



**VICTORIA UNIVERSITY**  
MELBOURNE AUSTRALIA

*The Impact of Impermanent Web-Located Citations:  
A Study of Scholarly Conference Publications*

This is the Published version of the following publication

Sellitto, Carmine (2005) The Impact of Impermanent Web-Located Citations: A Study of Scholarly Conference Publications. *Journal of the American Society for Information Science and Technology*, 56 (7). pp. 695-703. ISSN 1532-2882

The publisher's official version can be found at  
<http://dx.doi.org/10.1002/asi.20159>

Note that access to this version may require subscription.

Downloaded from VU Research Repository <https://vuir.vu.edu.au/336/>

**The impact of impermanent Web-located citations: A study of scholarly  
conference publications**

Carmine Sellitto

Lecturer

School of Information Systems  
Faculty of Business and Law (FP),  
Victoria University,

**Correspondence address:**

Faculty of Business and Law (FP),

Victoria University,

PO Box 14428 MCMC,

Melbourne, Victoria, 8001, Australia.

Telephone: 61 3 9688 4341 Fax: 61 3 9688 5024

**Contact email:**

[carmine.sellitto@vu.edu.au](mailto:carmine.sellitto@vu.edu.au)

# **The impact of impermanent Web-located citations: A study of scholarly conference publications**

## **Abstract**

*This paper reports on research that examined the viable nature of 1068 Web-located citations in 123 academic conference papers published between 1995 and 2003. The study appears to be one of the few but increasing number of investigations that examines the growing practice of authors citing URLs in their publications to support and argue their scholarly research. The research found that some 46% of all citations to Web-located sources could not be accessed— with the HTTP 404 message (61.5%) being the greatest cause of missing citations. Collectively, the missing citations accounted for 22.0% of all citations, which represents a significant reduction in the theoretical knowledge base underpinning many scholarly articles. Furthermore, this paper argues that the consequences of disappearing Web-located citations has led to diminished opportunities for future researchers to examination the underlying foundations of discourse and argument in scholarly articles.*

## **INTRODUCTION**

The Web has become the preferred medium for delivering business and organisational information to its constituency, allowing information to be effectively collated and presented in a useful form. The powerful publishing features of the Web are evident in the informational models that have been proposed by the general information systems and information science literature (Davenport 1997; O'Brien 2001; Turban *et al.* 2002). According to early work by Nunberg (1996), the migration of digital publications to the Web environment has resulted in enhanced information

availability— information published in the virtual domain being also independent of medium or location. Traditional publications, including written text, diagrams or images have relied on a physical medium (tablet, papyrus, scrolls, parchment, film, paper, etc) as the vehicle for information delivery, with access to that information being invariably tied to a point in time (Landlow 1996). Thus, the creation of these traditional documents based on time and media-type allows them to exhibit stability and permanence as an important and defining feature.

As the amount of available information has increased via the Web, so has the corresponding use of citations to Web-located sources by writers of reports and scholarly works (Spinellis 2003). Indeed, the permanency of these cited Web resources has been questioned by various authors, with disappearance rates of cited Uniform Resource Locators (URLs) for scholarly journal articles having been reported (Germaine 2000; Lawrence *et al.* 2001; Rumsey 2002; Spinellis 2003). The investigation of permanency associated with Web-located references in conference articles appears to have been overlooked, or has been a secondary focus— this may be due to a perception that conference publications do not appear to have the same prestige and scholarly value as journal items. The importance of conference publication is such that it provides avenues for less established or new researchers to publish their work, as well as fostering a formal and informal scholarly and collaborative environment. This study explores the permanence of Web-located references that have been cited in peer-reviewed conference articles and discusses the impact that missing citations may have on the theoretical base of an article as well as on future investigators that may reference that article.

## **BACKGROUND**

### *The advent of citation*

According to Czarniawska-Joerges (1998), the practice of referencing commenced to appear in written works by the end of the seventeenth century— a practice that was invariably associated with the nascent university and academic professions at the time. Previous to this time, books may have been commonly annotated with new additions and altered to keep them updated— people wanted books to be complete and originality had no real perceived value. The Middle Ages were also not favourable to new scientific exploration and ideas— it was a period of faith. Any search for understanding of why and how things worked was related back to the religious beliefs— it was so because God had made it that way (Prostan 1958). It appears that in this period God’s word was the only required reference.

Veyne (1988) indicates that part of the interchange amongst the seventeenth century scholarly community involved the soliciting of collegial comment on new ideas and theorems— possibly a preliminary practice of the modern-day academic peer-view applied to publications. New works might have been considered controversial, hence the onus was on the writer to justify arguments and inferences to support these new ideas. Consequently, the bibliographical reference and the practice of citation appears to have evolved as an important feature that tended to give credence and justification to an author’s original thoughts, theorems, models and/or assertions. Grafton (1997) chronicles the history of the citation and referencing inferring it is a subtle way of documenting the progress of knowledge, as well as portraying the evolution of modern scholarship.

