Persistence of information on the web: Analyzing citations contained in research articles

Steve Lawrence, Frans Coetzee, Eric Glover, Gary Flake, David Pennock
Bob Krovetz, Finn Nielsen, Andries Kruger, Lee Giles

NEC Research Institute, 4 Independence Way, Princeton, NJ 08540
{lawrence,coetzee,compuman,flake,dpennock,krovetz,fnielsen,akruger,giles}@research.nj.nec.com

ABSTRACT

We analyze the persistence of information on the web, looking at the percentage of invalid URLs contained in academic articles within the CiteSeer (ResearchIndex) database. The number of URLs contained in the papers has increased from an average of 0.06 in 1993 to 1.6 in 1999. We found that a significant percentage of URLs are now invalid, ranging from 23% for 1999 articles, to 53% for 1994. We also found that for almost all of the invalid URLs, it was possible to locate the information (or highly related information) in an alternate location, primarily with the use of search engines. However, the ability to relocate missing information varied according to search experience and effort expended. Citation practices suggest that more information may be lost in the future unless these practices are improved. We discuss persistent URL standards and their usage, and give recommendations for citing URLs in research articles as well as for finding the new location of invalid URLs.

1. INTRODUCTION

One of the limitations of the web is that link consistency is not enforced. There are many invalid links on the web, leading to annoyance and frustration for users. The use of URLs in research articles has been of particular concern. Some have argued that URL citations should not be contained in research papers because of the lack of persistence of URLs (and because the contents of the respective pages can change).

In this paper we analyze URLs contained in computer science research articles, looking at the evolution of how many URLs are cited, analyzing the type of URLs cited, analyzing the current validity of links, and analyzing the invalid links in detail. We provide recommendations for citing URLs in research articles, and for finding the new location of invalid URLs.

CIKM 2000, McLean, VA USA

2. PRESERVING INFORMATION ON THE WEB

The inclusion of adequate citations is arguably one of the major characteristics of serious scholarly work. Citations are used to assign credit, legitimize arguments, provide additional context, and as a way to summarize relevant background material. However, a citation is strictly valid only if it is possible for the reader to independently access the referenced material at a later date, and verify the writer's assertions.

Before the advent of the World Wide Web, adequate standards for citation emerged from centuries of experience and library studies. These standards addressed both style, and more importantly, minimum requirements regarding accessibility and permanence of materials that can be cited. The most widely accepted citations are for books, journals and conference proceedings from reputable publishers.

Some of the more plausible postulates for the widespread acceptance of the above citations are: first, a high cost of entry resulted in relatively few publishing sources, thereby making it easy to explicitly and implicitly enforce standards and provide repositories. Second, formal procedures and enforced centralization (such as ISBN number assignment, or Library of Congress centralization) have developed over centuries. Third, multiple copies of materials were distributed by these publishing sources. Finally, the materials to be disseminated were simple written materials.

In its present state, publishing on the World Wide Web has few of these characteristics. The web has removed the barriers to the publishing world for individuals. Everyone with access to the web has the ability to disseminate personal material widely and at low cost, simply by sharing a URL.

The benefits of being able to widely share multimedia materials at minimal cost, unfortunately, have been marred by the fact that most individuals or even organizations do not represent a reliable or stable publishing source. Web pages are abandoned, servers are shut down, and files may be arbitrarily renamed, for example. It is imperative that adequate mechanisms be found to ensure that citations to materials on the web not be needlessly lost.

Restricting the practice of citing URLs is not an answer; some materials and journals are now only published via the web. In fact, we were unable to obtain non-web instantiations for some of the references of this paper [in particular, those related to the Uniform Resource Name (URN) scheme].

There have been many proposals for improving the situation. Some authors focus on the problem from the web designer's perspective. For example, Creech [6] describes a link management technique to help automate the process of identifying and correct-

ing broken links throughout a site, or even across multiple sites controlled by a single enterprise.

Other authors propose augmenting existing web protocols to improve link persistence. Ingham et al. [9] suggest the development of an object-oriented network to exist in parallel with the current web that enforces referential integrity and performs garbage collection. Hyper-G [11] goes one step further; it is a proposed replacement for the current web, with built in mechanisms for enforcing link consistency. Other related work includes [17, 18, 21, 24].

A major promising effort is the Uniform Resource Name (URN) specification [25]. Partial implementations such as Persistent URLs (PURL) [22] and Handles [2] provide an indirection mechanism that allows for more location independence and hence better persistence. However, from our database, these protocols have seen very little use in citations. Unfortunately, existing URN services also require cooperation from the owner of the material to ensure that the URN is valid and that document contents do not change.

