# *A Study of Missing Web-Cites in Scholarly Articles: Towards an Evaluation Framework*

# A Study of Missing Web-Cites in scholarly articles: Towards an Evaluation Framework

**Carmine Sellitto**

*School of Information Systems*

*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*

*Email: carmine.sellitto@vu.edu.au*

**Abstract.**

***This paper reports on a study that examined citation practice in a set of scholarly papers. After evaluating 2162 bibliographic references it was found that 48.1% (1041) of all citations used in the papers referred to a Web-located resource. A significant number of references to URLs were found to be missing (45.8%) and an evaluation of these Web-located citations allowed the average half-life (4.8 years) for these missing resources to be determined. The study also examined the composition of the top-level domains associated with resource***

***loss as well as the proportional use of Web-located resources in individual articles. The proportional use of Web-located resources in individual articles and their corresponding evaluation for disappearance has not been previously documented. The paper utilised the proportional Web citation aspect of articles in proposing a Web-resource contribution index and a Web-citation use-loss grid that may aid future authors, editors and, in particular, researchers in investigating this growing aspect of citation behaviour.***

Keywords: Citation; impermanence; Web-located resource; URL; evaluation framework

## 1. Introduction

The advent of the World Wide Web (Web) has seen many organisations adopt this medium as their preferred document publication sphere— a practice that has resulted in an increasing use of the Web as the preferred information dissemination channel. As more and more organisations publish their documents electronically, there is an increasing tendency to identify and use electronic works as a substitute for traditional texts. As a new publishing medium, the Web has also altered the modus operandi of the academic researcher by providing a new way of searching and disseminating information, as well as being a significant tool for communication and collaboration (Zhang 2001). Citation accuracy has underpinned scholarship across all academic disciplines where the customary practice of authors of scholarly works has been to make use of citation and referencing to other resources in an attempt to support and buttress their own ideas— a process which also positions their work in context with others (Webster and Watson 2002; Zerby 2002). Indeed, citation practice in the academic literature review signals an awareness of ethical publishing principles and behaviour that is commensurate with recognising the knowledge ownership of other writers. Moreover, the contributory recognition of previous authors allows an author to *stand on the shoulders of others* in furthering their own work.

The central role of the Web as an electronic information medium has resulted in the reporting of both practical and theoretical research in the information science literature (Egghe 2000; Cronin 2001; Zhang 2001; Thelwall 2002; Vaughan and Thelwall 2003). And more recently, an important and pertinent area of research that has started to receive investigation is the manner in which Web-located resources cited in scholarly articles have a relatively high likelihood of eventually disappearing (Casserly and Bird 2003; Koehler 2004). Zhang (2001) indicates that citation behaviour associated with this shift away from traditional print media to electronic publishing has not been appropriately investigated. This study reports on the evaluation of a set of conference articles to determine the electronic citation base of those articles and the subsequent loss of citations that pointed to Web-based resources. Moreover, the paper extends findings and proposes an evaluation framework that can be

used by future investigators of this phenomenon in allowing them to examine the electronic citation behaviour of authors.

## 2. Background

The ability to access information via the Web has allowed authors to substitute some of the traditional paper-based resources such as books, journals, reports and notes with an electronic equivalent. Numerous style manuals elaborate on how Web-located references should be appropriately and technically cited with specific features that include the Uniform Resource Locator (URL) and date of resource access (APA 2001; Oxford 2003). With the vast quantity of easily accessible documentation available on the Web, authors often cite URLs as part of the attribution process when it comes to acknowledging supporting material in their publications (Rumsey 2002; Spinellis 2003). There is, however, an assumption that the resource, like traditional printed publications, has a permanency associated with its creation. Web permanency in the context of this study refers to a Web-located resource being easily located at the particular URL address cited in an article. Rumsey (2002) suggests that with the increase in scholarly citations to Web-located resources, there needs to also be caution due to the way that URL references tend to disappear. This concern with disappearing Web-resources is manifested through the growing incidence of missing or broken Web links— an issue first flagged by Kahle (1997) who suggested that the average lifetime of a URL might be just 44 days. Since 1997, various authors have reported the problem of disappearing URLs and alluded to the impact of this phenomenon on scholarly works (Koehler 1999; Germaine 2000; Koehler *et al.* 2000; Lawrence *et al.* 2001; Zhang 2001; Koehler 2002; Markwell and Brooks 2002; Rumsey 2002; Casserly and Bird 2003; Markwell and Brooks 2003; Spinellis 2003; Koehler 2004).

### *2.1.* ***The disappearance of URLs***

Koehler (1999) reported on aspects of Web page permanency indicating that both Web sites and pages underwent significant changes over a short period of time. Koehler examined 350 URLs over a 3 year period and concluded that 17.7% of web sites and 31.8% of web pages failed to respond when searched for after 12 months. Significantly, the study reported that all but one of the Web pages investigated had changed in either content or structure after three years and that some URLs at times disappeared, only to reappear some time late. This Web page disappearing-reappearing feature was referred to as a *comatosed* page by Koehler— a term that described a page temporarily disappearing, only to reappear on a later search— sometimes with different format and content. Koehler (2002) further examined the attrition and modification rates of Web sites/pages confirming many of his 1999 findings. Moreover, Koehler indicated 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. The 2002 study also reported *phantom* Web pages— a term that described a growing number of host-server generated error 404 pages that

