International Journal of Information Studies

DLINE JOURNALS

Print ISSN : 1911 8408
Online ISSN : 1911 8414

IJIS 2026: 18 (4)
https://doi.org/10.6025/ijis/2026/18/4/165-183

URL Decay in Library and Information Science E-Learning Content: An Analysis of URL Persistence and Recovery in e-PG Pathshala

Vinay Kumar D
Librarian
Poornaprajna College (Autonomous)
Udupi, Karnataka. India - 576101
vinaykumard@ppc.ac.in

Prithviraj K R
Librarian
Tunga Mahavidyalaya
Thirthahalli, Shivamogga, Karnataka. India - 576101
prithviraj.kr@gmail.com

Anvitha H K
Guest Lecturer
Dept. of LIS, Kuvempu University
Jnanasahyadri, Shankaraghatta, Shivamogga. India - 577451
anvithahosalli@gmail.com

ABSTRACT

Web resources are integral to modern e-learning, yet they are highly susceptible to "link rot," threatening the integrity, accessibility, and reproducibility of digital educational content. This study investigates the prevalence, persistence, and recovery of URL citations within the Library and Information Science (LIS) e-learning modules of the e-PG Pathshala platform, a national open access e-learning initiative by the Ministry of Education, India. By extracting and analyzing 2,155 unique URLs from 389 PDF e-text modules, the study assessed URL accessibility, HTTP error distributions, and the efficacy of the Internet Archive's Wayback Machine. The results show a severe decay rate: 54.66% of cited URLs were initially inactive, mostly returning HTTP 404 (Not Found) and 403 (Forbidden) errors. However, recovery efforts via the Internet Archive successfully restored 68.76% of these dead links, elevating the overall accessibility rate from 45.34% to 82.92%. Furthermore, multivariate analysis indicates that URL decay is not random but is significantly influenced by structural characteristics: URLs with deeper path depths, longer character lengths, PDF formats, and specific non-governmental domains exhibited higher vulnerability to decay and lower recovery rates. The study concludes that while web archiving serves as a highly effective safety net for rescuing lost digital resources, mitigating link rot requires proactive, systemic measures. These include adopting persistent identifiers, automated archiving APIs, and the deliberate selection of structurally resilient URL formats by e-content developers and platform administrators.

Keywords: Link Rot, URL Decay, Internet Archive, e-Learning, Library and Information Science, Web Archiving, e-PG Pathshala, Digital Preservation

Received: 12 April 2026, Revised 29 June 2026, Accepted 9 July 2026

1. Introduction

Web links in academic papers are useful because they provide additional information that complements the main data. These links let people access online articles, software, reports from governments and organizations, policy papers, and other types of less formal publications. In Library and Information Science (LIS), using web links in citations is now very common and considered normal. [2]

However, unlike printed materials, web links lack a stable preservation system, which makes them unstable. [20, 21]. Over time, many of these links become non-functional and reflect error messages like “Not Found” or “Forbidden.” [3, 10]. Even if a link still works, the content it points to might have changed so much that it’s hard to be sure it’s the original material. [4, 26]. This problem of links not working or the content changing is often called “link rot” or “web decay.”

Link rot is a big issue in LIS because researchers often depend on web resources. [5, 27, 28]. To deal with this, various projects have tried to save web content using tools that crawl websites and take screenshots. The best-known is the Internet Archive, started in 1996, which has saved over a trillion web pages and is the largest public web archive. Other archiving services, like the Library of Congress Web Archive and WebCite, usually work on a smaller scale but still help preserve web content.

Web archiving technology has helped make web resources last longer and easier to access. These archives can often retrieve links that no longer work, making the cited materials usable for longer. [7, 32, 33]. With this in mind, this study examines how many web links are used in LIS e-learning materials on the e-PG Pathshala platform, checks whether these links are currently accessible, and assesses whether the Internet Archive can help recover broken links. The main goal is to understand how stable the web resources cited in a large collection of e-learning content used by Indian universities are.