Replacing HTML by improved protocols (see [16], for example) could in the future result in interesting content-based solutions. These protocols could provide support for improved content based indexing and retrieval. In principle, content summarization and indexing can be used by search engines to recognize materials that move on the web. Phelps and Wilensky [19] have shown that most documents on the web can be uniquely identified based on a small set of words that no other document shares. These words can be used to augment URLs, and may be used to locate documents that move.

The above approaches can at best redirect attention to material that has moved, but not disappeared. The issue of material being totally lost on the web has to be faced. It is therefore imperative that reasonable estimates of the problem of invalid web citations be obtained, and reasonable policies should be instituted by academic societies and publishers to encourage good practices.

We note that a related issue is that the content of web citations can change, such that subsequent readers may not be viewing exactly the same material that has been cited. This issue can be addressed with version management as in Xanadu [17] and other proposals, or by periodically archiving the whole web. The Internet Archive stores snapshots of information from the web [1, 10]. However, it is still an open question whether this approach can fully solve the problem. Alexa has estimated that web pages disappear after an average time of only 75 days. Further, taking a snapshot of the web is non-trivial. The time required to download a snapshot means that many pages will change while the snapshot is being generated.

Another related issue is the format of information and the possibility of limited or no availability to the hardware and/or software required to read specific data formats.

Other efforts to improve permanence on the web include the Intermemory project at NEC Research Institute [5, 8], which aims to create highly survivable and available storage systems that are made up of widely distributed processors which may be unreliable and untrustworthy, and LOCKSS (Lots of Copies Keeps Stuff Safe) [20], in which multiple libraries work together to redundantly cache copies of specific documents.

3. METHODOLOGY

We analyzed the validity of URLs contained in computer science research papers from the CiteSeer (also known as ResearchIndex) database [12, 15]. CiteSeer is a system for indexing Postscript and PDF research articles on the web. A free service is available at `http://csindex.com/` (if this URL is invalid, try searching for CiteSeer or ResearchIndex in a search engine). The database currently contains about 300,000 research articles, including journal papers, conference papers, and technical reports. The CiteSeer database represents computer science papers that are available on the publicly indexable web [14].

URLs were extracted by searching for strings starting with (http:|https:|ftp:), and ending with whitespace or a quote. We removed trailing periods, commas, semicolons, parentheses, and brackets.

The CiteSeer database contained 270,977 articles at the time of our study, 100,826 of which were cited and linked within the database, and hence the publication year was known. For these 100,826 articles we extracted URLs from the header, body, and citation sections (67,577 URLs), and then attempted to access each URL. URLs that were redirected to another URL were followed to their new destination. These experiments were performed during May 3 - May 5, 2000.

4. RESULTS

4.1 Number of URLs in research papers

Figure 1 shows the average number of URLs contained in the research articles as a function of the year of publication, which has been increasing substantially since the inception of the web.

4.2 Percentage of invalid URLs

Figure 2 shows the percentage of invalid URLs in papers as a function of the year of publication of the source papers. The percentages are corrected so that they do not include URLs containing extraction errors (see the next section). The percentage of URLs that were invalid varies from 23% in 1999 to a peak of 53% in 1994. The lower percentage of invalid URLs for 1993 may be because many citations at this early stage of the web were to relatively well-known sites. However we note that the sample size for URLs is relatively small prior to 1994, leading to lower accuracy (only 608 URLs were extracted from papers published in 1993, while 21,056 URLs were extracted from papers published in 1998).

We also looked at the percentage of invalid links as a function of the year of publication for URLs located in the body of articles versus URLs located in the citation list. We found no substantial difference in the percentage of invalid URLs contained in these sections.

4.3 Classifying invalid URLs

We selected a random sample of the invalid URLs and manually classified them into the following classes:

1. The URL was extracted from the paper incorrectly, contains a syntax error such that it could never be valid, or is a sample URL that was not intended to be valid (e.g. `http://a.b.c/`).
2. The new location of the URL could be found by guessing an alternate URL or browsing the web.
3. The new location of the URL could be found with the use of a search engine. We used mostly Google, CiteSeer and Inquirus [13] when searching.

