COLLEGE LIBRARIES
Vol. 40 No. 4 December 2025 pp. 77-86

Received on 13-05-2025 Reviewed on 19-09-2025 Paper Accepted on 25-12-2025

The effectiveness of the Wayback Machine in Reviving Vanished Web Citations: A Webometric Approach

Gurusiddesh Muganavar[a] and B. T. Sampath Kumar[b]

Abstract

This study examines the availability, decay, and recovery of web citations in College & Research Libraries (C & R L) journal articles from 2015-2024. A total of 21,384 citations from 497 research articles were analyzed, of which 9,287 were web citations. Web citation usage increased significantly over time, rising from 21.49% in 2015 to 66.01% in 2023 ($r = 0.931$, $p = .000$). Concurrently, the proportion of vanished web citations declined ($r = 0.901$, $p = .000$), indicating improved link persistence. HTTP 403 errors formed 56.57% of the observed HTTP errors among vanished web citations. The Internet Archive's Wayback Machine successfully recovered 95.30% of these vanished web citations, highlighting its effectiveness as a tool to retrieve lost web citations. This study emphasizes the growing reliance on web sources in academic work and underscores the critical role of digital preservation tools in maintaining access to scholarly materials over time.

Keywords: Vanished Web Citations, Wayback Machine, Webometric Study, Web Citations

1. Introduction

The Internet has become the primary platform for scholarly communities to access and disseminate information, driving rapid growth in online resources (Loan & Shah, 2020). As a result, researchers worldwide increasingly rely on web-based materials, leading to a surge in web citations within academic publications such as books, research articles, and conference papers (Isfandyari Moghaddam et al., 2010; Gul et al., 2014). However, URLs linking to online content are inherently unstable and subject to decay over time, raising concerns about the reliability and persistence of web citations (Koehler, 1998; Germain, 2000). Broken links, domain expirations, server errors, and website restructuring contribute to the disappearance of web-based references, undermining scholarly communication (Lawrence et al., 2001; Markwell & Brooks, 2003). To mitigate these challenges, web archiving initiatives have emerged to preserve online content for long-

[a] Research Scholar, Department of Studies and Research in Library and Information Science, Tumkur University, Tumkur, India, https://vidwan.inflibnet.ac.in/profile/649332, and https://orcid.org/0000-0001-5077-1450, Mobile: +91 8861648474, Email: gurusiddesh75@gmail.com

[b] Professor, Department of Studies and Research in Library and Information Science, Tumkur University, Tumkur, India, https://vidwan.inflibnet.ac.in/profile/89383 and https://orcid.org/0000-0001-6031-2100, Mobile No. : +91 9448320187, Email: sampathbt2001@gmail.com

Mugannavar, G. & Sampath Kumar, B. T. (2025, December). The effectiveness of the wayback machine in reviving vanished Web Citations: A webometric approach. *College Libraries, 40*(4), 77-86

term access. Among these, the Wayback Machine-established by the Internet Archive in 1996 and archiving over 928 billion web pages as of May 01, 2025-stands out for its extensive coverage. Recent studies demonstrate the effectiveness of the Wayback Machine and other archival tools in recovering lost citations, ensuring continued access to scholarly references (Sife & Lwoga, 2017; Sharma et al., 2022; Loan et al., 2024; Singh & Devi, 2024). These efforts ensure continued access to scholarly references, mitigating the risks associated with the impermanence of web-based resources.

This research investigates the impermanent nature of web citations and utilizes the Wayback Machine to retrieve vanished web citations from College & Research Libraries (C&RL) journal articles published between2015 and 2024.

2. Literature review

This section reviews foundational studies on web citation decay and recovery, highlighting their contributions and identifying gaps for the current research.