mimicked content pages. These phantom pages fooled the software Koehler used to evaluate page presence into believing they were valid with specific content. Recent work by Koehler (2004) compared Web page attrition rates from previously published literature leading him to suggest that missing Web documents that were discipline specific showed variable stability that was reflected in different *half-lives*. Arguably, Koehler's recent work may highlight a disparity in research investigative methods that have examined this phenomenon. According to Markwell and Brooks (2002; 2003), Web-located science resources tend to lacked stability and permanence when compared to the science textbook. In their longitudinal study of 515 Web-located science resources they determined the URL half-life to be approximately 55 months. Furthermore, Markwell and Brooks (2002) found that some 14% of all URLs had ceased to function or changed their content in the first 14 months of the study. The authors indicate that the .gov domain was more reliable than other top-level domains with pages on these sites less likely to disappear. Conversely, the .com domain was the top-level that exhibited the highest degree of instability with almost 50% of these types of URLs not being accessible after 24 months

### *2.2. Loss of Web-located citations in scholarly articles*

One of the early investigations on the validity of citing Web-located resources in scholarly articles was undertaken by Germaine (2000). Germaine reported the persistence of 64 URL citations in 31 academic journal articles— persistence being interchangeable with permanency as a feature of URL citation and being inherently associated with resource accessibility after a period of time. Germaine found a declining accessibility of these types of references, reporting that nearly 50% of referenced Web pages could not be accessed after 3 years— leading her to question the appropriateness of using Web-located citations in scholarly articles.

The use of Web-located references in scholarly law articles was examined by Rumsey (2002) who found that such citations increased significantly between 1995 through to 2000. Rumsey reported that the average number of Web-located citations per article increased from 1.9 to 10.45 over a 6 year period, with some articles having a relatively high reliance on Web-located citations when compared to the total number of article citations. Rumsey concluded that, as a result of the significant number of non-accessible Web references cited in law articles, this type of citation lacked stability and authors in the legal disciple that utilise the Web as a primary information source tend to lose this support within a matter of years. Rumsey noted that authors who cited Web sources, instead of traditional based-paper references in an attempt to facilitate broader reader access to those resources may have actually achieved the opposite. Lawrence *et a.* (2001), evaluated citations to Web-located resources in 270,977 computer science articles concluding that citations to Web-located resources had increased dramatically over a 7 year period. The authors indicated that by 1999, each article had on average 1.6 citations that pointed to a Web-located resource however, they also noted a progressive decay of these citations from 1994.

Using a specifically developed software tool, Spinellis (2003) investigated several thousand Web-located references cited in the information systems based literature. Spinellis found an average of 1.71 Web citations per article, however many of these Web-located citations had disappeared and that the average *half-life* for URLs in the articles examined was approximately 4 years. Casserly and Bird (2003) investigated 500 citations to Web-located resources as a representative sample of over 35,000 citations in articles published in the library and information journals. They concluded that some 10% of all citations used in articles pointed to Web URLs (average of 2.5 per article), whilst in the investigated sample, 43.6% of all Web-citations were not found at author stated URLs. The study also found that the .edu and .org were the top-level domains exhibiting the greater degree of permanency, whilst URL directory depth and Web content were also important when evaluating for permanency. The study confirmed some of Lawrence et al's (2001) work, concluding that many references not found at the originally cited URL could be subsequently located by either using a Google or the Internet Archive to search for the appropriate citation, lifting the overall availability of electronically cited sources to 89.2%. One caveat ,however, was that the researchers faced many challenges in trying to determine whether content at a given URL *matched* that viewed by the author of the article in which it was cited. Arguably, when it comes to scholarship, the cited resources should not be best *matched* with what can be found somewhere else on the Google-searched Web or in an archive. The partnership that an author has with future readers is the integral conveyance of scholarship that is linked to the use of accurate citations— not best matched with another similar Web resource.

### 2.3. *The research questions*

The research is exploratory in nature and involved the investigation of scholarly Education and Training conference articles, published between 1995 and 2003 for features relating to missing Web-located citations. The study research questions were viewed as being—

What is the general citation profile of investigated articles with respect to Web-located citations?

With respect to cited Web-located citations:

- What proportion of Web-located citations have authors used in their articles?
- What top-level domains constitute the source of Web-located citations?
- What is the rate of citation loss (or impermanence) encountered in papers as an average across all publications and by individual papers?
- What top-level domains constitute the source of missing citations?

Further to answering the above research questions an evaluation framework based on the reported findings of the study is proposed.

## 3. Methodology

### *3.1. Selection of Conference*

The Australian-based AusWeb series of conferences between 1995 and 2003 was the source of scholarly articles for this study. AusWeb was the first regional conference to be endorsed by the International World Wide Web Conference Committee (IW3C2) that, since 1994 has been the central body organising academic conferences on Web technology. IW3C2 conferences engage the academic community by promoting Web-based research— both theoretical and practical— and addressing discussion and debate about all aspects of the Web and appropriate impacts (IW3C2 2000) .