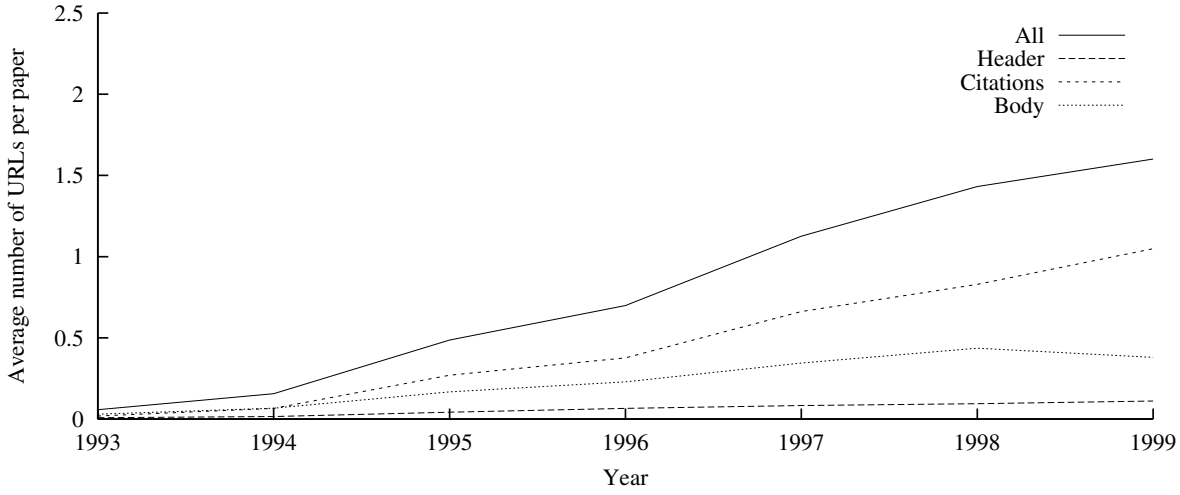


Figure 1: The average number of URLs contained in the papers as a function of the year of publication, for the header, body, and citation sections of articles, and for all sections combined.

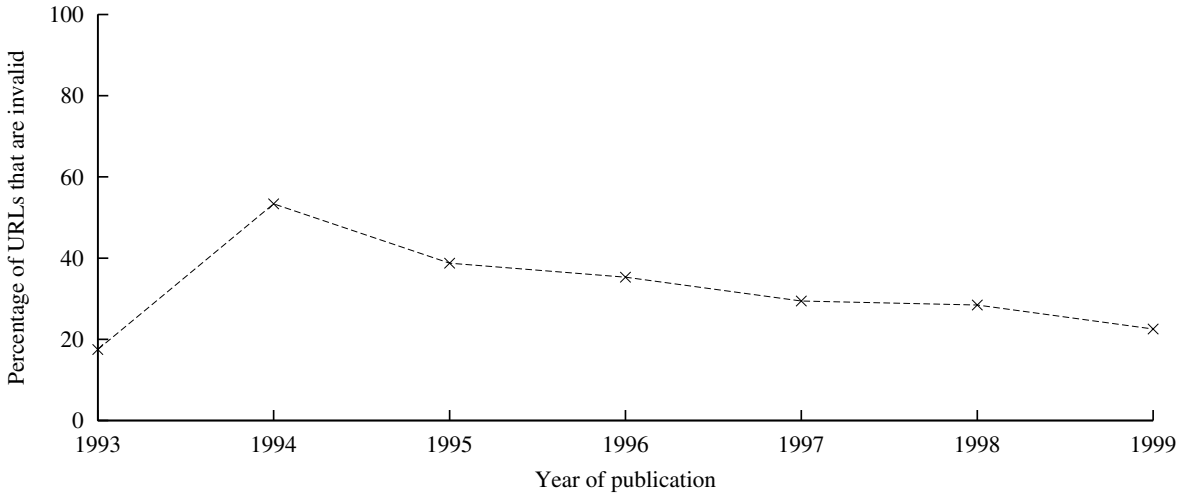


Figure 2: The percentage of invalid links contained in articles as a function of the year of publication of the articles.

4. The location of highly related information could be found with a search engine but it is not clear how good a substitute the related information is for the original URL.
5. The new location of the information could not be found but the URL was accompanied by a formal citation to a publication.
6. The new location of the information could not be found.

We manually classified 300 invalid URLs. Of these URLs, 31.7% were either extracted incorrectly from the papers, contained a syntax error such that they could never be valid, or were example URLs that we believe were never intended to be valid. Extraction errors were typically due to the `pstotext` conversion program not converting special characters correctly or inserting spaces within the URLs (our extraction routine corrects for some easily identifiable cases but not all). We removed these URLs from the dataset and the percentages reported are for the remaining URLs. Table 1 and Figure 3 show the percentage of the remaining invalid URLs contained in each category.

