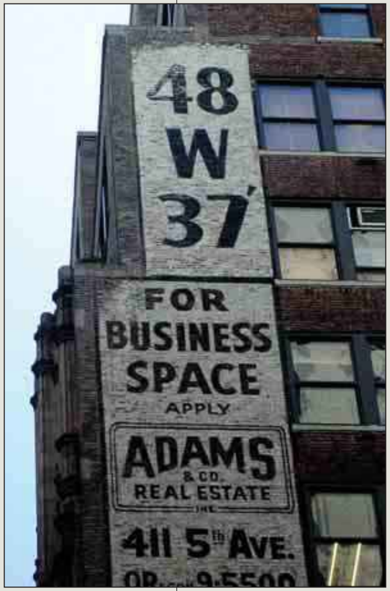
The Decay and Failures of Web References

Attempting to determine how quickly archival information becomes outdated.

By Diomidis Spinellis



The widespread adoption of the Web as a mechanism for sharing information has brought with it the corresponding ubiquity of URL references and citations. URLs regularly appear on billboards, packages, business cards, print advertisements, clothing, and as references in scientific articles. Most readers have probably experienced a "dead link": a Web reference that for a variety of reasons will not lead to a valid or correct Web page. A dead link stemming from a URL appearing in the context of everyday life is usually a minor inconvenience that can be resolved by using a Web index or a search engine; it will seriously affect only the future archeologists trying to untangle the web of our daily lives. On the other hand, a dead Web link appearing in a scientific article has wider implications. Citations in scholarly work are used to build upon existing work, substantiate claims, provide the context in which research is performed, and present, analyze, and compare different approaches or methodologies. Therefore, references that cannot be located seriously undermine the foundations of modern scientific discourse.

The objective of this article is to examine, quantify, and characterize the quantity and quality of Web links used in computing literature. Our aim is to provide definitive information related to the availability of URL references as a function of their age, their domain, the depth of the path used, as well as the technical reasons leading to failed links. Our research has been greatly aided by the emergence of online versions of traditional paper-based publications [4]. By tapping into the online libraries of the ACM and the IEEE Computer Society we were able to download, extract, and verify 4,375 Web links appearing in print articles during the period from 1995–1999. Here, we describe the technologies related to Web references and retrieval, outlining the methodology we followed, presenting the results obtained, and discussing their implications.

Internet resources are typically specified using the string representation of Uni-

PHOTOGRAPHS BY MICHAEL KLOSE

form Resource Locators. URLs are a subset of the Uniform Resource Identifiers (URIs) that provide an abstract identification of a resource location [3]. In general, URLs consist of a scheme (such as http, ftp, or mailto) followed by a colon and a scheme-specific part. The syntax of the scheme-specific part can vary according to the scheme. However, URL schemes that involve direct use of an IP-based protocol to an Internet host use the following common syntax:

//<user>:<password>@<host>:<port>/<url-path>

The double slash indicates the scheme data complies with the Internet scheme syntax. In our sample, over 98% of the URLs we encountered used the HTTP URL scheme that designates Internet resources accessible through HTTP (HyperText Transfer Protocol). The host is specified using the fully qualified domain name of a network host or its IP address. The default port for the HTTP scheme is port 80 and is usually omitted.

For a Web page of a given URL to appear on a browser's screen a number of different technologies and protocols must work in concert. In addition, the changing realities of the Web and the Internet have vastly complicated the simple end-to-end request-reply protocol that formed the basis of the early HTTP transactions. Any failure along the complicated chain of actions needed to retrieve a Web page will lead to a failed URL reference.

When a Web page is accessed for the first time, the name of the host must be resolved into a valid IP address. Although before the emergence of the Web there was typically a one-to-one correspondence between the IP address of non-routing hosts and a domain name the situation now is more complicated. Many hosts respond to different IP addresses, associating a different Web site with each address (virtual IP-based hosting). As IP addresses are becoming a limited resource it is also common to associate many domain names for the same host and IP address and serve different Web sites based on the host name used for a given page request (virtual name-based hosting). Finally, Web sites receiving a large amount of traffic may associate different IP addresses and hosts for the same domain name in order to distribute traffic among hosts.