Web Citations Decay: Bansal (2021) examined link rot in DJLIT articles (2014-2018), finding that 448 of 1,922 web citations (23.31%) were inaccessible, with 64.74% returning HTTP 404 errors and an estimated half-life of 6.55 years. Niveditha et al. (2022) analyzed URL degradation in LIS and CMS journals, reporting that 23.1% of LIS and 15.68% of CMS citations had decayed, predominantly due to HTTP 404 responses.

Web Citations Recovery: Howell and Burtis (2023) assessed Wayback Machine recovery of inactive URLs in healthcare management journals (2016-2018), achieving a 76.2% restoration rate. Loan et al. (2024) applied the same tool to Library Hi Tech articles (2004-2008), successfully recovering 786 out of 1,083 dead citations (72.58%).

Existing studies measure web citation decay and recovery but ignore key factors such as URL path depth, domain type, and file format, and lack decade-long, cross-journal comparisons. This study remedies these omissions by analyzing C&RL articles (2015-2024) to correlate decay with path depth, error type, and recovery performance.

3. Objectives of the Present Study

The following objectives are formulated for the study

- To identify the proportion of web sources used as citations in the journal articles.
- To find out the percentage of vanished and recovered web citations by the Wayback Machine.
- To identify the top-level domains with vanished and recovered web citations.
- To find out the correlation between the path depth and the recovery of vanished web citations.

4. Hypotheses

These following hypotheses are developed from the stated objectives

- Over a period of time from 2015-2024, there has been increased use of web citations.
- The path depth and percentage of recovered web citations are negatively correlated.

5. Materials and Methods

The present study investigates the accessibility, decay, and retrieval of web citations in College & Research Libraries (C&RL) journal articles, chosen for its comprehensive publication history since 1939 and its open access policy. A total of 21384 citations appended to 497 research articles were selected. Only research articles were included in the study, while editorial notes, book reviews, and short communications were excluded. A total of 9287 web citations were extracted from 21384 citations. The W3C Link Checker (http://validator.w3.org/checklink) was used to verify the accessibility of the extracted web citations. Following a W3C Link Checker testing, web citations were classified as active and vanished.

5.1 Recovery of vanished web citations

In the Wayback Machine search field (https://web.archive.org/), we entered the URL of the vanished web citations and clicked the 'Take Me Back' button to access the vanished web citations. If the submitted web citations were recovered, they were considered recovered web citations. The SPSS 29.0 version of Windows is used for statistical analysis and hypothesis testing (Figure 1).

Figure 1. Home page of Wayback Machine

6. Data Analysis and Interpretation

The following tables summarize the analyzed data collected during the study. A discussion of these findings highlights key trends and insights regarding the use, decay, and recovery of web citations in the College & Research Libraries journal.

6.1. Distribution of Articles, Citations and Web Citations

Table 1 presents the distribution of articles, citations, and web citations published from 2015-2024. A total of 497 articles were published, accumulating 21,384 citations, including 12,097 print citations and 9,287 web citations.

Table 1: Distribution of articles, citations, and web citations

Year	No. of articles	No. of citations	No. of print citations	%	No. of web citations	%
2015	57	2020	1586	78.51	434	21.49
2016	43	1508	1062	70.42	446	29.58
2017	49	1850	1273	68.81	577	31.19
2018	48	2464	1712	69.48	752	30.52
2019	49	2286	1553	67.94	733	32.06
2020	53	2434	1140	46.84	1294	53.16
2021	49	2180	1272	58.35	908	41.65
2022	51	2362	948	40.14	1414	59.86
2023	45	1915	651	33.99	1264	66.01
2024	53	2365	900	38.05	1465	61.95
Total	497	21384	12097	56.57	9287	43.43

The study reveals that web citations ranged from 21.49% (2015) to 61.95% (2024). Pearson's correlation analysis was conducted to assess the relationship between the year and the rise in web citation percentages. The results indicate a positive correlation (r = 0.931, p = .000), confirming that the increase in web citations over time is statistically significant. Therefore, the null hypothesis is rejected, and the alternative hypothesis is accepted. The figure highlights a growing reliance on web-based citations in scholarly communication during the observed period.