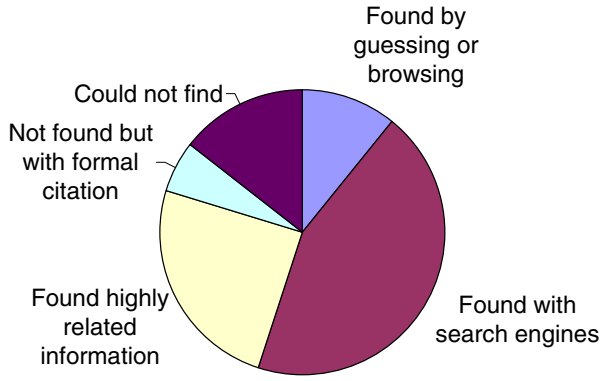
For 79.7% of the remaining invalid URLs we were able to find the new location of the page or highly related information (which is likely to be a good substitute for the original page, however we cannot guarantee this because we do not have access to the original page). For 5.9% of the URLs we could not find the new location but the URL was accompanied by a formal citation. The location for the remaining 14.4% of URLs was not found. A moderate amount of effort was put into locating moved or related information. We spent no more than about five minutes for each URL. More of the invalid URLs may be locatable given more search experience, more time, or better search tools.

URLs that were reported as lost were given to a second searcher. The second searcher was able to locate 80% of these lost URLs, bringing the overall percentage of lost URLs down to 2.9%. The revised percentages of URLs in each category after the second searcher can be seen in Table 1 and Figure 3.

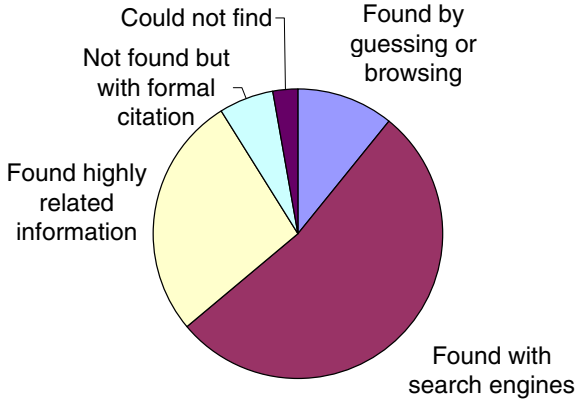
There was a significant difference between the five individuals that participated in the experiment for the success of locating URLs. The most successful individual located all URLs investi-

Category	Percentage	After second searcher
Found new location by guessing or browsing	10.7%	10.7%
Found new location with search engines	44.4%	53.2%
Found location of highly related information	24.6%	27.3%
Could not find but with citation	5.9%	5.9%
Could not find information	14.4%	2.9%

Table 1: Classification of invalid URLs.



(a) Initial classification



(b) Classification after second searcher

Figure 3: Classification of invalid URLs.

gated, while the least successful individual was unable to locate 16% of the invalid URLs. These differences are due to different degrees of persistence, differing search experience and abilities, and differences in opinions regarding whether or not located information was highly related in the case of related information.

For each of the URLs where relocated or highly related information was found, the searchers estimated the difficulty locating the URLs. The following classes were used:

- Easy
- Somewhat difficult
- Very difficult

Table 2 shows the percentage of lost URLs in each class.

5. LOST INFORMATION

For each invalid URL that could not be located, we examined the context of the citation in the respective paper, and estimated the importance of the URL with regard to the ability for future research to verify and/or build on the given paper. The following classes were used:

- URL is not very important
- URL is somewhat important
- URL is very important

Table 3 shows the percentage of lost URLs in each class before and after the lost URLs were sent to a second searcher.

Only 9% of the lost URLs were considered to be very important with regard to the ability for future research to verify and/or build on the given paper. After a second searcher looked for URLs that could not be found by the first searcher, none of the remaining lost URLs were considered very important.

The 6 remaining lost URLs after the second searcher were two author homepages, a web link verifying tool listed as one of two possible tools to use for verifying web links in future research, the homepage of the OSF formal methods project (many papers from the project could be found), a page listing search resources referred to in a paper listing URLs of interest to astronomers, and a result URL from a sample search found by a search tool.

Additional searchers may be able to locate some of the remaining lost URLs, or it may be possible to track some down by emailing individuals, for example.

5.1 Type of URLs contained in articles

Figure 4 shows the type of URLs as a function of the year of publication of articles, identifying HTTP versus FTP links, and links to Postscript and PDF files. As expected, the use of FTP links has been declining. The results for PDF files suggest that computer science researchers still use predominantly Postscript instead of PDF for posting research articles.

5.2 Cached URLs

The search engine Google [4] (http://www.google.com/) maintains cached copies of the pages that it indexes. For the random sample of invalid URLs as above, we tested to see if the page was cached by Google (and hence easily accessible). 22% of the invalid URLs were available from Google (excluding file types that Google does not index).

5.3 Invalid hosts

Tables 4 and 5 show the top 10 domains for invalid URLs, for sites with no valid links found (Table 4) and according to the number of invalid links (Table 5). ds.internic.net used to contain Internet RFCs (Request For Comments), which we consider very valuable.

Difficulty locating information	Percentage of located URLs
Easy	53.8%
Somewhat difficult	30.8%
Very difficult	15.4%

Table 2: Classification of invalid URLs.