### *3.2. Selection of articles*

Articles (N=123) were selected from the Education and Training component of the AusWeb conference archive (1995-2003) located at http://ausweb.scu.edu.au/aw04/archive/index.html. The Education and Training conference stream is one of the few streams constantly reported in AusWeb conference proceedings. By selecting a regularly appearing stream it was felt that citation behaviour and consistency amongst a group of discipline specific authors would add an element of 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), articles with author determined educational key words were used to select appropriate papers for testing. This selection process sacrifices probabilistic sampling, however, tends to embrace Koehler's (2004) recent observation that Web citation stability appears to be related to scholarly discipline.

### 3.3. *Testing of articles*

The AusWeb publication model involves the production of peer-reviewed articles in three formats— the traditional paper proceedings, electronically on a CD-ROM and interactively 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 navigation features of the Web to facilitate a broader reader access to cited resources.

The citations in an article were defined as the references that appeared as a list at the end of the article under the Bibliography and/or Hypertext Reference section. Any 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 bibliography and hypertext references sections— when this occurred they were only counted and evaluated a single time.

The World Wide Web Consortium's (W3C) online Link Checker (http://validator.w3.org/checklink) was used to evaluate links associated with a cited Web-located resource. The Link Checker service tests a submitted Web page for broken or non-valid hypertext links and reports the broken links encountered. Articles were submitted to Link Checker over a 17 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 broken links returned with Link Checker. Discrepancies between results from the two checking tools were investigated manually.

A random selection of identified broken links were rechecked several weeks later— testing for the possibility of transient network problems that may have occurred at the time of initial testing— no significant differences were found. It was assumed that if a link was active it led to the correct information resource cited by the author— no effort was made to check the integrity of Web-page content with respect to the context in which it was used in the article. Nor did the study check for Koehler's (2002) phantom web pages that would have masqueraded as content pages to Link Checker. The evaluation of broken links pointing to cited Web resources allowed the different type of top-level domains and the proportion of broken links to be determined.

## 4. Results and discussion

### *4.1. What is the general citation profile of investigated articles with respect to Web-located citations?*

123 conference articles for the years 1995-2003 were examined. The papers contained a total of 2168 references (average 17.6), 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 1 summarises findings for the 123 articles investigated.

*Table 1. Summary of article citations (1995-2003)*

| Year | Articles evaluated (N) | Total references (N) | Total Web-located references (N) | Average citations per paper | Average Web-located references per paper |
|---|---|---|---|---|---|
| 1995 | 18 | 236 | 166 | 13.1 | 9.2 |
| 1996 | 21 | 350 | 195 | 16.7 | 9.3 |
| 1997 | 12 | 181 | 63 | 15.9 | 3.5 |
| 1998 | 3 | 55 | 30 | 18.3 | 10.0 |
| 1999 | 15 | 212 | 90 | 14.1 | 6.0 |
| 2000 | 11 | 185 | 97 | 16.8 | 8.8 |
| 2001 | 15 | 359 | 184 | 23.9 | 12.3 |
| 2002 | 16 | 336 | 116 | 21.0 | 7.3 |
| 2003 | 12 | 254 | 102 | 21.2 | 8.5 |

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 who examined loss of these types of citations in journal articles (Casserly and Bird 2003; Spinellis 2003). This finding may indicate that when preparing for conference articles— which generally have a relative short preparation and submission period— authors may utilise Web content that is easily accessible and timely as an important source for grounding such articles. Arguably, the more prolonged and stringent journal submission process may allow an author to use more traditional citation sources which may be reflected in the lower average values that Spinellis (1.71) and Casserly and Bird (2.5) reported. Moreover, there was a relatively high reliance on electronic citations (48.1%) in the set of evaluated papers. This tends to reinforce the notion that Web may provide an important medium for conference authors to source information.

It was noted that authors first commenced documenting the date that a Web-located resource was accessed in 2000, however, this citation practice was not consistent amongst the year 2000 authors. Indeed, this inconsistency also extended to subsequent years which was a finding also reported by researchers Casserly and Bird (2003) in

their investigation of journal articles. Arguably, this may be symptomatic of poorly enforced electronic citation guidelines by article reviewers and editors.

*4.2. What proportion of Web-located citations have authors used in their articles?*

Table 2 provides a frequency distribution of the Web-located citations as a proportion of all citations. The results indicate that 10 authors chose not to use any citations to Web-located resources, whilst 16 articles draw entirely from the Web to substantiate theoretical argument. Table 3 indicates frequency groupings of the number of citations to Web-located resources in evaluated papers. A relatively high proportion (45.5%) of all articles referenced up to 5 citations as part of the attribution process, whilst 5.7% of articles sourced more than 20 references from the Web. The highest number of Web citations in an individual article was 41, with some authors not citing any Web-located resources.

*Table 2. Contribution of Web-located citations as a proportion of all citations*

| Web-located citations as a proportion (%) of each paper | 0 | 1-10 | 11-20 | 21-30 | 31-40 | 41-50 | 51-60 | 61-70 | 71-80 | 81-90 | 91-99 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Papers (N) | 10 | 10 | 10 | 8 | 9 | 21 | 5 | 10 | 13 | 7 | 4 | 16 |

*Table 3. Frequency of citations to Web-located resources in papers.*

| Frequency (N) of Web-located citations references in papers | 0-5 | 6-10 | 11-15 | 16-21 | +21 |
|---|---|---|---|---|---|
| Papers (N) | 56 (45.5%) | 33 (26.8%) | 14 (11.4%) | 13 (10.5%) | 7 (5.7%) |

Clearly, individual articles showed variability in the proportion of Web-located citations used by the author. A data spectrum (Figure 1) is used to visualise the overall contribution of Web-located references as a proportion of each individual paper. The shaded area in Figure 1 collectively represents all Web-located references cited in papers. A relatively high proportion of Web-located references in a paper are depicted by a spike or peak in the spectrum that may approach 100%. Notably in numerous papers it can be seen that the Web has been a source for 100% of all citations, whilst in others, there is an absence of Web references (non-spiking).