Today, the use of quoting and citation could be viewed as merely a technical function with appropriate rules of style and structure. However, citation practice in the academic literature review allows the author to support and buttress their own ideas—a process which also positions their work in context with others (Webster and Watson 2002; Zerby 2002). Citation also signals an author's awareness of ethical publishing principles and behaviour that is commensurate with recognising the knowledge ownership of other writers.

### ***Citation and the scholarly literature review***

A review of existing literature relating to a topic of interest is a fundamental first step to investigating an academic research project. The literature review allows the researcher to identify, analyse, discuss and critique previously published theoretical works applicable to their own research (Kellehear 1993). Furthermore, a review of relevant literature creates the foundations for advancing knowledge in that it can identify areas where research is needed through new initiatives, or it may act to reinforce existing theory (Hart 1998). The scholarly literature review, whether in a short monogram, descriptive report, academic paper or PhD thesis is, therefore, underpinned by previous works—a process enacted by correct use of referencing. Baker (2000) reiterates that the advancement of knowledge and ideas is incremental with investigations leading to new knowledge discoveries reliant on the work of previous researchers— hence, today's investigators *stand on the shoulders of yesterdays researchers to be able to see further*. The shoulders of others being directly evident in the citations used. Baker also indicates that the correct use of citations is vital in scholarly research, and calls for the unambiguous identification of

citations that allows the reader the opportunity to further investigate original or primary cited documents.

Newby and Ertmer (1997) suggest that referencing supports an authors thoughts and ideas by using previously documented statements that provide argumentative validity. Moreover, a list of citations provides an opportunity for readers to locate and explore further any topic of interesting. According to Ticehurst and Veal (2000) the practice of proper referencing displays a writer's scholarship and also provides evidence of association with a previous body of knowledge. Furthermore, Ticehurst and Veal propose that with appropriate citation readers have an opportunity to locate and check sources to verify the researcher's interpretation of previous work. Some authors suggest the specific use of citations can be used explicitly to expand and provide a theoretically grounded base on which to undertake research. For example, Webster and Watson (2002) strongly advocate academic scholars *go backwards* to review the citations found in leading and significant works of interest. Metcalfe (2003) proposes that a critical appraisal of the literature review's foundations can be achieved by treating any citation as a form of argumentative *evidence*, in much the same way that evidence is considered in a judicial process. This evidence-based analogy applied to references imparts rigour to a literature work and Metcalfe draws on the judicial constructs of expertise, authority and hearsay in which to view citations as supporting evidence.

Hence, the evaluation of the literature review's citation base is for many readers an important consideration in knowledge exploration and extension. Moreover, failure to acknowledge contributory works of others contravenes the ethical obligation that

authors have in making a form of symbolic payment to the owners of an idea or previous piece of work. A fundamental assumption is also made when considering any cited resource— book, journal article, government report or Web site — is that the source has a form of permanency in that it can be easily found, accessed, and evaluated.

### *The URL as an information resource reference*

The advent of the Web has altered the dynamics of information access in that information is no longer restricted to a physical presentation medium, nor is it reliant on geographical location. Consequently, any Internet linked computer allows researchers to access information that may have been previously unavailable to them. The ability to access information in this manner has allowed authors to substitute some of the traditional paper-based citations to books, journals, reports and notes with electronic alternatives. Moreover, with the vast quantity and easily accessible documentation available on the Web, many authors often cite Uniform Resource Locators (URLs) as part of the attribution process when it comes to acknowledging supporting material in their publications (Germaine 2000; Rumsey 2002; Spinellis 2003).

The URL represents a unique Internet locator of a digital information resource and can be written as four sets of dotted-decimal numbers (120.224.21.253) or as an alphanumeric string that constitutes an Internet protocol (IP) address (Powell 2003). As a machine-based addressing mechanism, the URL must be resolved to a valid Internet Protocol (IP) address, otherwise a Hypertext Transfer Protocol (HTTP) error message results. Powell (2003) lists numerous errors ranging from the often-found



Error 404 (Page not found), to the more complicated Error 500s which may be attributed to bad host name or server time-out. Table 1 outlines some of the common HTTP messages encountered (and used in this research) when attempting to access web pages.

**Table 1 HTTP messages encountered when retrieving web pages (Powell 2002 p. 514-517)**

Status Code	Description
301 (redirection code)	Moved Permanently. Indicates that the requested resource has been assigned a new address. This new address should be assigned to any future reference to this resource.
302 (redirection code)	Moved Temporarily. Indicates that the requested resource has been located at a new temporary address, however the original address is still suitable for accessing the resource.
403 (Client error code)	Forbidden: The URL request is understood, however, is disallowed.
404 (Client error code)	Not Found: Signifies a missing resource as a result of the URL not being found. Can be caused by an incorrectly typed URL string.
502 (Server error code)	Bad Gateway: The server encounters a problem associated with another gateway (network connection) that has passed the URL request along.
504 (Server error code)	Gateway Timeout: The server, after attempting to process the URL request, does not receive a timely response from another gateway (or some other DNS routing server) required to complete access to specific URL resource.

