formation of a "SLAC WWW Users Group." Then, in 1995, a World Wide Web Coordinating Committee was established by the Associate Directors' Committee on Computing, "to provide SLAC web authors, groups and committees with policy, guidance, standards, and further support." [12]

WHAT

One of the first Web events at SLAC was a revolution in the delivery of bibliographic information. Since 1974, the SLAC Library had participated in providing SPIRES-HEP, a 300,000 record bibliographic database, to the world particle physics community via the Internet and through clone sites in Europe and Japan. The 1991 introduction of the e-print archives at Los Alamos National Lab, coupled with the World-Wide-Web from CERN, suddenly made it possible for the Library to expand its service and to provide easy linkage between bibliographic database records and the actual full-text of papers.[13] A further improvement was realized when SLAC physicist Tony Johnson introduced "MidasWWW," the first graphical Web browser that could handle compressed files in PostScript, a page description language favored by high-energy physicists because of its ability to describe images in a device-independent manner.[14]

According to Berners-Lee (the Web's creator), the mounting of SPIRES-HEP on the World-Wide Web was a vitally-important factor in the rapid acceptance and utilization of the Web in the international high energy physics community [15]. Driven by increasing pressure to disseminate results as quickly as possible and to as wide a group as possible, the high energy physics community has embraced the World Wide Web because it a significant improvement in the communication of time-sensitive, much-sought-after information. While some of the records mounted on the Web at SLAC are also generated in paper form, many — in fact, less and less as time goes by — exist only in their Web incarnation. The authors see no need to produce the document or information in two formats, and the interactive electronic format of the Web is much preferred

WHERE AND HOW

Because the WWW Wizards group met under the sponsorship of the SLAC Library, and because the maintenance and delivery of the Web HEP Pre-Prints interface was spearheaded by Louise Addis, SLAC Associate Head Librarian, the Library has been able to provide the Archives with documentation of the Wizards' activity and of the HEP Pre-Prints site. Some documentation has also been preserved by individual WWW Wizards, and all of the computer backup tapes from the founding era of 1991 - 1994 still exist (although not on the current main-frame platform). Individual Wizards have conveyed to the Archives paper documentation that they had retained for "historical" purposes in their own files, including paper copies of e-mail messages that were exchanged about setting up the first server, getting it to work properly, and negotiating changes and additions to the SLAC home page. Wizards and members of the Web Coordinating Committee have also provided hard copies of written papers and presentations on the WWW and SPIRES-HEP, and, those papers usually have included examples of then-current SLAC Web pages

APPRAISAL

One thing IS certain at this point, however, and that is that SLAC World Wide Web pages are official records. They fit the statutory definition of records [16] because they are documentary materials, in machine readable format, made or received in the transaction of public business and appropriate for preservation "as evidence of the organization, functions, policies, decisions, procedures, operations, or other activities of SLAC." [17] As official records, SLAC's early World-Wide Web pages, which were created by elements of the Research Division, need to be appraised and

scheduled. Because it is a U. S. Department of Energy contractor, SLAC's records are appraised and scheduled for either temporary or permanent retention based on records disposition schedules negotiated between SLAC, the Department of Energy, and the National Archives and Records Administration

The DOE disposition schedule for Research and Development (R&D) records [18] encompasses R&D records generated within all DOE contractors and national laboratories. The DOE R&D Schedule designates permanent research and development records as "Level I" records, and defines them as the records of :

"Projects which received national or international awards of distinction; active participation of nationally or internationally prominent investigators; research which resulted in a significant improvement in public health, safety, or other vital national interests; scientific endeavors which were the subject of widespread national or international media attention and/or extensive congressional, DOE or other government agency investigation; show the development of new and nationally or internationally significant techniques which are critical for future scientific endeavors, or made a significant impact on the development of national or international scientific, political, economic or social priorities."[19] The World Wide Web fits two of these criteria: it was and is the subject of widespread national and international media attention; and it is a new and internationally significant technique which is critical for present and future scientific endeavors. Although either one of these attributes would alone be sufficient to establish the permanent value of SLAC's early Web documentation, the latter attribute, of course, quite eclipses the former in importance