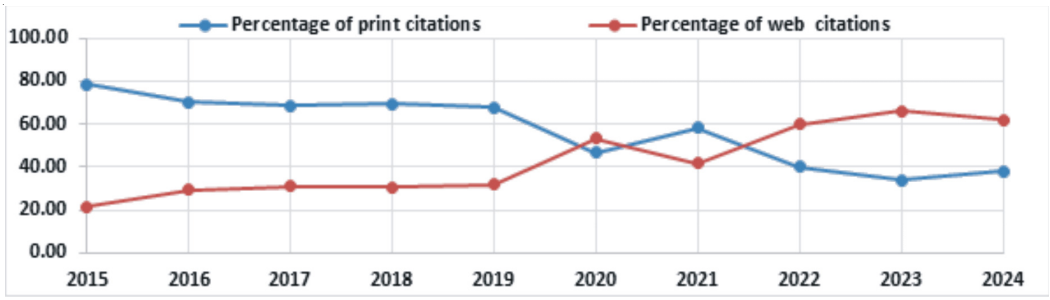

Figure 2. Percentage of print citations and web citations by year

6.2. Active and Vanished Web Citations Summary

Active and vanished web citations are presented in Table 2. Out of 9287 web citations, 6881 (74.09%) are active, and the remaining 2406 (25.91%) vanished. Active web citations ranged from 57.83% (2015) to 79.73% (2024).

Table 2: Active and vanished web citations

Year	No. of web citations	Active web citations	%	Vanished web citations	%
2015	434	251	57.83	183	42.17
2016	446	248	55.61	198	44.39
2017	577	354	61.35	223	38.65
2018	752	549	73.01	203	26.99
2019	733	555	75.72	178	24.28
2020	1294	880	68.01	414	31.99
2021	908	706	77.75	202	22.25
2022	1414	1144	80.91	270	19.09
2023	1264	1026	81.17	238	18.83
2024	1465	1168	79.73	297	20.27
Total	9287	6881	74.09	2406	25.91

Vanished web citations declined from 42.17% (2015) to 20.27% (2024). A positive correlation was found between web citation age and vanishing rate (r = 0.901, p = .000), which was statistically significant. These findings highlight the trend of older web citations being more likely to disappear, while newer ones remain accessible.

6.3. Summary of Vanished and Recovered Web Citations

Table 3 presents findings on web citation recovery through the Wayback Machine, showing restoration rates ranging from 88.55% (2024) to 99.28% (2020).

Table 3: Vanished and recovered web citations

Year	No. of web citations	Active web citations	Vanished web citations	Recovered web citations	%
2015	434	251	183	172	93.99
2016	446	248	198	181	91.41
2017	577	354	223	210	94.17
2018	752	549	203	195	96.06
2019	733	555	178	174	97.75
2020	1294	880	414	411	99.28
2021	908	706	202	198	98.02
2022	1414	1144	270	260	96.30
2023	1264	1026	238	229	96.22
2024	1465	1168	297	263	88.55
Total	9287	6881	2406	2293	95.30

Over 90% of vanished web citations from 2015-2024 were successfully recovered. Statistical analysis revealed an insignificant negative correlation between citation age and recovery rate ($r = -0.016$, $p = .966$), indicating age minimally affects retrievability. The figure shows the Wayback Machine's consistent effectiveness in recovering vanished web citations throughout the examined period, with most vanished web citations remaining accessible regardless of their age.