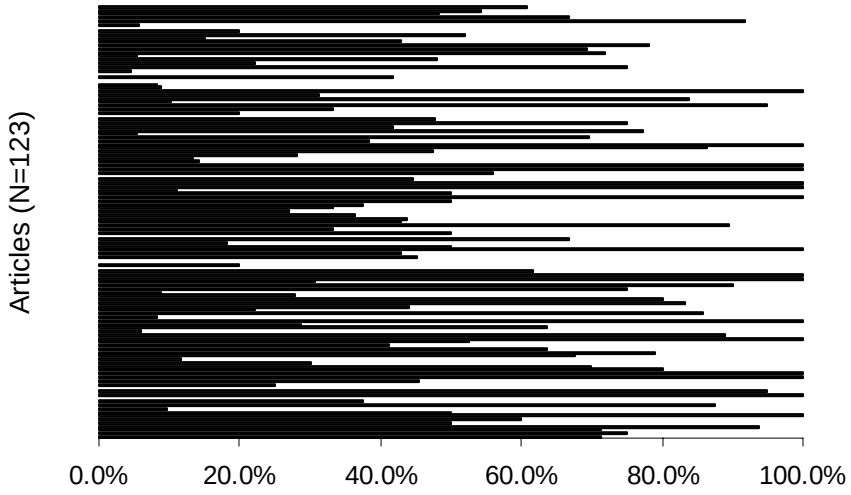


*Fig. 1 Data spectrum representing Web citations in each paper*

*4.3. What top-level domains constitute the source of Web-located citations?*

The top-level domain names of the Web-located citation published in articles between 1995-2003 is shown in Table 4.

*Table 4 Composition of the Web-located citation in papers (1995-2003)*

| Year | Domains associated with Web-located citations* | | | | | | Annual Totals (N) |
|---|---|---|---|---|---|---|---|
| | .edu | .com | .org | .gov | .net | ftp, gopher & IP | |
| 1995 | 89 (53.6%) | 13 (7.8%) | 5 (3.0%) | 8 (4.8%) | 4 (2.4%) | 47 (28.3%) | 166 |
| 1996 | 147 (75.4%) | 18 (9.2%) | 4 (2.1%) | 2 (1.0%) | 1 (0.5%) | 23 (11.8%) | 195 |
| 1997 | 37 (58.7%) | 10 (15.9%) | 2 (3.2%) | 1 (1.6%) | 3 (4.8%) | 10 (15.9%) | 63 |
| 1998 | 10 (33.3%) | 15 (50.0%) | 1 (3.3%) | 0 (0.0%) | 0 (0.0%) | 4 (13.3%) | 30 |
| 1999 | 58 (64.4%) | 12 (13.3%) | 3 (3.3%) | 4 (4.4%) | 0 (0.0%) | 13 (14.4%) | 90 |
| 2000 | 29 (29.9%) | 23 (23.7%) | 19 (19.6%) | 7 (7.2%) | 0 (0.0%) | 19 (19.6%) | 97 |
| 2001 | 77 (41.8%) | 42 (22.8%) | 32 (17.4%) | 7 (3.8%) | 10 (5.4%) | 16 (8.7%) | 184 |
| 2002 | 51 (44.0%) | 20 (17.2%) | 28 (24.1%) | 6 (5.2%) | 3 (2.6%) | 8 (6.9%) | 116 |
| 2003 | 58 (56.9%) | 13 (12.7%) | 18 (17.6%) | 4 (3.9%) | 2 (2.0%) | 7 (6.9%) | 102 |
| Domain Totals (N) | 556 | 306 | 166 | 54 | 112 | 39 | 1043 |

* Number of domains and relative percentage

The evaluation of top-level domains cited by authors indicates that the .edu was the leading domain (56.9%) referenced, with the .org being a distant second (17.6%). Indeed, the these two referenced domains were also found to be more permanent by previous work undertaken by Casserly and Bird (2003). The use of the .edu domain as the predominate source of online resources may reflect a perception by authors that such domains, by virtue of having similar characteristics to their own University environment, provided information content well suited for citation. Arguably, .edu domain associated with the knowledge-intense University environment will tend to publish documents that are considered accurate, authoritative and exhibit consistency that is generally associated reputable educational entities.

From 1999 the relative citation of sources located on .org servers increased significantly. The increased use of the .org domain may be due to an increased number of public and private associations now being online and making information more accessible to their constituencies. Indeed, authors may have discovered a perceived stability in resources located on the .org domain, and their citation behaviour reflecting a confidence in using information appearing on these domains. Another notable observation is the decreased citation of ftp, gopher and Internet Protocol (IP) domains, in that they have declined significantly to 6.9% in the last two years of the evaluation period. Indeed, the last citation of a resource found with the gopher and ftp file-transfer method occurred in 1996 and may reflect the migration of files and databases from legacy information systems to modern and standardised systems that invariably are based on the HTTP protocol. Koehler (2002) also observed this phenomenon of reducing numbers of non-specific top level domains and suggested this may be a result of migration of many Web sites to the more desirable and global domains such as .net, .com and .org.