The identification of which past Web documents fall into the "permanent retention" category is the next step in this process. Retained backups from the early days are being moved into a single storage location in the SLAC Computing Services' vast array of storage sites. However, these backups can not be used on the current computing platform, since SLAC migrated this past spring from a VM main-frame platform to a distributed computing paradigm. [20] Besides the access problem migration has created, there are several other problems as well. The first has to do with the inadequacy of using computer backups as permanent records, particularly for Web documents; the second with the issue of whether it is appropriate to attempt to permanently preserve all of the dimensionality of a permanently-retained Web document

THE "BACKUP" PROBLEM

The first problem with backups is: everything is there. A "backup" contains all of the files that were on a computer system at the time the backup was made, and at any given point in time, much less than half of what is backed-up is of permanent value.[21] For the early web documentation at SLAC, this problem is being handled by knowledgeable staff, who — as previously mentioned — are patiently reviewing old storage disks, as time permits, and moving the World-Wide Web files to a separate storage location. Not every organization will be willing to support such a time-consuming process, however, and it is not a practical long-term solution. Although there is and have been "archive" capabilities on the current and previous computing platforms at SLAC, documents in this Computing Services "archive" are unsystematically selected for retention by their individual creators, who are saving or discarding web documents based on de-centralized, changing, and certainly non-uniform criteria. In addition, although the volume of "archived" documents is smaller than the volume of backups, both share a further attribute that makes them an archival problem. Over time, odd things happen to Web document names: the content, or document attached to a particular "name" changes at regular intervals, because the name is both an address and a name. The "Web document name" is what external and internal systems — the "Internet" and the "intranet" — use to locate a document. In Web lingo, the document name is a "URL," or "Uniform Resource Locator." The "Web document name"/URL is also what you and I type when we want to view a document on the Web. When a Web page owner updates a page, the content, the information — and therefore, the "document" — is changed, but the URL, necessarily, stays the same. Otherwise, system would not be able to find the page, either through a directly-typed request, or through a previously-established link on another page. (See Figure 2.) This ability to link and to renew links constitutes the communicative and functional beauty of the Web, but it makes the World Wide Web an artifact of the present with a disappearing past