The path appearing in a URL will not necessarily match with a corresponding local file on the server. Web servers provide a number of mechanisms for managing namespaces; some include: the creation of a separate namespace for every local user, the definition of protection domains and access mechanisms, the support of aliases to map namespaces to local directories, and the dynamic creation of content—using technologies such the common gateway interface and active server pages. In addition, a feature of the HTTP protocol known as content negotiation allows a server to provide different pages based on technical or cultural characteristics specified in the incoming request (such as bandwidth, display technology, and languages the user can understand).

One final complication results from the fact that Web transactions seldom follow the idealized IP path from the user host to the server. Both ends are likely to be protected by firewalls actively blocking or even modifying content that passes through them. At the user end, routers utilizing network address translation (NAT) mechanisms as a way to manage a limited pool of IP addresses are likely to hide the IP address of the end host from the server. Finally, proxy servers—working either in cooperation with the end user or transparently intercepting requests—may cache documents and serve them without communicating with the original server.

Any failure along the complex path we described will often result in a failed request for a URL. The HTTP protocol defines 24 different errors that can occur within an HTTP exchange. In addition, some errors can occur before the client and server get a chance to communicate. In practice, while verifying thousands of URLs we encountered the following errors.

400 Bad Request. The syntax used for the request could not be understood by the server. In our case this error typically signifies incorrectly typed URLs.

401 Unauthorized. The request requires user authentication. Such an error can result when citations are given to URLs that exist within a domain of services that require registration, or when such services move from a free-access to a registration-based model. It is debatable whether this return code classifies an access as a failure. A number of digital library services are increasingly provided over the Web on a subscription basis; lacking authorization to access such a service is similar to material not being available in the local library.

403 Forbidden. The server is refusing to fulfill the given request, in this case, however, proper authorization cannot be used to retrieve the page. It is conceivable that URLs that are not part of the public Internet are used as citations when the authors fail to realize they have special privileges to access certain repositories that are not available to the global Internet population. As an example, our organization has transparent access to a collection of online journals with authentication based on the client IP address. URLs to this collection provided by unsuspecting users will typically generate a 403 error.

404 Not Found. This infamous and common response indicates the server has not found anything matching the requested URI. This error is typically generated when Web site maintainers change file names that are part of the given URL path or entirely remove the referenced material. This protocol error can be followed by customized content—typically HTML text informing the user of the problem and offering alternative navigation options.

500 Internal Server Error. The server encountered an unexpected condition that prevented it from fulfilling the request. This error can occur when a server is wrongly configured, or, more commonly, if a program or database that is used to serve dynamic content fails.

503 Service Unavailable. The server is currently unable to handle the request due to temporary overloading or maintenance of the server. Errors of this type sometimes appear on a misconfigured server, or servers overwhelmed by traffic.

504 Gateway Time-Out. The server, acting as a proxy or gateway, did not receive a timely response from the upstream server specified by the URI (HTTP, FTP) or some other auxiliary server (domain name server—DNS) it needed to access in attempting to complete the request. When HTTP requests are transparently intercepted by a proxy caching server, network connectivity problems are likely to appear as 504 errors.

901 Host Lookup Failure. The host name could not be mapped to an IP address. This error (which is not part of the HTTP protocol) signifies a problem in retrieving the IP address of the server using the DNS services. Likely causes include changes of host names and DNS server failures or connectivity problems.

Methodology

Our research involved the verification of URL references appearing in published material available from the ACM digital library and the IEEE Computer Society digital library; it involved the following steps:

1. "Crawl" each site and download published articles;
2. Convert the articles into text;
3. Extract URLs from the articles; and
4. Verify URL accessibility.

Although a large number of document collections are available on the Web, we limited our research to two publications: *IEEE Computer* (*Computer*), and *Communications of the ACM* (*CACM*). Our decision was based on the fact that both publications:

- Are available in electronic and paper formats;
- Are distributed to a wide audience and are widely accessible;
- Are published in regular and frequent intervals; and
- Contain articles from a diverse set of information and computer science disciplines and authors.