### *The published literature on URL permanency and disappearance*

The issue of URL permanency has become an important due to the way that Web-located information is being increasingly cited in both general and academic publications. When an author cites a resource located on the Web there is a fundamental assumption of resource permanency — that is, the particular information resource will be found at the cited location. However, given that Web-located resources are being increasingly cited, there has been concern on the way that URL references are disappearing (Rumsey 2002). The disappearing Web-resource is manifested through the growing incidence of broken links— or *link rot* as Neilsen (2000) coined— an issue that was flagged by Kahle (1997) who suggested that the average lifetime of a URL might be just 44 days. Various authors have subsequently reported on the problem of disappearing URLs (Koehler 1999; Davis and Cohen 2001; Koehler 2002; Markwell and Brooks 2002; Markwell and Brooks 2003), whilst some authors have examined the impact on scholarly works (Germaine 2000; Lawrence, Coetzee et al. 2001; Zhang 2001; Rumsey 2002; Spinellis 2003).

Koehler (1999) was one of the first to report on the concept of Web page permanency indicating that Web sites and pages underwent significant changes over a short period of time. Koehler's study tracked some 350 URLs from 1996 to 1998 and found that some 17.7% of web sites and 31.8% of web pages failed to respond when queried after 12 months. Moreover, the study also found that with the exception of all but one page in the sample, all pages experienced some form of content or structural change. The study also revealed that web sites and pages that had presumably disappeared sometimes reappeared, leading Koehler to conclude that some 5% of initially non-responding or missing pages would be found some time later. Subsequent studies by

Koehler (2002) examined the attrition and modification of Web sites/pages confirming many of the previous 1999 findings. Koehler indicates that the average web page *half-life* was approximately 2 years in 2002— that is, every two years, half the pages being tracked could not be accessed. Furthermore, Koehler introduced the notion of the *comatosed* Web page, a term that referred to a page temporarily disappearing, only to reappear on a later search— sometimes in a different format and with different content. The 2002 study also reported that navigational type pages had a better survival rate than pages rich in information content; and that younger Web pages are relatively unstable when compared to older published pages.

Markwell and Brooks (2002; 2003) examined Internet-based science information and suggest that even though Web-located science resources are freely available, they lack stability and permanence when compared to the traditional science textbook. The authors undertook a longitudinal study of Web-located science resources and determined the rate of disappearance of the 515 URLs associated with three science educational sites. Using a statistical coefficient equation they determined that the half-life of URLs on these sites was 55 months. The authors indicate that the .gov domain was more reliable than other top-level domains with pages on these sites less likely to disappear. Meanwhile, Dellavalle and colleagues (2003) examined the frequency, format and activity of web citations in over 1000 articles in three top level US scientific and medical journals finding that 13% of all cited Web references were inactive after 27 months. The authors concluded that URL citations occurred frequently, however, they were often not accessible even within months of first publication. Moreover, the authors indicate that the problem of impermanent Internet references requires immediate attention by the editorial and publishing community.

Davis and Cohan (2001) tracked bibliographic Web citations in research papers by a group of undergraduate students and reported that some 45% of the URLs cited in the bibliography had disappeared after 12 months, whilst 82 % of URL bibliographic references failed to directly lead to the cited document after 3 years.

Lawrence and colleagues (2001) analysed references to Web-located resources cited in computer science based literature published between 1993 and 1999, concluding that citations to Web-located resources had increased dramatically in that period. However, the authors noted that many Web references were invalid which was an impediment to allowing them to investigate and verify the content of these sources. One of the early researchers into the validity of using Web-located resources was Germaine (2000), who investigated the persistence of 64 URL citations in 31 academic journal articles. Germaine reported the declining accessibility of these types of references, finding that after 3 years some 48% of pages could not be accessed in the articles investigated— thus, leading her to question the use of Web-located citations for scholarly literature. Furthermore, Germaine highlighted that for the scholarly community to maintain its integrity, electronically cited works need to be retrievable and reliable.

Rumsey (2002) explored the way that Web-located references were cited in law articles finding that such citations increased significantly between 1995 through to 2000. Furthermore, the average number of Web-located citations per articles increased from 1.9 (1995) to 10.45 in 2000. Rumsey also reported a significant increase in non-accessible Web references cited over the six-year study period and suggested that, in general, Web citations lack stability which invariably leads to loss

of resource access. Rumsey ironically noted that authors who cited Web sources in an attempt to facilitate a broader reader access to those resources may actually have achieved the opposite than if they had used traditional paper based references.

Spinellis (2003) investigated several thousand Web-located references that were cited in 2471 articles from reputable information systems based publications, calculating that each article had an average of 1.71 citations to Web-located resources. Moreover, many of these Web-located sources cited in the articles by authors had disappeared. Spinellis was able to conclude that the average *half-life* for URLs cited in an article was approximately four years— that is, some 50% of cited Web-resources in an article would not be available at the specified URL after a four year period. Zhang (2001) investigated the scholarly use of Internet-based electronic resources and reported on the citation features found in articles published in eight different types of Library and Information Science journals. One of Zhang's conclusions was that citations of web-located resources increased over an eight-year period, although this practice still trailed author citations of traditional text resources. Zhang also found that Web-resources had become an important component of a scholar's research, however, many scholars indicated that the lack of permanency was a feature that prevented them from more widely citing electronic resources.