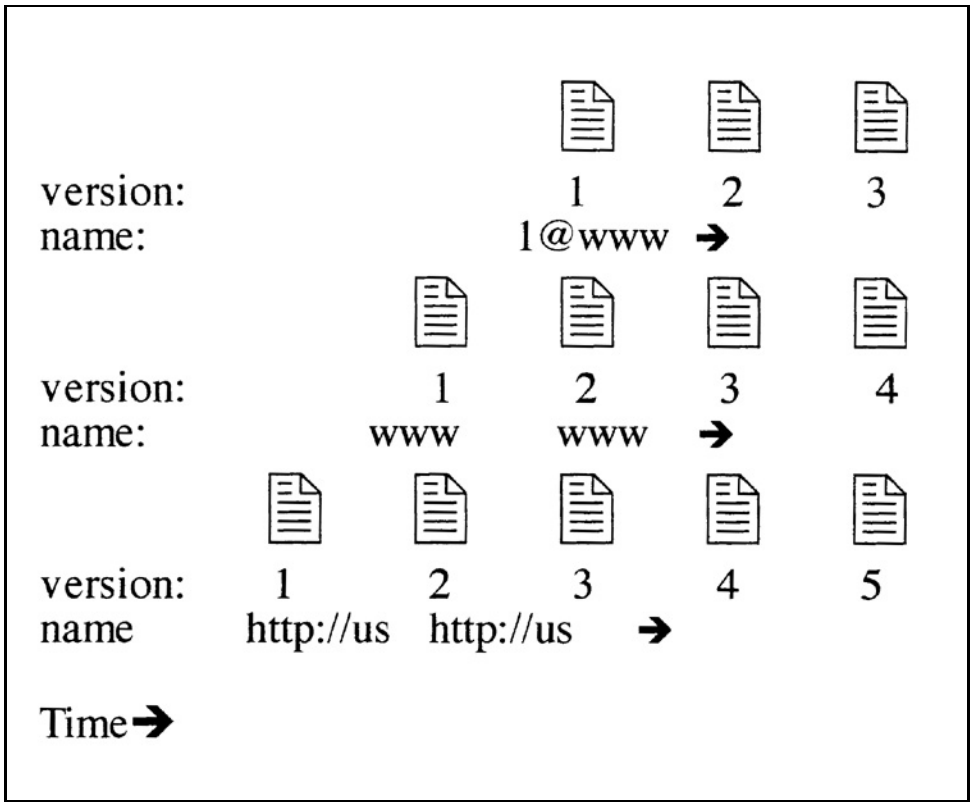

Figure 2

At SLAC, Web page changes have been automatically handled by the backup procedure on the HEP web site by means of a cascading name change that allows a limited number of iterations of a particular web page to be backed up and saved at any given time. In this process, the URL name travels to the newest iteration of the page; and backup name 1 is assigned to the immediately superseded iteration. At the time of the next backup, name 2 is assigned to the first iteration; name 1 is assigned to the second iteration, and the URL name is placed on the most current, or third, iteration. After the upper limit of backup documents of a single page have been stored (at SLAC, this number is six) the next subsequent backup cycle triggers the disposal of the oldest iteration in storage. Using file lists of the backup tapes, one can determine when a particular Web document was created ONLY IF a tape was taken out of the backup cycle, and therefore not overwritten, before an original Web document reached its seventh iteration. In fact, the earliest Web pages at SLAC were set aside through precisely such interventions. In the effort to locate SLAC's first web page on the "archived" backups, a combination of personal recollections of WWW Wizards and information from other electronic records (e-mail) was used to establish a rough estimation of the date of the page. "Archived" backups from that time period were then reviewed for the earliest-dated version of the document name. Using this approach, a digital copy of SLAC's first web page was located on the backups. This same approach will now have to be used to locate the other early web pages which will be appraised as permanent at SLAC, including the first web pages for each experimental group and, most likely, the first web pages for each department. Another historic page that will be sought in the early SLAC web backups is the code for the first World Wide Web page in China, which was set up with assistance from SLAC staff, in support of a collaboration with the Institute of High Energy Physics in Beijing. [22]

THE "DIMENSIONALITY" PROBLEM

Once Web documents have been appraised, and the electronic version of each permanent page has been located, the next issue becomes one of preservation. How should the Web documents be preserved, and who should preserve them? As a U. S. Department of Energy contractor, SLAC has an obligation to retire its permanent records to the

National Archives and Records Administration (NARA). NARA requirements for the retirement of electronic or digital records are quite clear: they must be retired "on one-half inch, seven or nine track reel-to-reel magnetic tape and 3480 class tape cartridges" in flat ASCII or EBCIDIC in a fielded format [23]; or on CD-ROM's which include fielded data files or text files and conform to the International Standards Organization (ISO) 9660 Standard and comply with the American Standard Code for Information Interchange (ASCII) [24] Any Web document transferred to NARA becomes "flattened," that is, loses its interoperability with other Web documents, both by its removal from active relation to the other Web pages to which it was designed to link, and by its conversion to one of the required transfer file formats. In a "flattened" Web document, the addresses of the links remain embedded, but the links themselves are broken, and can no longer be activated. However, Web documents on backups are already somewhat flattened, because they have been removed from the World Wide Web environment. In order to regain their dimensionality, such documents — AND all of the documents to which they are linked — would have to be once again mounted onto a Web server. As this issue arises for more and more organizations, one suspects that most will decide that they do not have the resources nor the institutional willingness both to set up and maintain permanent, fully-dimensional Web documents on an archival Web server on site AND to simultaneously fulfill their primary missions. It can be predicted that, unsurprisingly, most will opt, quite reasonably, to let their fully-dimensional Web documents expire