However, by concentrating our research on two publications we limited the generality of the obtained results. *CACM* and *Computer* do not represent the typical publication as, contrary to common practice, editors verify the URLs before publishing an article thus filtering out invalid URLs submitted by the authors or invalidated during the period leading to the publication. In addition, publications from other scientific domains, or with a different focus such as archival research journals or general circulation magazines, are likely to exhibit different characteristics regarding the appearance of URLs and their validity. To allow other researchers to build upon our work, we have made available the complete set of URLs, their source, and the programs used to verify their accessibility at: www.spinellis.gr/sw/url-decay.

We first downloaded all articles appearing in the two publications using a set of programs that crawled through the digital libraries of the two organizations. This phase occurred from February 2000 to May 2000. Over 9GB of raw material was downloaded during the process.

At the time of the study *CACM* articles were available from the ACM digital library in PDF format. In order to extract URLs we first converted the articles into text form. *CACM* articles appearing in the library before 1995 were scanned images; we did not attempt to OCR those items. Because 1995 was also the earliest year in which *Computer* was available online we decided to use the articles from the period 1995–1999. In total we used 2,471 items: 1,411 articles from *Computer* (38.2MB of HTML) and 1,060 articles from *CACM* (18.9MB of text).

We extracted URLs from the full-text body of each article. The IEEE Computer Society digital library provides articles in both HTML and PDF format. Articles that appear in HTML format have embedded URLs tagged as hypertext references, which can be easily extracted. The extraction of URLs from the text of the *CACM* articles proved more challenging; it was performed using a custom filter program and manual inspection. (Editors note: *CACM* articles are now also available in HTML form in ACM's digital library.)

After extracting the URLs we removed duplicate appearances of URLs in the same article (21 cases for *CACM*, 362 for *Computer*). In total we collected 4,224 URLs: 1,391 (33%) obtained from *CACM* and 2,833 (67%) obtained from *Computer*. We found a mean number of 1.71 URL references per article (median 0, mode 0) with a maximum of 127 URL references in a single article. A single complete URL was referenced by a mean number of 1.49 (median 1, mode 1) different articles in our sample with a maximum of 22 references for a single URL. The HTTP scheme was by far the most widely used: 4,158 URLs (98%) used the HTTP scheme and only 66 URLs (2%) used the FTP scheme.

Finally, we verified the accessibility of each URL by attempting to download each identified resource. We repeated this procedure three times, with a weekly interval between different runs, starting at different times, from two different networks and hosts to investigate transient availability problems. No substantial differences were found between the runs. Here, we report the results obtained on June 29 and 30, 2000. We did not merge positive results from different runs; our results reflect a model in which a reader tries to access a URL one single time. We did not perform any semantic processing on the retrieved URLs; we assume that if a URL could be accessed its contents would match the intent of the original reference.

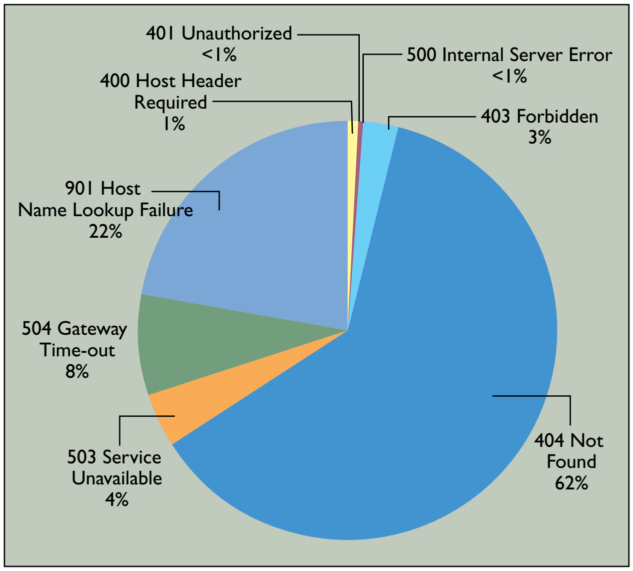


Figure 1. URL retrieval results.