## **METHODOLOGY**

AusWeb is an Australian based conference that was first held in 1995 and focuses on the World Wide Web as a new and important technology. The conference paper archive (1995 to 2003) was the source for articles to evaluate in this study.

The article publication model used by AusWeb involves the production of peer-reviewed articles not only in the traditional paper format, but also electronically on a CD-ROM and on the conference Web site. Moreover, authors are required to embed active hypertext links to URLs that may be referenced by a paper— allowing the reader to maximise the non-sequential navigation features of the Web and investigate cited resources. The AusWeb conference organisers provide freely accessible electronic versions of all papers, posters and articles on the conference Web site (<http://ausweb.scu.edu.au/aw04/archive/index.html>).

Articles (N=123) between 1995 and 2003 were selected from the Education and Training stream of the conference archive. This stream appears to be one of the few disciplines that has constantly appeared in conference proceedings and it was felt that some consistency in the type of references authors cited would add rigour to the evaluation process— more so than if cross-discipline paper selection was attempted where authors may have had different emphasis on citation sources. Where the stream was not specified (eg 2001) papers that had educational key works were used to select appropriate papers to test.

### *Testing of articles*

The citations in an article were defined as the references that appeared as a list at the end of the article under the Bibliographic and/or Hypertext Reference section. Expanded bibliographies, endnotes, footnotes, email links and annotations were not considered as citations and were not tested, or counted in the data collected. In some papers web-located citations were listed twice in the bibliographic and hypertext references sections— when this occurred they were only counted and evaluated a single time.

Each article was initially examined to check that all Web-located citations were active— that is, they were marked up as a hypertext link. Any references with non-hypertext URL links were noted and tested manually. The World Wide Web Consortium's (W3C) Link Checker was used to evaluate links associated with a cited Web-located resource. Link Checker is a freely available online service (<http://validator.w3.org/checklink>) that tests a submitted Web page for broken or non-valid hypertext links and reports the types of HTTP messages encountered. Articles were submitted to Link Checker to check for broken links over a seventeen day period (between 12 and 29 October 2003). A second link checking tool— Doctor HTML Report service (<http://www.doctor-html.com/RxHTML>)— was used in tandem to confirm the validity of messages associated with broken links returned with Link Checker. Any discrepancies between results from the two checking tools were investigated manually. Some papers caused one or both of the two checking tools to time-out— these papers for practical reasons were not included in the study.

Link Checker reports allowed documented broken links to be randomly checked for non-active links several weeks later— testing for the possibility of transient network problems, such as a server being temporarily unavailable at the time of initial testing. No attempt was made to verify or evaluate information content of cited Web-located resources— it was assumed that if a link was active it led to the correct information resource cited by the author. Data derived from Link-checker allowed the identification of non-active Web-located references, the HTTP messages associated with these references. Also identified were the different type of top-level domains (for example: edu, gov, net) that broken links referred to. Articles were manually checked for the total number of citations used.

## **RESULTS**

A total of 123 conference papers for the 1995-2003 period were examined. Collectively, the papers contained a total number of 2168 references, with 48.1% (1043) of references citing a Web-located resource. It was assumed that all URLs found in cited articles were originally retrievable. Table 2 summarises citation results for the 123 articles.

The average number of Web references per paper ranged from a low of 3.5 in 1997 to a high of 12.3 in 2001. The average number of Web-located references per paper was 8.5 across all articles— a result that appears significantly higher than results found by previous researchers (Germaine 2000; Spinellis 2003). Furthermore, 1995 papers— the first year the conference was run— had the highest number of Web-located references (70%) when compared to the total number of references. A noteworthy finding is the high use of web-located resources in early conference



articles (1995- 96) compared to the later articles (2001-2003). It might have been expected that as the Web matured as an established information source, authors may have accessed and cited more Web-located resources in their articles. Indeed, the high number of Web-located citations found in this study's early papers is contrary to the findings of Zhang (2001), Rumsey (2002) and Spinellis (2003) who reported a relative increase of Web citations in articles published in later years. The greatest number of Web-located references cited by a single paper was 41, with numerous authors not citing any Web-located references. Authors first commenced documenting the date that a resource was accessed on the Web in 2000, however, this citation practice was not consistent amongst the year 2000 authors.