Importance of lost URL	Percentage of lost URLs	Percentage after second searcher
Not very important	50.4%	50%
Somewhat important	40.6%	50%
Very important	9.0%	0%

Table 3: Classification of lost URLs.

It turns out that RFCs are quite easy to find in new locations on the web.

Interestingly, one of the broken links that we manually inspected was a link to RFC 1737 in a paper titled "Fixing the 'Broken-link' Problem: The W3Objects Approach" [9] (Doubtless the current paper will serve as an ironic footnote in a future publication).

We note that the site proceedings.www6conf.org ranked in the top 20 domains containing many invalid links and no valid links – the WWW conferences have been known to make the proceedings available on a given domain and discontinue the domain sometime after the conference.

6. PERSISTENT URL STANDARDS

Our initial belief was that authors may be motivated to use persistence mechanisms for their citations, and especially for their self-citations. We therefore wished to investigate the use of Uniform Resource Names (URNs) in citations.

The URN specification, produced by the Internet Engineering Task Force, provides a very flexible framework according to which unique identifiers can be assigned to both static and dynamic materials, and allows for such material to be tracked and located as it moves on the Internet. Partial implementations have been produced, notably the Persistent Uniform Resource Locator (PURL) [22] system, and the Handle [2] system. These systems both make use of distributed resolution servers to redirect attention to materials. Both of these implementations at present require cooperation on the part of the author of the material, both to generate the initial resource, and to maintain the validity of the indirection. Despite the obvious motivation of early users, already not all PURLs are valid. A search for url:purl.oclc.org at AltaVista turned up many PURLs that return a page stating that "The requested PURL has been deactivated and can not be resolved."

The major difficulty with evaluating URNs is the difficulty of detecting them. URNs (especially PURLs), are designed to be drop-in replacements for URLs. At the time of writing this paper, we are unable to detect URNs by resolving the redirection mechanisms. A second difficulty with evaluating these services is the relatively short time that they have been available.

As an initial attempt at obtaining an answer, we searched for all URLs that are resolved by the main PURL resolver, located at purl.org (we searched for all URLs containing the string "purl"). In principle, other PURL resolvers exist, but since the resolver software has not been available to others for long, we believe these to be much less popular.

The results obtained on our database show poor adoption of PURLs. Of the 67,577 URLs extracted from the papers in CiteSeer, we were only able to locate 11 PURLs (0.016% or URLs), all of which were to the same page http://purl.org/metadata/dublin_core).

In the future, we hope to perform a more extensive study to identify all possible URNs. However, we believe that there is still a long way to go in promoting the use of these approaches over standard URLs.

7. ANALYZING URL CITATION PRACTICES

Through our manual classification of invalid links, we have identified a number of reasons why URLs become invalid. Three problems that are likely to persist are:

1. Personal homepages tend to disappear when researchers move. This is particularly common for students, post-docs and sabbatical visitors. In some cases, old homepages are still available although a new homepage exists at another location.
2. URLs for academic software that is maintained on a personal machine tend to disappear. Machines come and go, and machine names change, leading to problems.
3. Sites may be restructured without maintaining old links.

There are also a number of possibly one-time problems due to the initial rapid growth of the web. Increased standardization should hopefully lead to fewer problems due to the following:

1. Most servers have completed the change from FTP to HTTP.
2. Early pioneers ran their own web servers (personal machines); however the infrastructure is now typically provided by universities and corporations, and paths are more stable. Servers set up on non-standard ports are becoming less common.
3. Certain conventions in setting up sites have become more commonplace (possibly due to Apache and a few other servers becoming so widespread). URLs for homepages, for example, have tended to become standardized to the ~ notation rooted at the domain name. This makes it easy to guess the new location of some invalid links. We found that the depth of the URLs therefore decreases (as shown in Figure 5), and the citation is more insulated from network reconfigurations.
4. With domain names easily available, software has moved from personal repositories to dedicated sites. Most successful academic software now has its own `.com` or `.org` domain.