Results

Despite our original reservations concerning the source material we used, the results we obtained have been corroborated by similar studies of Web-published documents [9]. Of the URLs we checked 72% could be retrieved without a problem. The successful retrieval rates differ depending on the URL source: 63% for *CACM* and 77% for *Computer* URLs. This difference can probably be attributed to the fact that *Computer* URLs are tagged as such in the HTML article text. The reasons for failed URL retrievals are classified in Figure 1. By far the most common reason was that the resource referenced no longer existed on the given server (error 404, 60% of the failures). The second most common (22%) failure reason was an invalid host name (error 901), while network problems (error 504) only represented 8% of the failures; a tribute to network availability. It is interesting to note that 83% of the failures can be attributed to invalid URL host names or paths (errors 901 and 404), indicating that addressing in all its forms is the predominant factor in URL failures.

The clustering of failure modes allows us from this point onward to classify failed URLs into just two different groups: the *network problems* that occur while trying to reach the host (errors 504 and the DNS access subset of the 901 errors) and the *server problems* that occur while resolving the host name and once the host is reached.

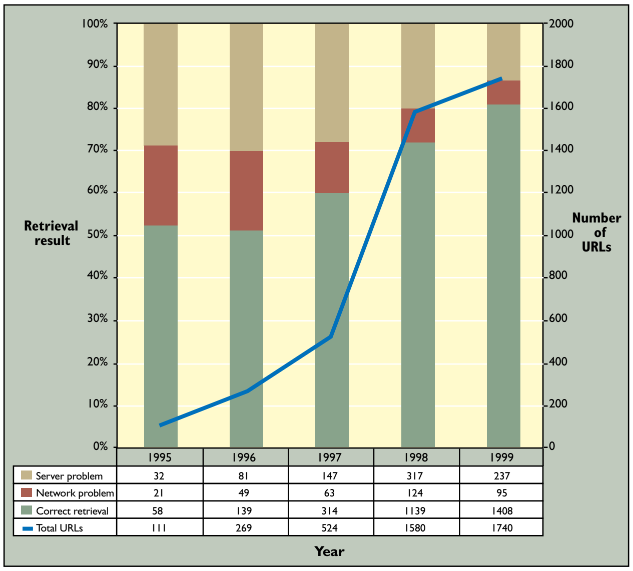
The temporal aspect of URL references and the respective failures is extremely interesting. As can be seen in Figure 2, URL references exhibited an exponential increase from 1995 to 1998 and appear to be leveling off afterward. This plateau is to be expected since the number of references in articles is constant over time (editors of the examined publications often impose limits); URL references apparently increased by displacing citations to printed sources. The most striking result that can be deduced from Figure 2 is that in our sample

the half-life of a referenced URL is approximately four years from its publication date

that is, that four years after their publication about 50% of the URLs are not accessible. It is also interesting to note that 20% of the URLs are not accessible one year after their publication and after the first year the URL decay is constant at about 10% per year for the next three years. Although URL decay appears to stabilize after that point, (a result that appeals to intuition—these will be URLs to authoritative sources on properly maintained servers) we have insufficient historical data to substantiate this claim. The 20% decay during the first year can either be attributed to a high infant URL mortality or the long period of time an article requires from its inception to its publication.

We were able to repeat the URL accessibility test two years after the 2000 exercise, in August 2002. The combined results of the two tests appear in Figure 3. What is apparent is the inexorable decline of the accessibility of the 1998 four-year-old URL references toward the 60% mark (as opposed to the 50% we originally predicted), the further decline of the

1995 and 1996 URL accessibility toward 40%, and a small but significant difference between the aging behavior of older and newer URLs. For the years for which we have comparable data (URL ages 3 to 5 years) more recent URLs (coming from the years 1997–1999 in the 2002 test) appear to be more accessible than their predecessors (years 1995–1997 in the 2000 test). This difference in URL aging over time can probably be attributed to increased author efforts to cite URLs that are less likely to disappear, and improved Web site maintenance practices.