OTHER OPTIONS

There are some independent repositories, though — most notably The Archive of the Internet [25] and The World Wide Web History Project [26] — which are proposing to preserve fully-dimensional archives of active Web pages on servers dedicated solely to these archival documents. The mission of The Internet Archive is "collecting and storing public materials from the Internet...the Archive will provide historians, researchers, scholars, and others access to this vast collection of data (reaching ten terabytes), and ensure the longevity of this information." [27] The goals of the World Wide Web History Project are somewhat more focused: "The World Wide Web History Project is a collaborative effort to record and publish the history of the World Wide Web and its roots in hypermedia and networking...producing a definitive history and historical archive of the Web." [28] Although both of these organizations have high-powered staff and impressive goals, their ability to achieve those goals remains to be proven. It is unclear, for example, how they will handle web pages that are search forms for internally produced and maintained databases. Search form pages become quite meaningless without active links to the underlying databases, which are not on the web, and are not in markup language format. Nevertheless, I do plan to recommended that, in addition to the required retirement of SLAC Web documents to NARA, a parallel transfer to one or both these repositories be considered, because of the promised full dimensionality. In fact, it is likely that many of SLAC's public web pages have already been collected by the Internet Archive web crawlers, and in 1996, the co-founders of the World Wide Web History Project conducted extensive interviews and videotaping sessions with SLAC WWW Wizards, and these tapes and transcripts are now held by that repository.

CURRENT AND FUTURE WEB DOCUMENTS

The approach that has been taken with SLAC's prior World Wide Web documents has, of necessity, been adopted after the fact. We have an opportunity with the present and future web documents, though, to expedite and streamline the appraisal and transfer process. For organizations which are not as large as SLAC, or which are not as polymorphously web inter-active, the use of backup tapes as the primary instrument of archival appraisal may not present a significant challenge or hardship. For SLAC, however, using backups does not appear to be a practical way of conducting the ongoing business of appraisal and transfer to the archives of permanent Web documents.

In our situation, it will probably be both more efficient and effective to identify which elements of the organization are most likely to generate Web pages that need to be scheduled for permanent retention. Some form of "web crawler" technology, like that used by commercial Web search engines, could then be used to gather iterations of pages from the designated URL's at regular intervals and deposit them in a specified storage location on site. The Web documents in the designated storage site would then be converted to one of the required file transfer formats, and deposited with NARA.

The task for the coming months, therefore, will be to develop and gain approval for a disposition schedule for SLAC Web documents based on this automated gathering approach. We will also need to investigate the mechanics of retiring permanent Web documents to NARA, as well as to one of the previously mentioned independent repositories, if such a transfer is deemed to be worthwhile. As we go about these tasks, we will also maintain a keen interest in developments in thinking about World Wide Web archival and records management issues coming from other projects currently underway, particularly the Syracuse University project to investigate the status of material posted on state agency World Wide Web Sites [29].

NOTES

1. SLAC Welcome Page, http://www.slac.stanford.edu/

2. Hobbes Internet Timeline, http://info.isoc.org/guest/zakon/Internet/History/HIT.html; also The World Wide Web Consortium's "A Little History of the World Wide Web, http://www.w3.org/History.html.

3. WWW History Project, http://www.webhistory.org/project/book.html:

   > "...Our goal is to establish a permanent process for recording and disseminating the ongoing history of the Web and networked information, in as close to real time as possible. Much of our work on the Project has been clearing up the six years worth of myths and half-truths which have accumulated around the Web's origins, precisely because there was no definitive history. We want to make it easier next time!"