**Table 2: Summary of article citations (1995-2003)**

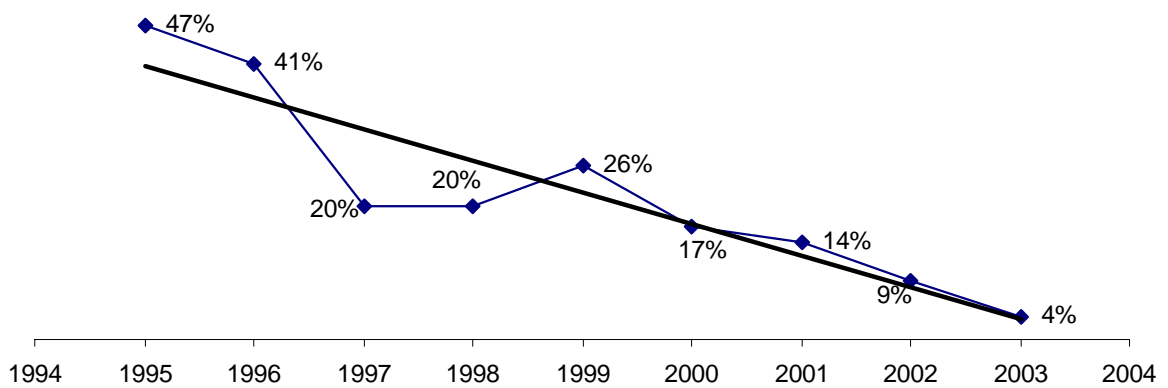
<b>Year</b>	<b>Articles (N)</b>	<b>Total references (N)</b>	<b>Web references (N)</b>	<b>Average citations per paper</b>	<b>Average Web citations per paper</b>	<b>Web references as percentage of all references</b>
1995	18	236	166	13.1	9.2	70%
1996	21	350	195	16.7	9.3	56%
1997	12	181	63	15.9	3.5	35%
1998	3	55	30	18.3	10.0	55%
1999	15	212	90	14.1	6.0	42%
2000	11	185	97	16.8	8.8	52%
2001	15	359	184	23.9	12.3	51%
2002	16	336	116	21.0	7.3	35%
2003	12	254	102	21.2	8.5	40%
Totals	123	2168	1043			
Average per paper		17.6	8.5			
As a percentage of all citations			48.1%			

### *Missing Web-located references cited in conference articles*

Web-located references were checked to see if they could be located at the specific URL cited in an article. Table 3 summarises the details of non-active Web-located references found in articles, whilst Figure 2 depicts the increased proportion and trend of missing Web references associated with older papers. Since 1995, there has been a progressive loss of Web-located references when compared to all the cited references. In that period 45.8% (478) of Web-located references could not be found at the documented Web address cited by authors. Indeed, the 2003 results indicate that as little as 4 months after papers were published some 4% of the Web-located references cited in papers were not active and had started to disappear. The results indicate that early-published papers have collectively a greater number of missing Web-located references— which is consistent with findings from many of the studies (see previous list) that have under taken similar investigations into missing Web-located resources. Moreover, in proportion to all citations (traditional and Web- located), some 22% of all references underpinning the theoretical foundations of articles could not be located at the author specified URL.

**Table 3: Summary of missing Web-located references (1995-2003)**

<b>Year</b>	<b>Total references (N)</b>	<b>Web references (N)</b>	<b>Active Web references (N)</b>	<b>Missing Web references (N)</b>	<b>Missing Web references as a percentage of all references</b>
1995	236	166	55	111	47%
1996	350	195	51	144	41%
1997	181	63	27	36	20%
1998	55	30	19	11	20%
1999	212	90	35	55	26%
2000	185	97	66	31	17%
2001	359	184	132	52	14%
2002	336	116	87	29	9%
2003	254	102	93	9	4%
Total				478	
Average missing URLs per paper (N)				3.9	
Loss of references (Web citations only)				45.8%	
Loss of references (all citations)				22.0%	



**Figure 2 Missing Web-located references as a proportion of all citations (1995-2003)**

*HTTP messages associated with missing Web-located references*

Four types of HTTP messages associated with missing Web references are summarised in Table 4. The 404 HTTP error message— Page not found— was the overwhelming message encountered and represented 62.5% of all HTTP messages. This suggests that Web-located resources cited in articles at the time of publishing had disappeared from the specified location as designated by the URL. Moreover, this finding tends to reinforce the transient nature of the Web as a publishing medium where information placed on a Web page is non-static, having a time dimension feature associated with it— a feature that allows information to be easily reformed, updated, altered or deleted. Although HTTP 403 was the least returned message— this finding suggests that authors have referenced Web-located resources that are within a restricted domain— not available to the general public. The reference to a restricted document may have been an inadvertent action by the author due to not understanding document access privileges they may have had within their University domain. Another possible explanation for this citation error is that previously

publicly available Web documents may have, with time, become restricted— only accessible through payment or subscription. Markwell and Brooks (2003) found this phenomenon in their investigations, reporting that various organisations had commenced charging for previously free educational material. Server side HTTP error codes 502 and 504 are indicative of server side problems and were almost equally split between a network timeout signal error being received and the host server not being found.