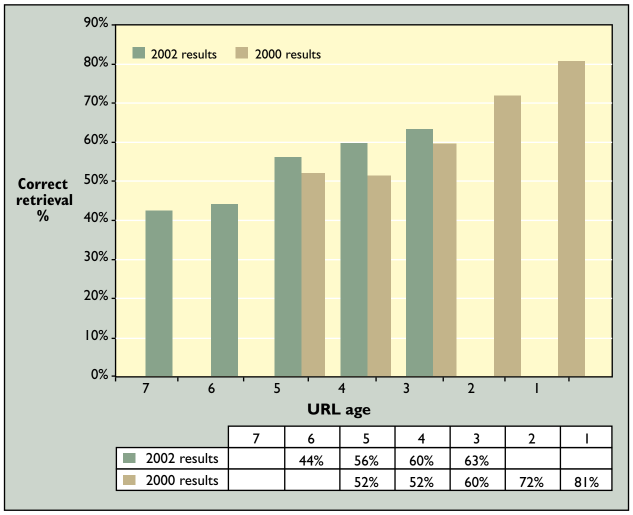
	1995	1996	1997	1998	1999
Server problem	32	81	147	317	237
Network problem	21	49	63	124	95
Correct retrieval	58	139	314	1139	1408
Total URLs	111	269	524	1580	1740

Figure 2. URLs and retrieval results by year.

One result that appears to have both predictive and prescriptive value concerns the relationship between the path depth of a given URL and its failure rate. As can be seen in Figure 4, the number of published URLs of a given path depth is linearly decreasing between path depths 0 and 2; it appears to decrease at a exponential rate after that point. What is significant is that while the network-related problems are, as expected, approximately constant relative to the path depth, server-related problems (mainly 404 errors) increase as the depth increases. While no direct causation can be deduced, we can point out that

deep path hierarchies are linked to increased URL failures.

This result is not immediately intuitive. A deep hierarchy is a sign that someone made an effort to organize content in a structure that should persist over time. We have two explanations for this result:



	7	6	5	4	3	2	1
2002 results		44%	56%	60%	63%		
2000 results			52%	52%	60%	72%	81%

Figure 3. URL decay from different time perspectives.

1. Each element of a URL path has a constant probability to fail due to changes of names, organizational structure, and maintenance personnel. These failures accumulate over all elements of the path.
2. URLs with a short path are more likely to be cited than those with a longer path, therefore site administrators try to keep them alive.

We also examined the relationship between two other URL properties: references to specific files and user directories, and the respective failure rates. We identified URLs that referenced specific files (for example, www.acm.org/pubs.html) rather than directories (for example, www.acm.org/ cacm) assuming that if the last part of the URL contained a dot it referred to a file. The difference between the two URL classes is noteworthy: in total, 40% of URLs referring to files could not be retrieved whereas only 23% of URLs referring to directories had the same problem.

Some HTTP servers allow the specification of separate content directories maintained by end users using the ~username convention. We hypothesized that URLs to such user directories (which we found to be 13% of the total) were more likely to fail than others due to the higher mobility of individuals. In fact only 24% of these URLs had retrieval problems; the respective figure for the rest was 28%.

In Figure 5 we list the retrieval result of the referenced URLs according to their top-level domain. The domains .com, .edu, and .org represent 74% of all referenced URLs. Other studies [7] have estimated that on a global scale only 6% of the Web servers have scientific or educational content. Since in our sample .edu domain URLs comprise 23% of the total, we can deduce that these URLs are referenced in computing articles three times more frequently than what would be expected by their population. It is interesting to note that URLs in the .com and .edu domains are equally likely to fail; a startling result given the radically different management models prevalent in educational establishments and companies. Also remarkable is the fact that URLs in the .org domain are less likely to fail than the other two categories; contrary to intuition, it appears that the management structures used by the volunteer efforts typically hosted in .org domains result in slightly more stable Web sites.

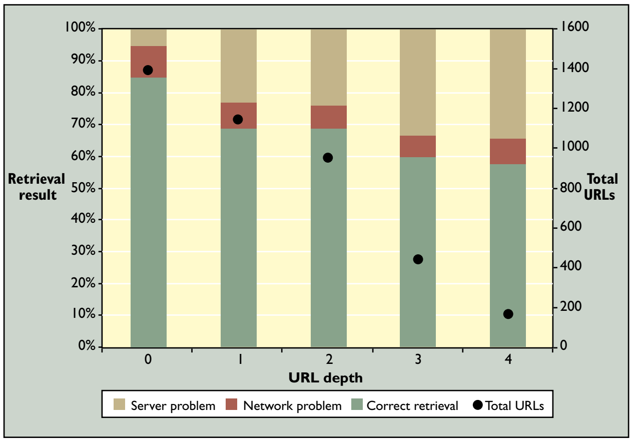


Figure 4. URLs and retrieval results by URL depth.