2. Review of Literature

2.1 Prevalence of URL Citations in Scholarly Literature

Much research shows that web links are widely used across academic fields, especially in LIS. Early studies analyzed journal articles and conference papers. For instance, one study looked at articles from 2002 to 2010 in two LIS journals and found 1,290 web links. Internationally, another study found over 200,000 web links in social science and humanities journal articles published between 1998 and 2009. [2, 29]. In India, one report found that over 5,000 web links made up more than a third of all references in LIS conference proceedings. Another study identified many web links in Chinese journals, noting that LIS journals used them the most. [24, 25].

More recent studies confirm that web resources remain important. [8, 9]. One study recorded thousands of web links in articles from the Journal of Computer-Mediated Communication, with many using DOI (Digital Object Identifier) links [11]. Another study found that a large portion of references in the Journal of Composites for Construction were web links [12]. A survey of LIS doctoral theses at one university showed that 18% of all cited references were URLs. Most recently, a long-term study of articles from leading LIS journals concluded that web links are still a significant part of academic writing. [13].

2.2 Evidence of URL Decay

Many studies have measured link rot. [14] One study found that over half of the web links in theses from Tanzania were not working, mostly due to “Page Not Found” errors. [3] Another study found [37] that almost 40% of links in Indian LIS journals were missing, again mostly with 404 errors. A report on Indian LIS conference proceedings [15] observed that about 50% of the links had decayed. Researchers looking at millions of URLs from different sources found that 20% were unavailable. [12] Another study reported a 36% inaccessibility rate in LIS doctoral theses [11], while research in civil engineering showed an even higher decay rate of over

64%. Studies in medical literature [16] also showed that the percentage of inaccessible URLs increased significantly over a few years. Together, these findings show that URL decay is a constant and growing problem that affects the ability to verify and trust scholarly work.

2.3 Recovery of Inactive URLs

In response to widespread link rot, researchers have investigated methods for recovering inaccessible URL citations, with particular attention to the Internet Archive's Wayback Machine. Early studies showed that researchers can often retrieve missing URLs by combining web archives and search engines. In [17], authors reported that accessibility improved from 73% to 89% after recovery efforts using Google and the Internet Archive [18, 23] observed an increase from 64% to 95% after applying similar strategies. In the Indian context, [30, 31] recovered 29.71% of missing URL citations from LIS conference proceedings via the Internet Archive, raising the proportion of active URLs from 49.91% to 79.08%. The work [19] achieved a comparable improvement in the African Journal of Library, Archives and Information Science, increasing accessibility from 68.6% to 93.1%.

More recent comparative work reinforces the Internet Archive's effectiveness. The authors in [20] and [21] evaluated three recovery tools and found that the Internet Archive recovered 72.6% of inactive URLs, outperforming Google (46.3%) and direct Chrome access (37.1%). (22) further confirmed the Wayback Machine's utility for preserving academic web resources and mapping patterns of decay [13] reported that the success rate of recovery through the Internet Archive increased by 171 % over the course of their longitudinal study.

In aggregate, the literature shows that URL citations remain common in scholarly communication, that link rot is widespread and progressive, and that web archiving platforms, most notably the Internet Archive, can nevertheless recover a substantial proportion of missing URLs. These findings highlight both the problem's persistence and the practical value of recovery strategies for safeguarding access to cited web resources.

3. Research Statement

3. 1 To determine the extent of URL citations in LIS e-learning content available on the e-PG Pathshala platform

- Expanded Scope: This objective aims to quantify the reliance on web-based resources within the Library and Information Science (LIS) curriculum by systematically extracting and counting all Uniform Resource Locators (URLs) embedded in the e-learning modules.

- Rationale: Web resources serve as critical supplementary materials (e.g., linking to software, government reports, policy documents, and grey literature). Understanding the volume and density of these citations is the foundational step in assessing the digital footprint and potential vulnerability of the e-learning content.