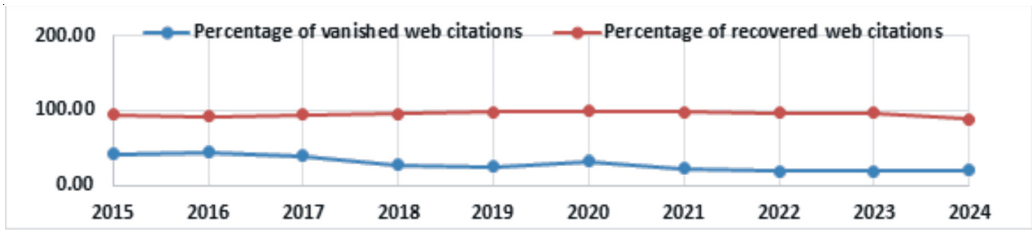

Figure 3. Percentage of decayed and recovered web citations

6.4.Summary of Vanished and Recovered Web Citations by Domain

Table 4 displays the domain-wise distribution of vanished and recovered web citations through Wayback Machine. The highest vanishing rates were for .int (50%), country code (34.67%), and .ac (33.93%).

Table 4: Vanished and recovered web citations by domain

Domain	No. of web citations	Active web citations	%	Vanished web citations	%	Recovered web citations	%
ac	112	74	66.07	38	33.93	38	100.00
country code	300	196	65.33	104	34.67	97	93.27
com	848	628	74.06	220	25.94	198	90.00
edu	935	663	70.91	272	29.09	248	91.18
gov	257	202	78.60	55	21.40	55	100.00
info	14	10	71.43	4	28.57	4	100.00
int	8	4	50.00	4	50.00	4	100.00

net	107	89	83.18	18	16.82	16	88.89
org	6686	4997	74.74	1689	25.26	1633	96.68
others	20	18	90.00	2	10.00	0	0.00
Total	9287	6881	74.09	2406	25.91	2293	95.30

The lowest were others (10%), .net (16.82%), .gov (21.40%), and .org (25.26%). Recovery rates were highest for .ac, .gov, .info, and .int (100%) and lowest for .net (88.89%), .com (90%), and .edu (91.18%).

6.5. Summary of File Formats of Vanished and Recovered Web Citations

Table 5 summarizes file formats linked to vanished and recovered web citations. The study shows .ppt files had the highest vanishing rate (100%), trailed by .doc (80%) and .pdf (46.07%). Conversely, .jsp (14.29%), others (16.67%), and .cgi (18.52%) had the lowest disappearance rates.

Table 5: Vanished and recovered web citations by file formats

File formats	No. of web citations	Active web citations	%	Vanished web citations	%	Recovered web citations	%
.asp	86	48	55.81	38	44.19	29	76.32
.cfm	38	23	60.53	15	39.47	15	100.00
.cgi	27	22	81.48	5	18.52	4	80.00
.doc	5	1	20.00	4	80.00	4	100.00
.html	8086	6175	76.37	1911	23.63	1818	95.13
.jsp	7	6	85.71	1	14.29	1	100.00
.pdf	814	439	53.93	375	46.07	368	98.13
.php	217	162	74.65	55	25.35	52	94.55
.ppt	1	0	0.00	1	100.00	1	100.00
others	6	5	83.33	1	16.67	1	100.00
Total	9287	6881	74.09	2406	25.91	2293	95.30

For recovered web citations, .cfm (100%), .doc (100%), .jsp (100%), and .ppt (100%) were most successfully restored, while .asp (76.32%), .cgi (80%), and .php (94.55%) had the poorest recovery rates.

6.6. Path Depth Associated with Vanished and Recovered Web Citations

Table 6 presents the path depth of vanished and recovered web citations. The highest vanishing rates occur at PD=8 (83.54%), PD>8 (81.82%), and PD=5 (38.43%), while the lowest are at PD=0 (13.24%), PD=2 (23.08%), and PD=3 (24.19%). Pearson's correlation confirms a strong positive relationship between PD and vanishing rates ($r = 0.840$, $p = .002$), indicating longer paths correlate with higher disappearance.