Improving URI Longevity

Problems with links becoming out of date are not new, they have been with us since the emergence of hypertext structures. A number of schemes have been suggested for maintaining error-free hypertext structures [1]. However, the unique problem posed by URLs appearing in print publications stems from the uneasy coexistence of two radically different affordances paradigms: hypertext-based electronic publishing and paper media. Once a URL is committed to paper it cannot be modified and it might be difficult to locate and trace.

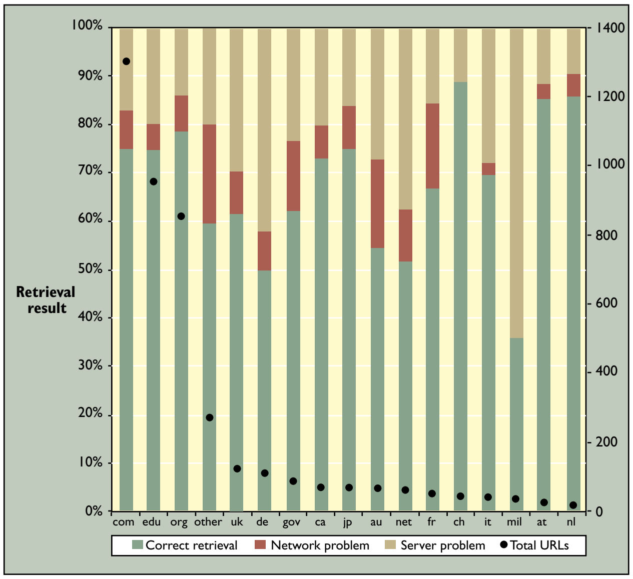


Figure 5. URLs and retrieval results by top-level domain.

The emergence of publications that appear both electronically and on paper [6] can help alleviate the tensions between the two formats. It has certainly helped us trace URL references, and we can envisage a system that would keep hypertext references up to date in the non-paper part of hybrid publications. In the future, citation linking [5], provided as a service by publishers or public-service efforts [8], may lead to publication formats that actively support hypertext links across time.

Uniform Resource Names (URNs) have been proposed as a way to provide persistent, location-independent, resource identifiers [11]. However, URNs—typically consisting of an authority identifier followed by a string—are at a low level similar to URLs. Organizations that find it difficult to organize URLs will face the same problem with URNs and vice versa [2]. While URNs can solve the problem of maintaining connection with a moving target, they cannot solve the problem of accessing deleted material.

A technology specifically targeting the persistent and interoperable identification and exchange of intellectual property in the digital environment is the Digital Object Identifier (DOI) system [12]. A DOI, consisting of a publisher prefix and a suffix identifying the work, is registered together with the corresponding URL and metadata in a central DOI directory working as a routing system. DOI-based requests are forwarded to the directory and are resolved to a current and valid URL supplied by the respective rights holder. Based on this technology, a large number of publishers (including the ACM and the IEEE) accounting for over 3,500 journals have teamed up to provide CrossRef, a reference linking service.

One alternative way for reestablishing contact with invalid URL references is to use one of the Web's main search engines. However, research indicates that search engine coverage is low (around 16%) and decreasing, indexing varies considerably between different sites (with educational and non-U.S. sites more likely to be disadvantaged), while the use of metadata that could be used to automatically locate citations is quite low [7]. The same research estimates the amount of data in the publicly indexable Web at 15Tbytes; we therefore believe the creation of frequent historical "snapshots" of the Web is not within the realm of our current technical capabilities.

Based on our results, ways to alleviate the specific problem of invalid URL references appearing in print articles can be identified by concentrating on the distinct roles of the principal stakeholders involved in the process. Professional societies and publishers should draft and encourage the use of sound Web citation practices and verify referenced URLs as part of the article editing process. Both should also work toward establishing online repositories for Web material (such as the ACM Computing Research Repository) and endow them with policies to guarantee their longevity in the future. Publishers of archival research journals have an additional obligation toward the future generations that will access their material. Although some URLs are less important than others, a guideline limiting Web citations in archival publications to organized collections with concrete long-term retention policies may be the only responsible action consistent with the current state of the art.