**Table 4 Summary of HTTP messages associated with missing Web-located references (1995-2003)**

<b>Year</b>	<b>Web references (N)</b>	<b>Missing Web references (N)</b>	<b>HTTP 404 (Not found)</b>	<b>HTTP 504 (Timeout)</b>	<b>HTTP 502 (Bad Gateway)</b>	<b>HTTP 403 (Restricted)</b>
1995	166	111	41 (8.6%)	27 (5.6%)	39 (8.2%)	4 (0.8%)
1996	195	144	110 (23.0%)	15 (3.1%)	18 (3.8%)	1 (0.2%)
1997	63	36	23 (4.8%)	9 (1.9%)	4 (0.8%)	0 (0.0%)
1998	30	11	5 (1.0%)	2 (0.4%)	3 (0.6%)	1 (0.2%)
1999	90	55	43 (9.0%)	4 (0.8%)	4 (0.8%)	4 (0.8%)
2000	97	31	17 (3.6%)	8 (1.7%)	4 (0.8%)	2 (0.4%)
2001	184	52	27 (5.6%)	16 (3.3%)	8 (1.7%)	1 (0.2%)
2002	116	29	21 (4.4%)	3 (0.6%)	4 (0.8%)	1 (0.2%)
2003	102	9	7 (1.5%)	0 (0.0%)	2 (0.4%)	0 (0.0%)
<b>Totals</b>	<b>1043</b>	<b>478</b>	<b>294 (61.5%)</b>	<b>84 (17.6%)</b>	<b>86 (18.0%)</b>	<b>14 (2.9%)</b>

### *Redirection of URLs associated with Web-located references*

A specific Web-located resource may be repositioned on an organisational Web site or it may also be moved from one Internet host server to another. Enquiries to a repositioned Web document can be achieved seamlessly through server redirection—a function that generally is imperceptible to a user. Table 5 summarises two HTTP messages generated when Web-located references were being investigated. HTTP 301 and 302 messages indicate that a redirection of a submitted URL has occurred from an originally specified location. The results indicate that for active Web-located references investigated some 17.1% (96) of the 563 active Web-located references were redirected from their originally cited URL. Hence, some 55% (478 missing + 96 redirected) of all cited URLs had either a missing web-located reference (45.8%) or had a reference that could be found but had moved from the original cited address (9.2%). Furthermore, 5.9% of all redirected URLs resulted in an error after redirection. A noteworthy observation was the high incidence of successful redirection rates in the years 2000 through to 2003, possibly suggesting that Web server/technology managers may be recognising the need to maintain the integrity of existing web pages and indirectly the affiliated links to those pages.

**Table 5 Summary of HTTP messages associated with Web page redirection**

HTTP Message/Year	1995	1996	1997	1998	1999	2000	2001	2002	2003	Total
<b>HTTP 301 &amp; 302 (Redirection) (N)</b>	11	9	4	8	6	12	20	16	10	96 (17.1%)
<b>HTTP 302 -&gt; 404/502/504 (Redirection -&gt; error) (N)</b>	2	10	6	0	4	2	3	1	0	28 (5.9%)

***Domains associated with missing Web-located resources***

Table 6 summarises the types of domains identified with missing Web-located references. The top-level domain having the greatest number of missing URLs was the education domain (edu). The edu domain is associated with educational organisations with Australian and International Universities being prominently represented in the URLs investigated. The high degree of missing education domain resources was also documented by Markwell and Brooks (2002) in their research.

Many missing resources in the years 1995 and 1996 referred to resource files that used the ftp and gopher file transfer method. Presumably as computer servers supporting these file transfer methods were phased-out, a relatively large number of ftp and gopher based references have disappeared. The number of missing ftp and gopher domains in recent times is negligible by virtue of HTTP having become the dominant file transfer method on the Web. This phenomenon of reducing numbers of non-specific top level domains was also reported by Koehler (2002) who suggests this



may be a result of migration of many Web sites to the more desirable and global domains such as .net, .com and .org.

**Table 6 Summary of domain types associated with missing Web-located references**

Year	Missing Web references (N)	Domains associated with missing Web references					
		.edu	.com	.org	.gov	.net	ftp, gopher & no domain
1995	111	71	7	3	1	0	29
1996	144	114	9	2	2	1	16
1997	36	21	4	1	1	3	6
1998	11	1	7	0	0	0	3
1999	55	35	7	3	4	0	6
2000	31	11	5	4	2	0	9
2001	52	30	9	4	3	3	3
2002	29	17	4	4	1	1	2
2003	9	6	2	1	0	0	0
Totals	478	306	54	22	14	8	74
Average as a percentage of all domains		64.0%	11.3%	4.4%	2.7%	1.7%	16.7%

## **DISCUSSION**

The study has provided evidence of the impermanent nature of URL citations in a set of conference papers. A comprehensive and direct comparison of results obtained in this study to previously published works (Germaine 2000; Lawrence, Coetzee et al. 2001; Zhang 2001; Rumsey 2002; Spinellis 2003) on scholarly use of Web-located resources was not appropriate due to different evaluation methodology used in these works. However, some comparison with Germaine's finding of 48% URLs loss after 3 years— compared to 45.8% after 9 years in this study— tends to suggest that the conference papers in this study have Web-located citations with a greater degree of stability when compared to Germaine's journal article results. Spinellis's finding of an average 1.71 Web citation per article when compared to 8.5 in this study, also suggests that authors of conference papers may have a greater inclination to cite Web-located resources than journal authors. Indeed, with an increasing incidence of research and investigation of Web page/site permanency there appears a need for some standard or consistent evaluation method amongst investigators to allow cross study and/or interdisciplinary comparison.