Table 6: Path depth associated with vanished and recovered web citations

Path Depth	No. of web citations	Active web citations	%	Vanished web citations	%	Recovered web citations	%
PD=0	204	177	86.76	27	13.24	25	92.59
PD=1	661	471	71.26	190	28.74	176	92.63
PD=2	5957	4582	76.92	1375	23.08	1326	96.44
PD=3	926	702	75.81	224	24.19	206	91.96
PD=4	654	452	69.11	202	30.89	189	93.56

PD=5	445	274	61.57	171	38.43	162	94.74
PD=6	195	127	65.13	68	34.87	64	94.12
PD=7	111	73	65.77	38	34.23	37	97.37
PD=8	79	13	16.46	66	83.54	63	95.45
PD>8	55	10	18.18	45	81.82	45	100.00
Total	9287	6881	74.09	2406	25.91	2293	95.30

For recovered citations, the highest rates are at PD>8 (100%), PD=7 (97.37%), and PD=2 (96.44%), while the lowest are at PD=3 (91.96%), PD=1 (92.63%), and PD=4 (93.56%). Pearson's analysis also reveals a positive correlation between PD and recovery ($r = 0.724$, $p = .018$), suggesting longer paths are linked to both higher vanishing and recovery. Therefore, the null hypothesis is rejected, and the alternative hypothesis is accepted. These findings highlight the complex dynamics of web citation persistence, where deeper paths influence both loss and restoration.

6.7. HTTP Errors Associated with Vanished and Recovered Web Citations

Table 7 shows HTTP errors linked to vanished and recovered web citations. HTTP 403 has the highest vanishing rate (56.57%), followed by HTTP 404 (29.51%) and HTTP 500 (11.31%).

Table 7: Details of HTTP errors associated with vanished and recovered web citations

HTTP errors	Vanished web citations	%	Recovered web citations	%
301	1	0.04	1	100.00
302	17	0.71	15	88.24
400	5	0.21	3	60.00
403	1361	56.57	1334	98.02
404	710	29.51	663	93.38
410	9	0.37	9	100.00
500	272	11.31	238	87.50
502	23	0.96	22	95.65
503	8	0.33	8	100.00
Total	2406	100.00	2293	95.30

The lowest rates are for HTTP 301 (0.04%), HTTP 503 (0.33%), and HTTP 410 (0.37%). Over 90% of vanished citations are recovered annually via the Wayback Machine. HTTP 301, 410, and 503 achieve 100% recovery, while HTTP 400 (60%), HTTP 500 (87.50%), and HTTP 302 (88.24%) have the lowest recovery rates.

6.8. Testing of Hypotheses

Table 8 presents the results of hypothesis testing for the study. The hypotheses were tested using correlation analysis. For each hypothesis, the calculated p-value was compared with the standard significance level (? = 0.05). When the p-value was less than 0.05, the null hypothesis was rejected, and the alternative hypothesis was accepted, indicating a statistically significant relationship between the variables. Accordingly, for H1, the p-value ($0.000 < 0.05$) suggests that the null hypothesis is rejected, confirming a significant positive trend in the increased use of web citations from 2015 to 2024. Similarly, for H2, the p-value ($0.018 < 0.05$) indicates rejection of the null hypothesis, establishing a significant negative correlation between the path depth and the percentage of recovered web citations. Therefore,

hypotheses H1 and H2 are statistically supported by the data.

Table 8: Details of testing of hypotheses

Sl. No	Hypotheses	Statistical test	p-value	Result
H1	Over a period of time from 2015-2024, there has been an increased use of web citations.	Co-relation	.000	Supported
H2	The path depth and percentage of recovered web citations are negatively correlated.	Co-relation	.018	Supported