During the years 2000 and 2001 there was an increase in the citation of commercial pages (.com) which corresponds to the peak in the dot com boom. The general hype surrounding Internet adoption in this period may have contributed to the increased availability of commercially located information resources, whereby many authors perceived this new avenue of information as accessible, reliable and permanent.

*4.4. What is the rate of citation loss (or impermanence) encountered in papers as an average across all publications and by individual papers?*

Citations that pointed to Web-located resources were evaluated to see if they could be located at the specified URL. Table 5 summarises the details of missing Web-located citations found in articles and Figure 2 depicts rate of decay of Web references.

*Table 5: Summary of missing Web-located references (1995-2003)*

| Year | Total citations (N) | Web citations (N) | Missing Web citations (N) | Missing Web citations as a percentage of Web citations | Missing Web citations as a percentage of ALL citations | Annual half –life of Web citations (years) |
|---|---|---|---|---|---|---|
| 1995 | 236 | 166 | 111 | 66.9% | 47% | 5.9 |
| 1996 | 350 | 195 | 144 | 73.8% | 41% | 4.7 |
| 1997 | 181 | 63 | 36 | 57.1% | 20% | 5.3 |
| 1998 | 55 | 30 | 11 | 36.7% | 20% | 6.8 |
| 1999 | 212 | 90 | 55 | 61.1% | 26% | 3.3 |
| 2000 | 185 | 97 | 31 | 31.9% | 17% | 4.7 |
| 2001 | 359 | 184 | 52 | 28.3% | 14% | 3.5 |
| 2002 | 336 | 116 | 29 | 25.0% | 9% | ** |
| 2003 | 254 | 102 | 9 | 8.8% | 4% | ** |
| | | | | | | |
| Average missing Web-located citations per paper (N) | | | | | | 3.9 |
| Loss of references as a proportion of Web citations | | | | | | 45.8% |
| Loss of references as a proportion of all citations | | | | | | 22.0% |
| Average half life of Web-located citations in articles (years) | | | | | | 4.8 |

** Half-life values for 2002 and 2003 were not calculated due to insufficient time having elapsed to allow for meaningful evaluation.

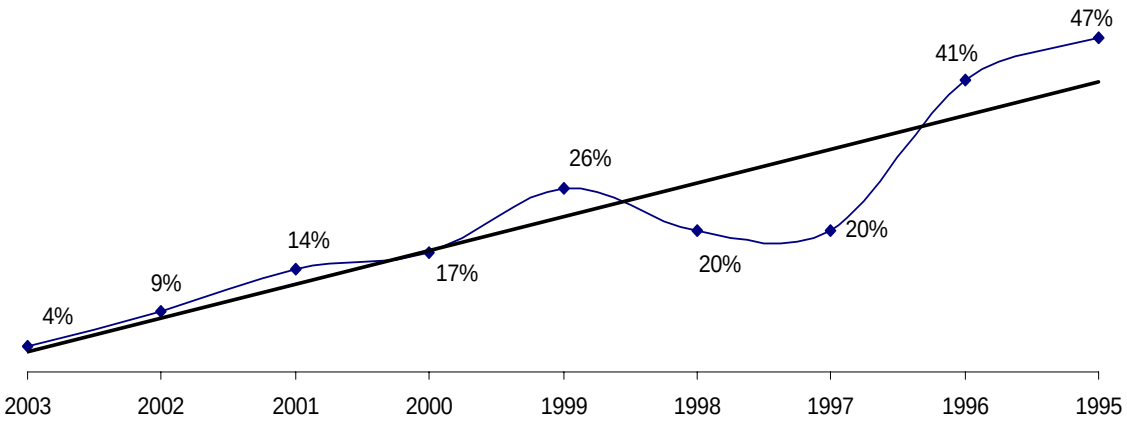


*Fig. 2. Missing Web references as a proportion of all citations (1995-2003)*

For the 1995 to 2003 evaluation period, 45.8% (478) of Web-located references could not be found at the documented URL cited by authors. The missing Web citations in proportion to all citations (traditional and Web-located) accounted for a total loss of 22.0% of all references— indicating that nearly one quarter of supportive works was unable to be located. Notably, as little as several months after papers were published some 4% of the Web-located references were found to be missing. The half-life for Web citations were calculated— this time value represents the years passed after which half the number of Web citations would be found to be missing (it was assumed a linear relationship). The half-life values allowed a year-by-year comparison as well as an average half-life (4.8 years) for the cited Web-located citations in articles to be determined. A comparison of the average half-life for Web-located resources to previously published works by Koehler (half-life of 2 years), Spinellis (4 years) Markwell and Brooks (55months) and Germaine (circa 3 years) suggests that the articles in this study have electronic citations sources that have lasted longer and could be viewed as having a greater degree of stability.