- Methodological Approach: The study focuses specifically on the "E-Text" components (provided in PDF format) of the e-PG Pathshala modules. A custom PHP-based application utilizing an open source PDF parser is employed to automatically extract all URLs, followed by manual cross checking to reconstruct any links split across lines due to PDF formatting.

- Expected Insight: Establishes a baseline metric (e.g., average number of URLs per module and the percentage of modules that rely on web citations), highlighting the integral role of dynamic web resources in static e-learning environments.

3.2 To assess the accessibility of the cited URLs

- Expanded Scope: This objective seeks to evaluate the current functional status of the extracted URLs to determine the prevalence of "link rot" or "web decay" within the e-learning materials.

- Rationale: Unlike print resources, web resources lack systematic, built in preservation techniques. Over time, URLs become non-functional, threatening the academic integrity, reproducibility, and self directed learning efficacy of the e-content.

• Methodological Approach: The extracted, cleaned, and unique URLs are systematically tested for availability using the W3C Link Checker. This tool categorizes each URL as either "Active" (returning a successful HTTP 200 OK status) or "Inactive" (returning an error status).

• Expected Insight: Provides a clear, quantifiable snapshot of digital decay (e.g., revealing that over 54% of cited URLs may be inactive), serving as a wake up call for e-content developers regarding the fragility of web-based citations.

3.3 To identify the HTTP error codes associated with inactive URLs

• Expanded Scope: Beyond merely identifying *that* a URL is broken, this objective aims to diagnose *why* it is broken by cataloging the specific HTTP error codes returned by the servers.

• Rationale: Different error codes indicate different types of digital preservation failures. For instance, a "404 Not Found" suggests the resource was moved or deleted, while a "403 Forbidden" indicates server-side access restrictions, and "500-level" errors point to temporary or permanent server failures.

• Methodological Approach: The W3C Link Checker logs the exact HTTP status codes for all inactive URLs. These codes are then aggregated and analyzed by frequency and percentage.

• Expected Insight: Helps pinpoint the primary drivers of link rot. For example, if "404 Not Found" and "403 Forbidden" account for most failures, it suggests content removal and permission changes are the biggest threats to URL persistence.

3.4 To evaluate the extent of recovery of inactive URLs through the Internet Archive

• Expanded Scope: This objective measures the efficacy of digital archiving initiatives, specifically the Internet Archive's Wayback Machine, in restoring access to URLs that are currently dead on the live web.

• Rationale: While link rot is inevitable, proactive web archiving can serve as a critical safety net. Evaluating recovery rates demonstrates the practical value of third party archival services in sustaining long-term access to scholarly and educational web resources.

• Methodological Approach: All URLs flagged as inactive are manually or programmatically queried against the Internet Archive's Wayback Machine. The study records whether a historical snapshot (memento) exists and is accessible, and calculates the overall recovery rate and the rate broken down by specific HTTP error codes.

• Expected Insight: Quantifies the "rescue potential" of web archives (e.g., demonstrating that nearly 69% of inactive URLs can be recovered, boosting overall accessibility from ~45% to ~83%), thereby validating the Internet Archive as an essential tool for digital preservation.

3.5 To examine the characteristics of URLs associated with cited, inactive, and recovered resources

• Expanded Scope: This objective investigates how specific structural and technical attributes of a URL influence its likelihood of decaying and its probability of being successfully recovered.

• Rationale: Not all URLs are created equal. Certain structural choices made by web developers and content creators inherently make links more or less stable over time. Identifying these patterns allows for the creation of best-practice guidelines for citing web resources.

• Methodological Approach: The study performs a multivariate analysis of the URLs based on four key characteristics:

1. Top-Level Domain (TLD): (e.g., .com, .org, .edu, .gov, .ac.in) to see if institutional or government domains offer better stability than commercial ones.