4. Fermi-News August 16, 1996, p. 1 and ff. (It should be reiterated here that this Fermi-News article did not, in fact, claim that FermiLab was the first U.S. WWW site.)

5. http://www.w3.org/pub/WWW/People/Berners-Lee/.

6. http://www.w3.org/pub/WWW/History.html

7. ibid. Work on installing the server at SLAC was undertaken by both Terry Hung and Paul Kunz. (Paul F. Kunz, e-mail to Jean Deken 30 March 1998, Re: History of WWW at SLAC:

   > "The time between September 1 1991 when I brought a Web browser to SLAC and source code for the server and December when [the] server came up is very long. The explanati[o]n is that Terry Hung .... did not know the SLACVM system very well so dropped the project...Finally in December, I put myself to finishing the server installation since I knew how to handle VM and the interface to SLAC Spires...")

8. Addis, Louise.

   > "A Brief and Biased History of Preprint and Database Activities at the SLAC Library 1962-1994," p. [26] Appendix to: Addis, Galic, Kreitz and Johnson, "The Virtual Library in Action." Presented at the American Chemical Society (ACS) National Meeting "Chemical Information Symposium" (CINF) "The Library of the Future," Anaheim California, 4 April 1995 (unpublished manuscript, SLAC Archives, unprocessed papers of the WWW Wizards.)

9. 21 September 1994 Memorandum, To: Appendix to WWW Wizards Report to C. Dickens, From: WWW Wizards Committee, Subject: Brief background info on Web at SLAC (SLAC Archives, unprocessed papers of the WWW Wizards.)

10. 21 Sept 1994 Memorandum, To: Appendix to WWW Wizards Report to C. Dickens, From: WWW Wizards Committee, Subject: Brief Background info on Web at SLAC (SLAC Archives, unprocessed papers of the WWWizards; also Addis, Galic, Kreitz and Johnson, "The Virtual Library in Action" p. 5.)

11. http://www.slac.stanford.edu/slac/www/wwwtech/wwwtech.html

12. http://www.slac.stanford.edu/slac/www/wwwcc/charge.html

13. Addis, Galic, Kreitz and Johnson, "The Virtual Library in Action" p.1

14. http://www.cs.indiana.edu/docproject/programming/postscript/ what-is-it.html

15. Berners-Lee, Tim. Keynote Address, April 11, 1997, History/Developers' Day, Sixth International WWW Conference, Santa Clara, CA. (videotape) (Slides available at http://www.w3.org/Talks/9704WWW6-tbl/) Also: Johnson, Tony J. "Spinning the World-Wide Web" SLAC Beamline, vol. 24, no. 3/4, Fall/Winter 1994, pages 2-9.

16. 44 U.S.C. 3301; http://www.law.cornell.edu/uscode/44/3301.html

17. ibid.

18. National Archives Disposition Job N1-434-96-9 (pending)

19. ibid.

20. http://www.slac.stanford.edu/comp/vm/vmmigr.html

21. Estimates for the ratio of permanent to temporary records in an organization's paper-based system place the permanent records at less than 5% of the total volume. (See: Records Management Handbook, "Disposition of Federal Records" National Archives and Records Administration. 1989. p. 4)

22. Cottrell, R. L. A. et al. "Networking With China" SLAC-PUB-6478, April, 1994.

23. 36 CFR 1228.188 "Transfer of Machine Readable Records to the National Archives

24. NARA Bulletin 94-4, Use of Compact Disk-Read Only Memory (CD-ROM) medium to transfer records to the National Archives.

25. http://www.archive.org

26. http://www.webhistory.org/home.html

27. http://www.archive.org

28. http://www.webhistory.org/home.html

29. "Managing Web Sites is focus of Syracuse Study" NAGARA Clearinghouse v. 13, n. 3, Summer 1997, page 1; also http://istweb.syr.edu/~mcclure/