### *4.5. Missing Web-located resources as a proportion of individual articles*

The data spectrum (Figure 1) indicated a proportional use of Web-located citations in each article and as such does not give an indication of the loss of an article's citation base. In some papers there may have been a high use (HU) of Web citations and a corresponding high loss (HL), or low loss (LL) of these references. Conversely, a paper may have a low use (LU) of web citations with either a low loss (LL) or high loss (HL). A plot of the proportional contribution of Web-located citations against the rate of citations loss for each paper (N=123) can be depicted as a scatter plot (Figure 3)— where each point on the scatter reflects a paper's relationship between the number of Web citations contributing to the theoretical foundations of an article verses the citations now missing.

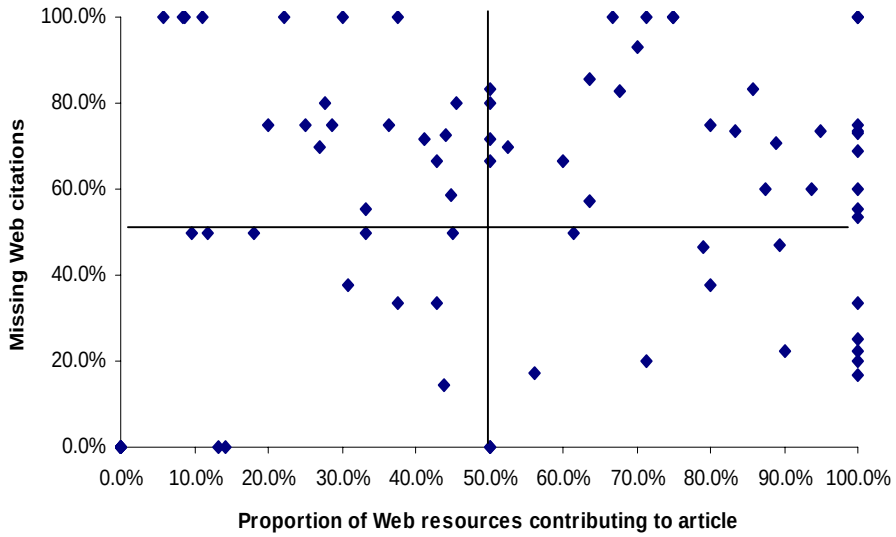


*Figure 3* *Use-loss representation grid for article Web citations*

The use-loss plot can be viewed as a visualisation of article stability with respect to Web citations. For example, papers located in the top right hand quadrant of the plot correspond to papers that have a relatively high use of Web-location citations and of which a high corresponding number are missing (HUHL)— a combination that would significantly impact on the theoretical underpinnings of an article. Papers in the lower right hand quadrant that have a high use with a corresponding low loss of Web-located citations (HULL) have stood the test of time with respect to electronic resource stability. Indeed, the HULL group of papers may be an investigative source providing exemplary citation practice for authors. This would allow authors to emulate permanency of electronic citations in future articles. An examination of articles (N=14) that fell into this quadrant published between 1995 and 2000 (papers for the years 2000 to 2003 were omitted due to insufficient time having elapsed for meaningful evaluation) had the following characteristics:

- Average number of citations used was 13.1 per article (study average 17.6)
- Average number of Web-located citations used was 10.2 per article (study average 8.5)
- Average number of Web-located citations as a proportion of all citations was 77.7% (study average 48.1%)— hence, High Use (HU)
- Average number of missing references as a proportion of Web-located was 27.9% (study average 45.8%)— hence, Low Loss (LL)

- Average number of missing references as a proportion of all citations for this group 21.7% (study average 22.0%)

This group of articles were also evaluated in for hypertext transfer protocol (http) messages associated with missing Web resources. The following errors were encountered:

- 5% of missing URLs were attributable to restricted host server access (HTTP message 403) compared to the overall study result of 2.9%
- 57.5% of missing URLs encountered a page not found message (HTTP message 404). Overall study result was 61.5%.
- 13% of URLs had been re-directed to a different host server location containing the cited resource (HTTP message 301 & 302) compared to the overall study result of 17.1%)
- 5% of missing URLs were attributable to incorrect host server names (HTTP message 502) compared to the overall study average of 18%.

The HULL articles have a lower number of average citations (13.1) compared to the study average (17.6), however, paradoxically these papers have used a relatively higher proportion (10.2) of electronic citations compared to the other papers (8.5). The higher degree of redirection to a new host and the decreased incidence of bad or incorrect host server names associated with missing Web-located resources is a notable finding. This suggests that the authors of HULL articles have cited documents located on host servers that appear to have been managed in a manner that has reduced the incidence of electronic document loss. Indeed, superior host server management practice appears to be one of the features that may result in improved permanency of Web-located electronic resources.

*4.6. What top-level domains constitute the source of missing citations?*

Table 6 summarises the top-level domains associated 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 and in this study Australian and International Universities were prominently represented in the URLs investigated. The high degree of missing education domain resources was also documented by Markwell and Brooks (2002) in their study. A notable finding is the proportionally low level of loss associated with the .org domain (19.6%) consolidating findings (Table 5) that the .org domain has not only been increasingly used by authors, but is also perceived as a reputable information source that appears to have relatively high permanency characteristics.