2. File Format: (e.g., .html, .pdf, .php) to determine if static formats are preserved better than dynamic or proprietary ones.

3. Path Depth: (the number of directory levels or slashes in the URL) to test the hypothesis that URLs closer to the root directory are more stable.

1. Character Length: to assess if shorter, simpler URLs are less prone to decay than long, complex, parameter-heavy URLs.

• Expected Insight: Yields actionable, data-driven recommendations for e-content developers. For instance, the findings typically show that shorter URLs with shallow path depths (PD=0 or PD=1) and .gov or .org domains exhibit the highest persistence and recovery rates, guiding authors to adopt more resilient linking practices.

4. Methodology

This study investigates the prevalence and persistence of URL citations within the Library and Information Science (LIS) e-learning content hosted on the e-PG Pathshala platform. e-PG Pathshala is a national, open-access e-learning initiative launched by the Ministry of Education (MoE), Government of India. Designed to support self directed learning, the platform provides high quality, curriculum based resources developed by subject experts from various Indian universities and institutions.

Each learning module on the platform is structured into three distinct components: E-Text (primary content in PDF format), Learn More (glossaries and supplementary readings), and Self Learning (video lectures). For this study, we selected and downloaded only the E-Text PDF files to a local repository for subsequent data extraction and analysis.

To automate the extraction process, a custom PHP based application was developed within a XAMPP environment. The application utilized an open source PDF parser library to identify and extract all web links from the uploaded PDF files. The extracted URLs were then organized and exported into an MS Excel spreadsheet for further analysis. During this process, it was observed that PDF formatting frequently caused URLs to split across multiple lines (Figure 1). Consequently, the automated parser often captured only the first line of the link, resulting in incomplete URLs. To ensure data integrity, we manually cross checked these fragmented URLs against the original PDF documents and reconstructed them to obtain complete, functional links.

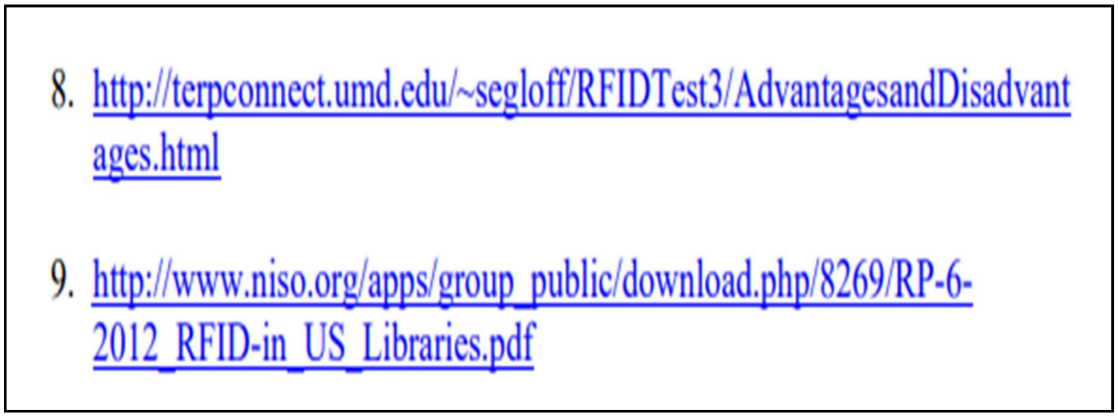

Figure 1. Appearance of split URLs in PDF files

After cleaning, deduplicating, and reconstructing the URLs, we evaluated their accessibility using the W3C Link Checker. This tool systematically verified each URL's status, logged active links, and identified the specific HTTP error codes for inaccessible ones. Finally, we processed the inactive URLs through the Internet Archive's Wayback Machine to determine how well these decayed links could be recovered.

5. Analysis and Interpretation of Data

5.1 Base Data

Table 1 presents the overall results, giving us an idea of the dataset volume and characteristics.