7. Discussion

This study identified a rising trend in web citations within journal articles, increasing from 21.49% in 2015 to 66.01% in 2023, aligning with prior studies conducted by Kumbar & Niveditha (2020), Shah & Anayat (2018) and Satyanarayana & Damodar (2022). However, web citation decay poses a challenge, with 25.91% of links vanishing. Active web citations accounted for 74.09%, while vanished web citations declined from 42.17% in 2015 to 20.27% in 2024. Notably, URLs with a path depth exceeding 8 had the highest decay rate (85%), consistent with earlier findings revealed by McCown et al. (2005); Goh & Ng (2007). The study also revealed that HTTP 403 errors were the most prevalent (56.57%) among error types. To address web decay, web archiving tools like the Wayback Machine were utilized, successfully recovering over 90% of vanished web citations. Previous research studied by Sharma et al. (2022), Howell & Burtis (2023), and Loan et al. (2024) proved its effectiveness in preserving web-based citations.

Key findings demonstrate web citation usage increased from 21.49% to 66.01%, with declining proportions of vanished citations despite growing usage, predominance of HTTP 403 errors among failures, and remarkably consistent 95.30% recovery rates via the Wayback Machine regardless of citation age or complexity. These outcomes confirm that while longer path depths contribute to decay as prior research suggested (Goh & Ng, 2007; Kumar & Sampath Kumar, 2017), robust archival tools effectively mitigate access risks.

8. Conclusion

This study reveals a growing trend of researchers using web citations in scholarly works, yet their decay over time poses significant access challenges. Through systematic analysis of C&RL journal articles from 2015-2024, this research achieved comprehensive documentation of web citation patterns, decay rates, and recovery effectiveness, providing novel decade-long evidence demonstrating both increasing reliance on web sources and the critical importance of preservation strategies. The study's novelty lies in establishing empirical correlations between URL path depth and citation decay while documenting the Wayback Machine's exceptional recovery capabilities, offering actionable insights for citation design and preservation policy. To address citation decay, publishers, editors, and authors should thoroughly review web citations before publication, maintain digital backups, adopt standardized file formats and stable domains, and utilize web archiving tools including WebCite and DOIs. Future research should extend this framework across disciplines to compare decay signatures and evaluate preservation mandate impacts. The academic community must foster preservation awareness through diverse repositories while pursuing international policies to safeguard scholarly web content.

References

Bansal, S. (2021). Decay of URL References cited in DESIDOC Journal of Library & Information Technology. *Library Philosophy and Practice* (e-Journal), 1-13. https://doi.org/https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=10743&context=libphilprac

Germain, C. A. (2000). URLs: Uniform Resource Locators or Unreliable Resource Locators. *College & Research Libraries, 61*(4), 359-365. https://doi.org/10.5860/crl.61.4.359

Goh, D. H., & Ng, P. K. (2007). Link decay in leading information science journals. *Journal of the American Society for Information Science and Technology, 58*(1), 15-24. https://doi.org/10.1002/asi.20513

Gul, S., Mahajan, I., & Ali, A. (2014). The growth and decay of URLs citation: A case of an online Library & Information Science journal. *Malaysian Journal of Library & Information Science, 19*(3), 27-39. https://doi.org/https://mjlis.um.edu.my/article/view/1781/2525

Howell, S., & Burtis, A. (2023). The continued problem of URL decay: an updated analysis of health care management journal citations. *Journal of the Medical Library Association, 110*(4), 463-470. https://doi.org/10.5195/jmla.2022.1456

Isfandyari Moghaddam, A., Saberi, M. K., & Mohammad Esmaeel, S. (2010). Availability and Half-life of Web References Cited in Information Research Journal: A Citation Study .International *Journal of Information Science and Management, 8*(2), 57-75. https://ijism.isc.ac/article_698149_2f3b10898eabe799e626be2f8eeae18b.pdf

Kumar, V., & Sampath Kumar, B. T. (2017). Finding the unfound: Recovery of missing URLs through Internet Archive. *Annals of Library and Information Studies, 64*(3), 165-171. https://doi.org/http://op.niscair.res.in/index.php/ALIS/article/view/16709