**A study of missing web-cites in scholarly articles: towards an evaluation framework**

*Table 6. Top level domains associated with missing Web-located resources*

| | Domains associated with missing Web references | | | | | |
|---|---|---|---|---|---|---|
| | .edu | .com | .org | .gov | .net | ftp, gopher & IP number |
| Missing Web references (N) | 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% |
| As a percentage of domain group | 55.0% | 32.5% | 19.6% | 35.9% | 34.8% | 50.3% |

## 5. Towards a citation framework for Web resources

This section develops and articulates an evaluation framework that is based on aspects of the findings reported. The evidence from this research, and the findings of others, (Kahle 1997; Koehler 1999; Germaine 2000; Davis and Cohen 2001; Lawrence, Coetzee et al. 2001; Koehler 2002; Markwell and Brooks 2002; Rumsey 2002; Casserly and Bird 2003; Markwell and Brooks 2003; Spinellis 2003) points to a continuing use of Web citations by authors in scholarly articles with a corresponding loss of some of those references. Thus, the intent is to provide a means for:

- Classifying scholarly articles with respect to author citation of Web-located resources to reflect the proportional electronic Web citation base underpinning a paper (web-resource contribution index), and
- Allowing articles to be grouped into categories based on the proportional use of Web-located citation and subsequent loss of these citations (use-loss citation grid), which can be used for analysing citation behaviour.

### *5.1. A Web-resource contribution index for scholarly papers*

The data spectrum (Figure 1) depicted an articles proportional contribution of Web citations. Evaluating papers for citations that point to Web-located resources allows a Web contribution index ($I_w$) to be determined. The $I_w$ is a value calculated by dividing the number of cited Web-located references ($R_w$) in a paper by the total number of references cited ($R_t$). Hence,

Web-resource Contribution Index ($I_w$) = Web-located references ($R_w$)/ Total References Cited ($R_t$).

The $I_w$ of a paper can have a value between zero and one— papers with high citations to Web-located resources will score an Index value closer to 1.0 than those that have none or minimal Web citations. For example, an article containing 20 Web citations in a total of 25 would have an $I_w$ of 0.8, at the high end of the index; a paper where the author used no URLs would be devoid of electronic citations and have an $I_w$ of zero. Moreover, the higher the index value the greater the contribution of Web-located resources to the theoretical underpinnings of a research paper. Considering the relatively high rates of disappearing Web-located resources cited in scholarly papers— as identified in the research literature as well as from this study's results— it can be assumed that a paper that has a high index ($I_w$) value at the time of publication will have a relatively higher likelihood that its theoretical underpinnings will disappear with time. This index becomes a useful tool for editors and reviewers of scholarly articles allowing them to gauge the contribution of Web-located citations and subsequently assess the potential decay of the articles' support base. Zhang (2001) indicates many editors that adjudicate scholarly publications do not have specific and clear guidelines allowing them to evaluate articles that have referred to electronic sources. The proposed $I_w$ could be used by a publisher/editor as a predictor to gauge the stability of the theoretical foundations of an article. Moreover, a system such as this would allow authors to conform to editorial expectations when citing Web-located references.

### *5.2. The Web-citation use-loss citation grid*

The move to the Web environment as a preferred publishing medium for many organisations has provided scholarly authors with increased opportunities to source and use a higher number of Web-located citations. The high use (HU) of Web-located resources by authors reflects a citation shift from the traditional paper based paradigm to the electronic format. Arguably, the increasing use of electronic citation could be viewed as an indicator of both a personal and general acceptance of this type of citation amongst the academic community.

This study reported that various scholarly articles had an associated high use of cited Web-located resources with minimal or low loss (LL) of these citations. The low loss of citations reflects electronic resource permanency and may be an indicator of the stability of Web technology and/or site management practices. The use-loss scatter plot (Figure 3) allowed articles to be visually depicted on a grid system to represent the use (U) of Web-located citations in a paper compared to the proportional loss (L) of those citations over time. Conceptualising this scatter plot representation to a more general model it is possible to view the citation of Web-located citations within a use/loss grid as in Figure 4.