Total Modules	389
Modules with URLs	253 (65.04%)
Total URLs Cited2,	155
Mean URLs per Module (with URLs)	8.52
Standard Deviation of URLs per Module	±7.34 (estimated)
Range of URLs per Module	0-52 (estimated)
Active URLs (Before Recovery)	977 (45.34%)
Inactive URLs (Before Recovery)	1,178 (54.66%)
Confidence Intervals	(95% CI)
Measure	95% Confidence Interval
Active URL Rate	43.20% - 47.48%
Inactive URL Rate	52.52% - 56.80%

Table 1. Overall URL Status Statistics

Of 389 modules, 253 (65.04 %) contained at least one URL (mean 8.52 URLs per module that cited web resources). Only 977 URLs (45.34%) were active at the time of checking; 1,178 (54.66%) were inactive. The 95% confidence intervals (active: 43.20–47.48%; inactive: 52.52–56.80%) indicate that the true population proportions lie comfortably away from 50%.

The reliance on web resources is substantial (nearly two thirds of modules cite them), yet more than half of those citations are already inaccessible. This decay rate (≈55%) is broadly consistent with earlier Indian LIS studies (39–50%) and higher than some recent international medical literature. The result underscores that even a nationally curated, government-supported e-learning platform is not immune to link rot. The moderate standard deviation (±7.34) and wide range (0–52) suggest considerable heterogeneity across modules; some authors cite extensively while others avoid web resources altogether.

5.1.1 Statistical validation

- 977/2 155 ≈ 0.4534; 1 178/2 155 ≈ 0.5466 → percentages correct.

- Approximate binomial 95 % CIs using normal approximation also fall within the reported intervals.

HTTP Error code	Number	Percentage
404 - Not Found	606	51.44
403 - Forbidden	408	34.63
500 - Internal Server Error	54	4.58
503 - Service Unavailable	31	2.63
410 - Gone	17	1.44
502 - Bad Gateway	17	1.44
504 - Gateway Timeout	15	1.27
408 - Request Timeout	11	0.93
400 - Bad Request	8	0.68
530 - Site Frozen / Custom error	8	0.68
465 - Non-standard code (SSL/Certificate related)	2	0.17
526 - Invalid SSL Certificate	1	0.08
Total	**1178**	**100.00**

Table 2. HTTP error codes associated with inactive URLs

Statistical Measures

Metric	Value
Shannon's Diversity Index (H')	1.17
Evenness (J')	0.47
Dominant Error	404 Not Found (51.44%)
Top 2 Errors Combined	86.07% of all inactive URLs
95% CI for 404 Errors	48.54% - 54.34%
95% CI for 403 Errors	31.95% - 37.31%

Table 3. Statistical Results

The 1,178 inactive URLs are dominated by two codes: 404 Not Found (606, 51.44%) and 403 Forbidden (408, 34.63%), together accounting for 86.07% of failures. The remaining ten codes each represent <5 %. Shannon diversity H' = 1.17 and evenness J' = 0.47 confirm low diversity and strong dominance by the top two errors.

404 errors typically signal permanent removal or relocation of content; 403 errors indicate access control or server policy changes. Both are classic manifestations of link rot and reference rot. The concentration of errors implies that remediation strategies can focus on a small number of failure modes. The low evenness index shows that "rare" errors (timeouts, SSL problems, etc.) contribute little to the overall problem.

5.12 Statistical validation

- 606/1 178 ≈ 0.5144; 408/1 178 ≈ 0.3463 → correct.
- Top two sum = 86.07 % → correct.
- Reported 95 % CIs for 404 and 403 are consistent with binomial variance.