Kumbar, M., & Niveditha, B. (2020). Permanence and characteristics of URLs cited in the journal scientometrics: A study using PHP script. *International Journal of Information Dissemination and Technology, 10*(4), 201-205. https://doi.org/10.5958/2249-5576.2020.00037.0

Lawrence, S., Pennock, D. M., Flake, G. W., Krovetz, R., Coetzee, F. M., Glover, E., Nielsen, F. A., Kruger, A., & Giles, C. L. (2001). Persistence of Web references in scientific research. *Computer, 34*(3), 26-31. https://doi.org/10.1109/2.901164

Loan, F. A., Khan, A. M., Andrabi, S. A. A., Sozia, S. R., & Parray, U. Y. (2024). Giving life to the dead: role of the Wayback Machine in recovery of dead URLs. *Data Technologies and Applications, 58*(2), 201-213. https://doi.org/10.1108/DTA-06-2022-0242

Loan, F. A., & Shah, U. Y. (2020). The decay and persistence of web references. *Digital Library Perspectives, 36*(2), 157-166. https://doi.org/10.1108/DLP-02-2020-0013

Markwell, J., & Brooks, D. W. (2003). "Link rot" limits the usefulness of web?based educational materials in biochemistry and molecular biology*. *Biochemistry and Molecular Biology Education, 31*(1), 69-72. https://doi.org/10.1002/bmb.2003.494031010165

McCown, F., Chan, S., Nelson, M. L., & Bollen, J. (2005). The Availability and Persistence of Web References in D-Lib Magazine. *ArXiv Preprint*. https://doi.org/https://doi.org/10.48550/arXiv.cs/0511077

Niveditha, B., Kumbar, M., & Sampath Kumar, B. T. (2022). Rotten web citations cited in scholarly journals: use of time travel for retrieval. *Aslib Journal of Information Management, 74*(2), 225-243. https://doi.org/10.1108/AJIM-05-2021-0139

Satyanarayana, D., & Damodar, P. (2022). Web Citation Analysis on Journal of Travel Research: A Study. *Library Progress (International), 42*(2), 412-420. https://doi.org/10.5958/2320-317X.2022.00039.3

Shah, U. Y., & Anayat, S. (2018). Web referencing in online scholarly world: a case study of library and information science research. International Journal of Information Movement, 2(9), 104-112. https://doi.org/https://www.ijim.in/wp-content/uploads/2018/01/Vol-2-Issue-IX-104-112-Paper-2014-Ufaira-Yaseen-Shah-WEB-Referencing-in-online-scholarly-world.pdf

Sharma, N., Agarahari, A., & Singh, S. N. (2022). 'Availability And Persistency Of Web Resources On Assam Movement?: A Study Of Scopus Indexed Research Articles. *Webology, 19*(4), 57-73. https://doi.org/https://www.webology.org/abstract.php?id=3279

Sife, A. S., & Lwoga, E. T. (2017). Retrieving vanished Web references in health science journals in East Africa. *Information and Learning Science, 118*(7/8), 385-392. https://doi.org/10.1108/ILS-04-2017-0030

Singh, T. G., & Devi, K. S. (2024). Web decay analysis and digital archiving of websites of technical institutions: a view from Wayback Machine. *C ollege Libraries, 39*(I), 1-10.https://collegelibraries.in/index.php/CL/article/view/142

Gurusiddesh Mugannavar, Librarian at Christ Academy Institute for Advanced Studies, Bangalore, has 8+ years of experience in Academic Library Systems andresearch. He holds MLISc and PGDLAN degrees and is pursuing a PhD atTumkur University. He has published, presented widely, and supportsteaching-learning through modern Library Services.

Dr. B. T. Sampath Kumar is a Senior Professor in the Department of Library and Information Science, Tumkur University. He has published over 230 research papers in reputed National and International journals and conferences, and has presented papers at international conferences in the UAE, Thailand, Vietnam, Cambodia, and Sri Lanka.