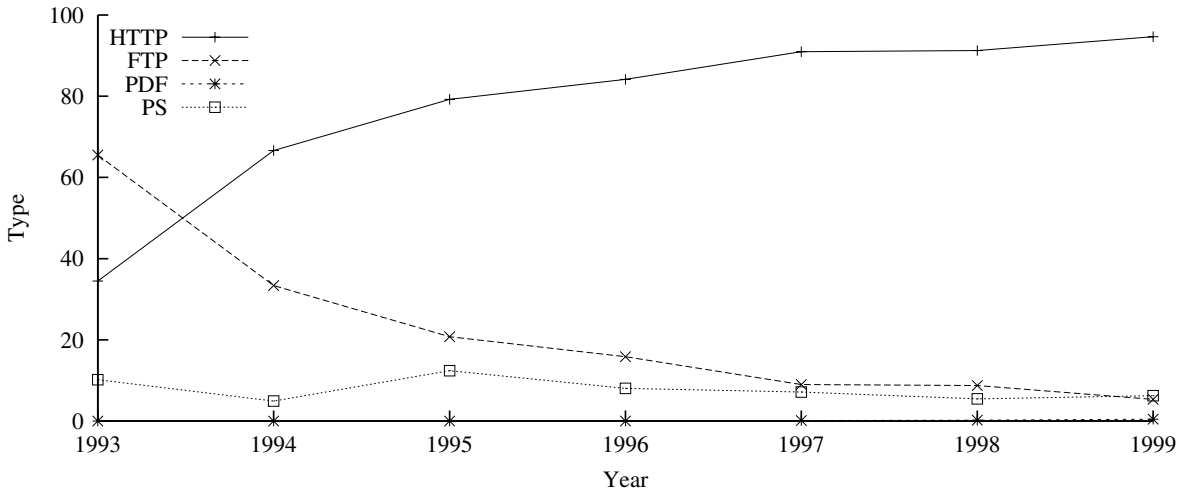


Figure 4: The type of URLs as a function of the year of publication of the articles.

Rank	Domain
1	ds.internic.net
2	harvest.cs.colorado.edu
3	netlib.att.com
4	www.cs.cmu.edu:8001
5	um.org
6	warp.cs.cmu.edu
7	www.my.edu
8	rakaposhi.eas.asu.edu:8001
9	snapple.cs.washington.edu:600
10	ftp.tac.mta.ca

Table 4: The top 10 domains containing only invalid links ranked by the number of invalid links.

8. RECOMMENDATIONS FOR CITING AND GENERATING URLS

Based on our experiences in labeling missing URLs, we have formulated a number of good citation practices that should improve the chances of future readers finding information that may have moved. We wish to emphasize that researchers have a vested interest in following good citation practices. A side-effect of dead links is that they may have a negative impact on the ranking of the containing material. For example, formal approaches have been proposed to bypass such pages during browsing [3], or reduce their ranking when presenting search engine results to users [23].

We note that proposals have been made to update style manuals (such as those of the Modern Language Association [7]) to specify how to cite URLs. However, these manuals typically only specify how to format the text representation. While helpful, adequate formatting does not help in locating moved or missing material.

The first set of recommendations relate to all authors:

1. We recommend that URL citations be accompanied by formal citations to published works whenever possible. We also feel that URL citations deemed valuable to the reader should be included even when a formal citation is not available. Although some percentage of URLs may become invalid over time, the percentage of links that are left out that are unavailable is 100%. Even if a formal citation is available, the existence of an accompanying URL can significantly improve the accessibility of the information.
2. Enough context information should be provided at the location of citation to enable readers to pose adequate queries to search engines in order to track down invalid links. For example, when giving the URL for a preprint, the full title of the document should be given, along with full details of the authors (as opposed to using "et al." for example). We found many examples where the contents of URLs could not be inferred from the context.

We found that many URLs cite repositories controlled by the author. In such a case we feel the author should invest in adequately preparing material for citation.

1. If at all possible, place materials in a reliable central repository, such as a preprint or software archive. We believe that this is particularly important for links to complete versions of papers, supporting data or results, and omitted proofs.
2. Repositories should be named, and the name of the repository should be given in the citation. The name can then be used for later searches. Software distributions should include a file with the name of the software package; this file can be indexed by some search engines. The name of the file should not change as the distribution changes. Software should have a documented homepage. Establishing a domain name for the software also helps.
3. Unless absolutely required, when referencing software manu-

Rank	Domain	Percentage of all invalid URLs
1	www.cs.cmu.edu	2.5%
2	ds.internic.net	1.1%
3	www.research.att.com	0.9%
4	www.netlib.org	0.9%
5	www.zyxel.com	0.6%
6	www.cs.kuleuven.ac.be	0.6%
7	www.ics.uci.edu	0.5%
8	www.genmagic.com	0.5%
9	harvest.cs.colorado.edu	0.5%
10	www.mcs.anl.gov	0.5%

Table 5: The top 10 domains for invalid URLs.