Researchers should appreciate the limitations of Web citations regarding their probable lifespan and use them sparingly rather than gratuitously, keeping in mind the Web is not an organized library. Where possible, they should prefer citing the published version of a work to its online version, and citing material in organized collections over material in corporate or personal Web pages. In some cases they could even persuade authors of material they would like to reference to move it to an organized online repository. In addition, researchers should reference items using the shortest possible path, and avoid references to ephemeral data (such as non-archived news reports) and proprietary material.

Finally, maintainers of Web sites should try to preserve the validity of existing URLs and clearly indicate ephemeral pages that should not be cited (or linked). The standardization of appropriate HTML meta tags for indicating the projected longevity of a particular page will help all the stakeholders unambiguously identify the extent to which a page can be cited.

A more ambitious scheme would involve having all stakeholders cooperate to establish a long-term archive of referenced material, similar in nature to existing Internet archival efforts such as www.archive.org. Citations would reference the archived version of Web material: archive.acm.org/2002.11.17/www.ibm.com/ai.htm. Under such a scheme libraries and professional societies would establish and promote the use of archival services perpetually keeping copies of referenced material—subject to intellectual property restrictions. Researchers would cooperate with Web-site producers to obtain copies of their material for archiving while Web-site producers should be encouraged to draft and implement liberal policies for placing cited material under long-term archival custody.

Conclusion

Approximately 28% of the URLs referenced in *Computer* and *CACM* articles between 1995 and 1999 were no longer accessible in 2000; the figure increased to 41% in 2002. In addition, after four years 40–50% of the referenced URLs become inaccessible. A noteworthy parallel can be observed between the four years we calculated as the half-life of referenced URLs and five years given as the median citation age for computer science [10]. One could claim that the self-organizing nature of the Web filters out irrelevant URLs at approximately the same rate as those have traditionally been rendered obsolete in articles appearing in print.

The Web has revolutionized on a global scale the way we distribute, disseminate, and access information and, as a consequence, is creating a disruptive paradigm shift in the way human scientific knowledge builds upon and references existing work. In the past, libraries could provide reliable archival services for books and other printed publications; the emergence of the Web is marginalizing their role. In the short-term none of the approaches toward solving the general problem of dangling URL references is likely to be a panacea. It is therefore important to appreciate the importance of Web citations and invest in research, technical infrastructures, and social processes that will lead toward a more stable scientific publication paradigm. C

References

1. Ashman, H. Electronic document addressing: Dealing with change. *ACM Computing Surveys 32*, 3 (Sept. 2000), 201–212.
2. Berners-Lee, T. Cool URIs don't change. *Current* June 2002, 1998; www.w3.org/Provider/Style/URI.
3. Berners-Lee, T., Masinter, L., and McCahill, M. RFC 1738: Uniform resource locators (URL), December 1994.
4. Denning, P.J. The ACM digital library goes live. *Commun. ACM 40*, 7 (July 1997), 28–29.
5. Hitchcock, S., Carr, L., Harris, S., Hey, J.M.N. and Hall, W. Citation linking: improving access to online journals. In *Proceedings of the 2nd ACM International Conference on Digital Libraries*, (Philadelphia, PA, July 1999), 115–122.
6. Kling, R. and Covi, L. Electronic journals and legitimate media in the systems of scholarly communication. *The Information Society 11*, 4 (1995), 261–271.
7. Lawrence, S. and Giles, C.L. Accessibility of information on the Web. *Nature 400* (1999), 107–109.
8. Lawrence, S., Giles, C.L., and Bollacker, K. Digital libraries and autonomous citation indexing. *IEEE Computer 32*, 6 (June 1999), 67–71.
9. Lawrence, S. et al. Persistence of Web references in scientific research. *IEEE Computer 34*, 2 (Feb. 2001), 26–31.
10. Meadows, A.J. *Communicating Research.* Academic Press, 1998, 221–222.
11. Moats, R. RFC 2141: URN syntax, May 1997.
12. Paskin, N. E-citations: Actionable identifiers and scholarly referencing. *Learned Publishing 13*, 3 (July 2000), 159–168.

Diomidis Spinellis (dds@aueb.gr) is an assistant professor in the Department of Management Science and Technology at Athens University of Economics and Business in Greece.