The loss of a large number of Web-located resources cited in scholarly articles has important implications for authors and the academic community. The disappearance of previously cited Web-located references challenges the reader's traditional assumption of reference availability and access. Furthermore, missing references tend to stymie the ability of a reader to further investigate interesting or significant aspects of an article. Indeed, the loss of cited sources tends to weaken an article's theoretical foundation— one of the fundamental objectives associated with scholarly literature. From Webster and Watson's (2002) perspective, the reader is denied the opportunity

to go backwards to the primary argumentative sources that might underpin new ideas, theory and model synthesis. Lost references as a result of missing URLs also tends to uncouple the incremental knowledge contributions of previous authors who may have *provided the shoulders* for an author to formulate an argument or premise, develop a model or accordingly place their work in theoretical context. Arguably, with the disappearance of cited Web resources the *shoulders climbed to be able to see further* by authors have actually diminished in number— clearly unintended and a consequence of the unstable publishing medium. Missing Web-located citations also impact on Metcalf's (2003) judicial approach to evaluation of citations by viewing them as providing supportive evidence for a scholarly work. For investigators wanting to take a judicial approach to the examination of articles, the issue of non-findable citations results in an inability to engage the simple task of examining the supportive evidence— the missing Web citations— on which a case (prosecution or defence) can be mounted or conclusion (verdict) reached. In essence, judgement and interpretation of past research is limited and potentially muddled— curtailment of future investigation and discourse may also be compromised. Hence, in a Web-enabled world, Metcalf's judicial evidence-based appraisal of an article's knowledge base should also have permanency of evidence added to the other constructs of expertise, authority and hearsay. Missing Web citations are also inconsistent with Baker's (2000) proposal of unambiguous identification of cited works to allow primary sources to be further examined. Hence, loss of cited URLs invariably will impact on the review of original material, methods and quotations, as well as semantic relevance to previous works.

A scholarly list of citations associated with an article should always be accessible, available and, in the electronic environment, retrievable. When citing Web-located resources authors have assumed that Web-resources will exhibit permanency— that is, they will always be available at the specifically cited URL. This assumption has been based on citation habits associated with publications that were created at a specific point in time, and generally in tangible print format. Indeed, considering the newness of the Web, author citation behaviour is still undoubtedly based on traditional citation teachings and experiences— which are invariably associated with hundreds of years of library practices where documents have been printed, purchased, indexed, shelved and then finally archived. The advent of the Web has circumvented this process to a certain extent allowing documents to be easily published, but, unlike traditional publications, Web documents can be pliantly altered, updated, relocated or deleted from their original Web posting. Consequently, the powerful publishing flexibility the Web medium offers also creates an environment where there is an almost certainty that cited URLs will disappear, and be lost with time. Thus, an important component supporting the longevity and integrity of scholarly works is the ability to access permanent citations. Clearly, scholarly literature citations must stand the integral test of time for the incremental advancement of knowledge to be upheld, as has been the case with traditional publications. The growing occurrence of missing URL citations as exemplified in this study is reflective of findings that should be viewed as impeding this incremental advance.

## **CONCLUSION**

The study investigated the permanency of Web-located references in 123 conference articles. As such, it is one of the few, but growing, number of studies that has

examined permanency of Web-located citations in academic articles. Of the 1043 Web-located resources cited, 45.8% of these resources were not found at the specified URL. The major reason for missing Web references was that the page was not found (HTTP 404)— a finding that not only reinforces the notion of Web pages being ephemeral and time reliant, but also reinforces that Web pages lack a sense of stability when it comes to scholarly citation.

It was argued that the inability to locate a resource at the specified URL weakens some of the theoretical foundations that the author has used to underpin their scholarly work. Indeed, the way that authors cite others — *standing on their shoulders allowing them to see a little further*— although valid at the time of publication, becomes invalid once the citations disappear. In the Web environment even though a large volume of resources are easily accessible, there appears to be a certain degree of predictability that cited URLs will disappear, be redirected, become restricted or altered from original published form— making them unsuitable as a scholarly citation source.

### ***Future Research***

Future research will incorporate two phases. The next stage of this research will be to examine the types of URLs that are ‘standing the test of time’. There has been little work done on this and the author is interested to see if certain sites, such as commercial online publishers of journals, provide a more secure ‘base’ for quoting a URL. The second phase will undertake studies that expand on these findings in other articles published in academic conference proceedings. These proceedings may be conferences that use the Web as the primary publishing venue, or those that publish

proceedings using the traditional non-electronic press.. A further avenue of research should examine possible solutions to addressing the loss of Web-located resources—some solutions have been proposed, however, presently lack critical mass in implementation. Indeed, it was noted that many of the newly proposed electronic referencing methods such as DOI (digital object identifiers) and PURL (Permanent Uniform Resource Locator) were not used in any of the later year papers— allowing the conclusion that URLs are the preferred citation style.