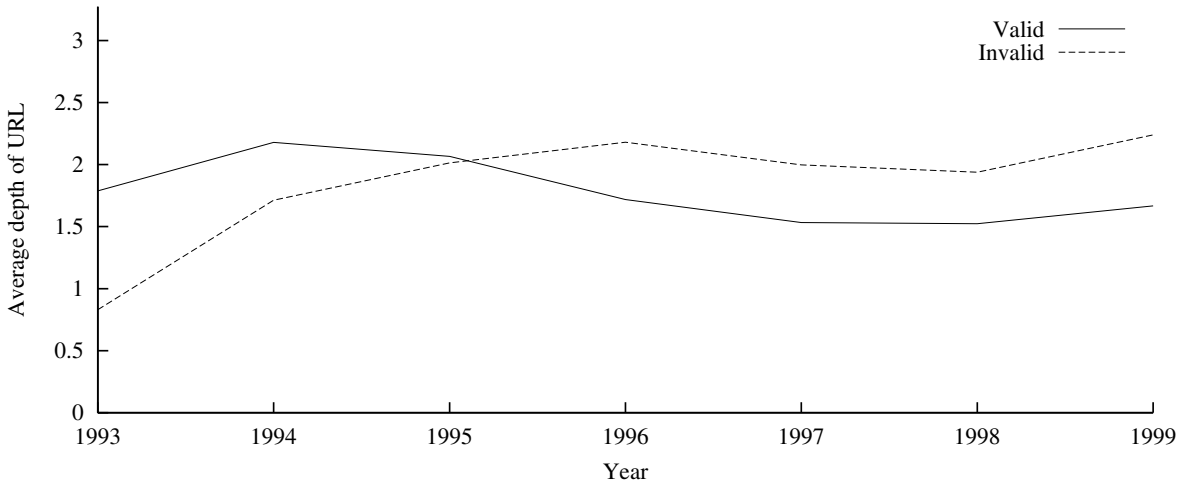


Figure 5: The average depth of URLs as a function of the year of publication of the articles.

als the URL should not provide a reference to only a specific version of the manual. Version files often become unavailable when the software or manual is updated.

4. URLs that depend on a personal directory should be avoided.
5. URLs that depend on a machine or subnet name should be avoided. (Try to obtain a virtual redirect directly from the domain name).

9. FINDING THE NEW LOCATION OF INFORMATION THAT HAS MOVED

Individual searchers used different strategies when attempting to find relocated and related information. The Google, ResearchIndex, and Inquirus search engines were most commonly used. Inquirus is a metasearch engine that combines the results of several regular search engines. Different engines tend to index different sets of web pages, and combining the results of multiple search engines can significantly improve coverage of the web [14]. Other search engines used include Northern Light and AltaVista.

Some common methods that we found useful include searching for the title of a document (typically as a phrase) if known, and/or the authors if known. The context of citations can be examined to generate alternative queries, for example company, institution, or project names. Browsing within a site may be attempted from an alternative starting point (e.g. the top level page or a researcher homepage), in order to locate pages that may have moved to a different location on the same site. Guessing possible new locations may be attempted, for example academic software may have moved from a specific machine to its own domain, or the homepage of an individual may have changed to the standard ~ notation. If a site has its own search engine this can be tried. If a URL has a relatively unique component then a search for this component may be attempted.

10. ENFORCING LINK CONSISTENCY

Alternatives to the World Wide Web such as Xanadu [17, 18] enforce the consistency of links. However these systems are not used widely like the web [26]. We argue that part of the reason for the success of the web may be the relative lack of requirements on the part of authors. A system that includes features such as link consistency may impose too much overhead and added complexity that limits acceptance.

Rather than enforcing link consistency, which may make participation in the web more difficult by increasing resource requirements or system complexity, we recommend promotion of improved practices for citing URLs, use of services like PURL, the availability of archives of the web which can be very valuable (see Brewster Kahle's Internet Archive (www.archive.org) [10, 1]), and the introduction of services that attempt to track and monitor URLs that move (such a service might be integrated with

an archive that has the full text of invalid links in order to locate new pages).

Our personal view is that it is not practical in the long term to expect individuals or small organizations to provide on-line resources, either passively, or actively, by maintaining indirection services to guarantee access to materials. Such material is likely to move or disappear eventually.

The problem of materials that move on the web can in principle be solved using technical means. Search engines could use signatures based on the content of materials to detect material that has moved [19]. In fact, the ability of search engines to perform resolution of dead links could become an added feature for differentiation. However, to solve the general problem of persistence and disappearance, we believe that technical solutions and peer policies will have to be combined. Professional societies such as the IEEE and ACM could help by proposing and enforcing acceptable standards for citations. Ideally, it should be required that all cited materials (or those important to building on or verifying the work) be available from a stable repository, such as the aforementioned Internet archive. These societies could also aggressively promote preprint and software repositories.

11. CONCLUSIONS

We performed a study that showed that many URLs referenced in academic articles are invalid, even for articles published within the last year. Fortunately, much of the information referenced in these invalid URLs can be located by guessing URL modifications, using search engines, or in accompanying formal citations. Very few URLs that were important to a given paper were lost, suggesting that scientific progress is not currently being significantly affected by lost URLs. However, there were significant differences in the ability of different searchers to find relocated information. Additionally, current citation practices suggest that more important URLs may be lost as time goes by. We detailed methods by which link consistency can be encouraged, as well as improved practices for citing URLs in research articles.