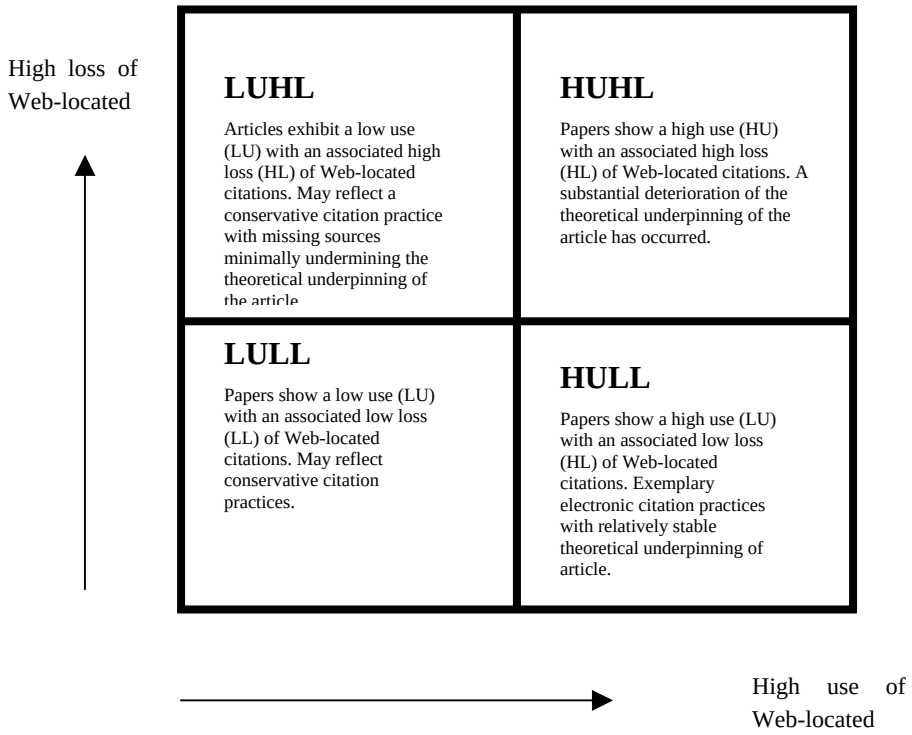


*Fig. 4. The use-loss Web citation grid for mapping articles*

As such, the grid allows a means of classifying or grouping articles to represent the decay of the theoretical base underpinning articles as a result of missing Web-located citations. The grid classification allows the identification of authors that have achieved high stability of electronic citations in their publications over time. Indeed, the authors of papers that classify in the HULL sector of the grid offer the researchers of this phenomenon the opportunity to further investigate Web citation behaviour in an endeavour to evaluate best practice. Studies of academic electronic citation behaviour in the Web enabled environment is nascent, hence this grid approach allows the suitable identification of exemplary articles in both the high use and limited loss of electronic citations

for further exploration. The grid may also allow researchers a vehicle to implement consistency in reporting findings that may allow easier cross-discipline comparison— a feature that has been missing from many of the published works on this topic.

## 6. Conclusion

This study examined the citation of Web-located resources in a set of scholarly Education and Training conference papers. The paper reported on the composite nature of electronic citations used in these articles and found that a relatively high degree of missing (45.8%) citations that pointed to URLs. The average *half-life* of missing electronic citation was determined to be 4.8 years and when compared to previously reported journal *half-lives*— the citations in this group of conference articles appear to have a greater degree of electronic stability. The study demonstrated proportional use of Web-located resources in individual articles and their corresponding disappearance by using a scatter plot visualisation.

The results of this study add further to the growing evidence that scholarly discourse is being impacted and eroded by the instability associated with citations to Web-located resources. Moreover, the author takes the view that enough evidence has emerged to suggest that the loss of Web-located citations in scholarly articles will continue and needs to be constantly and further investigated. Unlike other authors who have recommend a set of citation guidelines (Germaine 2000; Lawrence, Coetzee et al. 2001; Rumsey 2002; Casserly and Bird 2003), digital archives (Kahle 1997; Casserly and Bird 2003), or identifiers such as PURL or DOI (Spinellis 2003), this study proposed an evaluation framework that may assist information science researchers in the investigation of the stability and permanency of using Web-located citations in scholarly articles.

## 7. Reference

APA (2001). *Publication Manual of the American Psychological Association*, 5th Edition. APA.

Casserly M. F. and Bird J. E. (2003). Web Citation Availability: Analysis and Implications for Scholarship. *College and Research Libraries*, 64 (7): pp. 300-317.

Cronin B. (2001). Bibliometrics and Beyond. *Journal of Information Science*, 24 (1): 1-7.

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

Egghe L. (2000). New Informetric Aspects of the Internet: Some Reflections- Many Problems. *Journal of Information Science*, 26 (5): pp. 329-335.

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

IW3C2 (2000). *International World Wide Web Conference Committee (IW3C2) Home Page* Available at Internet Archive

http://web.archive.org/web/20000522233140/http://www.iw3c2.org/Welcome.html [Accessed 22 April 2004].

Kahle B. (1997). Preserving the Internet. *Scientific American*, 276 (3): 72-74.

Koehler W. (1999). An Analysis of Web Page and Web Site Constancy and Permanence. *Journal of the American Society for Information Science*, 50 (2): 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*, 53 (2): 162-171.

Koehler W. (2004). A Longitudinal Study of Web Pages Continued: A Consideration of Document Persistence. *Information Research*, 9 (2). Available at http://informationr.net/ir/9-2/paper174.html (accessed 4 May 2004).

Koehler W., Anderson A. D., Dowdy B. A., Fields D. E., Golden M., Hall D., *et al.* (2000). A Profile in Statistics of Journal Articles: Fifty Years of American Documentation and the Journal of the American Society for Information Science. *International Journal of Scientometrics, Informetrics and Bibliometrics (Cybermetrics)*, 4 (1,Paper3). Available at http://web.archive.org/web/20030623200359/http://www.cindoc.csic.es/cybermetrics/articles/v4i1p3.html (Accessed 16 May 2004).

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): 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: 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: 69-72.

Oxford (2003). *The Oxford Style Manual*, 2nd Edition. Oxford: Oxford 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: 27-39.

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

Thelwall M. (2002). A Comparison of Sources of Links for Academic Web Impact Factor Calculations. *Journal of Documentation*, 58: pp. 60-72.

Vaughan L. and Thelwall M. (2003). Scholarly Use of teh Web: What are the Key Inducers of Links to Journal Web Sites. *Journal of the American Society for Information Science and Technology*, 54 (1): pp. 29-38.

Webster J. and Watson R. T. (2002). Analyzing the Past to Prepare the Future: Writing a Literature Review. *MIS Quarterly*, 26 (2): 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): 628-654.