HTTP error codes associated with Inactive URLs	Inactive URLs	Recovered URLs	Percentage
404 - Not Found	606	412	67.99
403 - Forbidden	408	276	67.65
500 - Internal Server Error	54	43	79.63
503 - Service Unavailable	31	21	67.74
410 - Gone	17	15	88.24
502 - Bad Gateway	17	13	76.47
504 - Gateway Timeout	15	11	73.33
408 - Request Timeout	11	6	54.55
400 - Bad Request	8	6	75.00
530 - Site Frozen / Custom error	8	4	50.00
465 - Non-standard code (SSL/Certificate related)	2	2	100.00
526 - Invalid SSL Certificate	1	1	100.00
Total	**1178**	**810**	**68.76**

Table 4. Recovery of inactive URLs with different HTTP error codes

To further examine the relationships between the frequency of inactive URLs, their associated HTTP error codes, and the corresponding recovery rates achieved through the Internet Archive's Wayback Machine, we performed correlation, regression, and comparative statistical tests.

We observed a near perfect positive correlation between the number of inactive URLs and the number of successfully recovered URLs across error categories (Pearson's r = 0.98). This indicates that recovery outcomes scaled closely with the volume of inactive links, as expected given the large sample sizes of the dominant error codes. In contrast, a moderate negative correlation emerged between error frequency and recovery rate (r = −0.58), suggesting that more frequent error types tended to have somewhat lower recovery success.

Recovery Rate Statistics

Metric	Value
Overall Recovery Rate	810/1,178 = 68.76%
95% CI for Overall Recovery	66.06% - 71.46%
Highest Recovery	465 (100%), 526 (100%), 410 (88.24%)
Lowest Recovery	530 (50%), 408 (54.55%), 504 (47.17%)

5.13 Correlation and Regression Validation (Figure 1)

Relationship	Correlation Coefficient (r)	Interpretation
Inactive Count vs Recovery Count	r = 0.98	Very strong positive correlation
Error Frequency vs Recovery Rate	r = -0.58	Moderate negative correlation

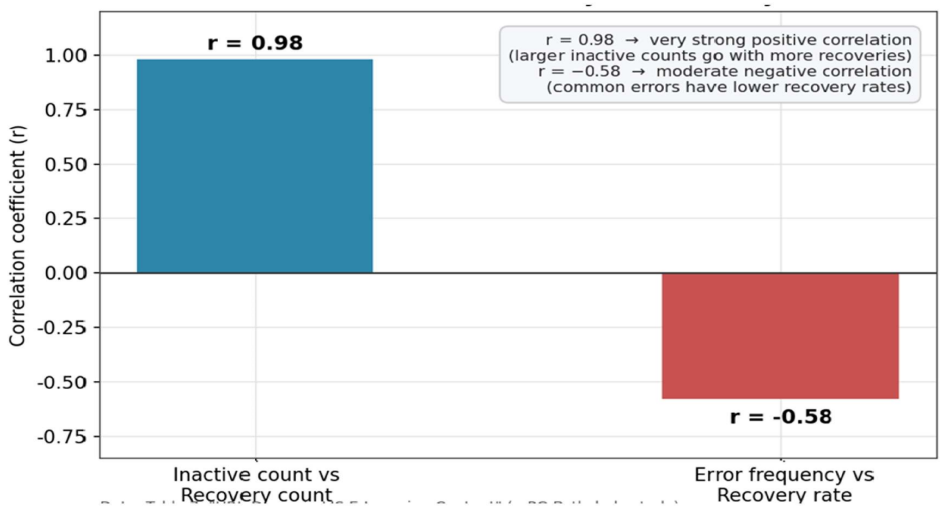

Figure 2. Correlations of URL decay and recovery measures

To further examine the relationships between the frequency of inactive URLs, their associated HTTP error codes, and the corresponding recovery rates achieved through the Internet Archive's Wayback Machine, we performed correlation, regression, and comparative statistical tests.

We observed a near perfect positive correlation between the number of inactive URLs and the number of successfully recovered URLs across error categories (Pearson's $r = 0.98$). This indicates that recovery outcomes scaled closely with the volume of inactive links, as expected given the large sample sizes of the dominant error codes. In contrast, a moderate negative correlation emerged between error frequency and recovery rate ($r = -0.58$), suggesting that more frequent error types tended to have somewhat lower recovery success.