## **BIBLIOGRAPHY**

Baker M. J. (2000). “Writing a Literature Review.” *The Marketing Review*, 1: pp. 219-247.

Czarniawska-Joerges B. (1998). *A Narrative Approach to Organizational Studies*, Thousand Oaks, CA: Sage.

Davenport T. H. (1997). *Information Ecology: Mastering the Information and Knowledge Environment*, New York: Oxford University Press.

Davis P. M. and Cohen S. A. (2001). “The Effect of the Web on Undergraduate Citation Behavior 1996-1999.” *Journal of the American Society for Information Science and Technology*, 52 (4): pp. 309-314.

Dellavalle R. P., Hester E. J., Heilig L. F., Drake A. L., Kuntzman J. W., Graber M., *et al.* (2003). “Going, Going, Gone: Lost Internet References.” *Science*, 302 (11): pp. 787-788.

Germaine C. A. (2000). “URLs: Uniform Resource Locators or Unreliable Resource Locators?” *College & Research Libraries*, 61 (4): pp. 359-365.

- Grafton A. (1997). *The Footnote\*: A Curious History*, Cambridge, Massachusetts: Harvard University Press.
- Hart C. (1998). *Doing a Literature Review*, London: Sage Publication.
- Kahle B. (1997). "Preserving the Internet." *Scientific American*, 276 (3): pp. 72-74.
- Kellehear A. (1993). *The Unobtrusive Researcher: A Guide to Methods*, St Leonards: Allen & Unwin.
- Koehler W. (1999). "An Analysis of Web Page and Web Site Constancy and Permanence." *Journal of the American Society for Information Science*, 50 (2): pp. 162-180.
- Koehler W. (2002). "Web Page Change and Persistence: A Four-Year Longitudinal Study." *Journal of the American Society for Information Science and Technology (JASIST)*, 53 (2): pp. 62-171.
- Landlow G. P. (1996). Twenty Minutes into the Future: How we are Moving Beyond the Book. *The Future of the Book*. Berkley LA: University of California. pp. 296-313.
- Lawrence S., Coetzee F., Glover E., Pennock D. M., Flake G. and Nielsen F. (2001). "Persistence of Web References in Scientific Research." *IEEE Computer*, 34 (2): pp. 26-31.
- Markwell J. and Brooks D. W. (2002). "Broken Links: The Ephemeral Nature of Educational WWW Hyperlinks." *Journal of Science Education and Technology*, 11: pp. 105-108.

Markwell J. and Brooks D. W. (2003). "Link Rot Limits the Usefulness of Web-based Educational Material in Biochemistry and Molecular Biology." *Biochemistry and Molecular Biology Education*, 31: pp. 69-72.

Metcalfe M. (2003). "Author(ity): The Literature Review as Expert Witnesses." *Forum: Qualitative Social Research*, 4 (1)

Newby T. and Ertmer P. (1997). *Practical Research: Planning and Design*, Columbus, Ohio: Prentice Hall.

Nielsen J. (2000). *Designing Web Usability: The Practice of Simplicity*, New York: New Riders Publishing.

Nunberg G. (1996). Farwell to the Information Age. *The Future of the Book*. Berkley, LA: University of California. pp. 103-138.

O'Brien J. A. (2001). *Management Information Systems: Managing Information Technology in the Internetworked Enterprise*, 5th edition. Boston: McGraw-Hill.

Powell T. (2003). *HTML & XHTML: The Complete Reference*, 4<sup>th</sup> edition. Berkeley: Osborne/McGrawHill.

Prostan M. (1958). Why Was Science Backward in the Middle Ages. *The History of Science: Origins and Results of the Scientific Revolution (A Symposium)*. Carlton: Melbourne University Press.

Rumsey M. (2002). "Runaway Train: Problems of Permanence, Accessibility, and Sustainability in the use of Web Sources in Law Review Citations." *Law Library Journal*, 94: pp. 27-39.



Spinellis D. (2003). "The Decay and Failures of Web References." *ACM*, 46 (1): pp. 71-77.

Ticehurst G. W. and Veal A. J. (2000). *Business Research Methods*, Frenchs Forest: Longman.

Turban E., Mclean E., Wetherbe J., Bolloju N. and Davison R. (2002). *Information Technology Management: Transforming Business in the Digital Economy*, New York: John Wiley.

Veyne P. (1988). *Did the Greeks Believe in Their Myths*, Chicago: University of Chicargo.

Webster J. and Watson R. T. (2002). "Analyzing the Past to Prepare the Future: Writing a Literature Review." *MIS Quarterly*, 26 (2): pp. xiii-xxiii.

Zerby C. (2002). *The Devils Advocate: A History of Footnotes\**, NY: Touchstone.

Zhang Y. (2001). "Scholarly Use of Internet-Based Electronic Resources." *Journal of American Society for Information Science and Technology*, 52 (8): pp. 628-654.