12. REFERENCES

[1] A snapshot of cyberspace. Library of Congress Bulletin 57(11), Nov 1998.

[2] W. Arms, C. Blanchi, and E. Overly. An architecture for information in digital libraries. D-Lib Magazine, February 1997. `http://hdl.handle.net/cnri.dlib/february97-arms`, `http://www.dlib.org/dlib/february97/cnri/02arms1.html` (Search for CNRI, Magazine of Digital Library Research).

[3] Paul De Bra and Geert-Jan Houben. A formal approach to analyzing the browsing semantics of hypertext. In *Proceedings of Computation and Neural Systems (CNS94)*, Monterey, CA, July 1994.

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Seventh International World Wide Web Conference*, Brisbane, Australia, 1998.

[5] Yuan Chen, Jan Edler, Andrew Goldberg, Allan Gottlieb, Sumeet Sobti, and Peter Yianilos. A prototype implementation of archival intermemory. In *Proceedings of the Fourth ACM Conference on Digital libraries (DL '99)*, 1999.

[6] M.L. Creech. Author-oriented link management. *Computer Networks and ISDN Systems*, 28(7–11):1015–25, 1996.

[7] Joseph Gibaldi. *MLA Style Manual and Guide to Scholarly Publishing*. MLA: New York, 1999. ISBN 0873526996.

[8] A. Goldberg and Peter N. Yianilos. Towards an archival intermemory. In *Proceedings of IEEE Advances in Digital Libraries, ADL 98*, pages 147–156, Santa Barbara, CA, 1998. IEEE Computer Society.

[9] D. Ingham, S. Caughey, and M. Little. Fixing the "broken-link" problem: the W3Objects approach. *Computer Networks and ISDN Systems*, 28(7–11):1255–68, 1996.

[10] Brewster Kahle. Preserving the Internet. *Scientific American*, March 1997.

[11] F. Kappe, K. Andrews, and H. Maurer. The Hyper-G network information system. *Journal of Universal Computer Science*, 1(4):206–220, 1995.

[12] Steve Lawrence, Kurt Bollacker, and C. Lee Giles. Indexing and retrieval of scientific literature. In *Eighth International Conference on Information and Knowledge Management, CIKM 99*, pages 139–146, Kansas City, Missouri, November 1999.

[13] Steve Lawrence and C. Lee Giles. Context and page analysis for improved web search. *IEEE Internet Computing*, 2(4):38–46, 1998.

[14] Steve Lawrence and C. Lee Giles. Accessibility of information on the web. *Nature*, 400(6740):107–109, 1999.

[15] Steve Lawrence, C. Lee Giles, and Kurt Bollacker. Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6):67–71, 1999.

[16] S. Mace, U. Flohr, R. Dobson, and T. Graham. Weaving a better web. *Byte*, 23(3):58, 1998.

[17] T. Nelson. The Xanadu project. *Byte*, pages 299–300, September 1990.

[18] Theodor Nelson. *Literary machines*. Mindful Press, Sausalito, CA, 1993. ISBN 089347052X.

[19] T. Phelps and R. Wilensky. Robust hyperlinks cost just five words each. Technical Report UCB//CSD-00-1091, University of California, Berkeley, January 2000.

[20] David S. H. Rosenthal and Vicky Reich. Permanent web publishing. In *Freenix*, San Diego, CA, June 2000.

[21] R. Schwartz. Uniform resource identifiers and the effort to bring "bibliographic control" to the web: an overview of current progress. *Bulletin of the American Society for Information Science*, 24(1):12–14, 1997.

[22] Keith Shafer, Stuart Weibel, Erik Jul, and Jon Fausey. Persistent Uniform Resource Locators. `http://www.purl.org/` (OCLC Online Computer Library Center).

[23] J. Shavlik and T. Eliassi-Rad. Intelligent agents for web-based tasks: An advice-taking approach. In *AAAI/ICML Workshop on Learning for Text Categorization*, 1998.

[24] J. Simonson, D. Berleant, X. Zhang, M. Xie, and H. Vo. Version augmented URIs for reference permanence via an Apache module design. *Computer Networks And ISDN Systems*, 30(1-7):337–345, April 1998.

[25] K. Sollins and L. Masinter. Functional requirements for uniform resource names. Internet Request for Comments, RFC1737, `http://www.cis.ohio-state.edu/htbin/rfc/rfc1737.html`, December 1994.

[26] G. Wolf. The curse of Xanadu. *Wired*, pages 137–202, June 1995. `http://www.wired.com/wired/archive/3.06/xanadu.html`.