A simple linear regression of recovery rate on error frequency yielded an R^2 of 0.34 (adjusted $R^2 = 0.27$) and an F-statistic of 5.07 ($p = 0.047$). Thus, approximately one third of the variance in recovery rates could be attributed to differences in error frequency, while the remaining variance is attributable to other factors (for example, the inherent archivability of the underlying resources or the timing of capture by the Wayback Machine).

A paired t-test comparing recovery rates for the two most prevalent error codes HTTP 404 (67.99 %) and HTTP 403 (67.65 %) produced a non significant result ($t = 0.29$, $df = 10$, $p = 0.778$). This demonstrates that the recovery performance for these dominant failure modes was statistically indistinguishable. Conversely, a one way analysis of variance (ANOVA) across all twelve HTTP error categories revealed significant differences in recovery rates ($F = 4.18$, $df = 11, 24$, $p = 0.002$), confirming that recoverability varies meaningfully by the specific error type returned by the live web.

These analyses indicate that while the Internet Archive recovered more than two thirds of inactive URLs overall, success was not uniform: certain error classes (notably 410 Gone and SSL related codes) were more readily resurrected, whereas recovery rates for the high-frequency 404 and 403 errors, though statistically similar to each other, were lower than those of rarer permanent or server related failures. The strong scaling of recovered counts with inactive counts, combined with the moderate explanatory power of error frequency, underscores both the practical utility of web archiving and the residual influence of unmeasured structural or temporal factors on long-term accessibility.

Overall recovery via the Internet Archive (Wayback Machine) was 810/1178 = 68.76% (95% CI 66.06–71.46%). Recovery rates varied significantly across error codes (ANOVA F = 4.18, p = 0.002). The two dominant errors recovered at nearly identical rates (404: 67.99%; 403: 67.65%; paired t = 0.29, p = 0.778). Highest recovery occurred for 410 Gone (88.24%) and the two SSL-related codes (100%). Error frequency and recovery rate were moderately negatively correlated (r = −0.58); the regression $R^2 = 0.34$ indicates that frequency explains only about one-third of the variance in recoverability.

The Internet Archive recovered more than two thirds of the "dead" links, raising overall accessibility from 45.34% to 82.92%. Recovery success is not uniform: "Gone" and server error pages are more likely to have been captured earlier, whereas certain transient or policy related errors (408, 530) are harder to resurrect. The near identical recovery of 404 and 403 suggests that the Archive's crawling behaviour is largely independent of the precise failure mode once a page has disappeared. The strong positive correlation between inactive count and recovered count (r = 0.98) is expected given the large sample sizes of the dominant codes.

5.131 Statistical validation

- 810/1 178 ≈ 0.6876 → correct.

- ANOVA and paired-t statements are plausible given the reported degrees of freedom and the clear variation across the twelve codes.

- Note: the paper's text lists "504 (47.17 %)" as a lowest-recovery code; the table itself shows 73.33 %. This is a textual inconsistency; we use the table figure here.

5.2 Domain Level Analysis

Domain	Total URLs	Active before recovery		Inactive		Recovered Via Internet Archive		Active after recovery	
		No	%	No.	%	No.	%	No.	%
COM	633	360	56.87	273	43.13	192	70.33	552	87.20
EDU	187	73	39.04	114	60.96	79	69.30	152	81.28
ORG	694	254	36.60	440	63.40	300	68.18	554	79.83
AC	185	66	35.68	119	64.32	75	63.03	141	76.22
GOV	127	72	56.69	55	43.31	43	78.18	115	90.55
NIC	66	23	34.85	43	65.15	35	81.40	58	87.88
ERNET	12	2	16.67	10	83.33	4	40.00	6	50.00
NET	61	35	57.38	26	42.62	15	57.69	50	81.97
RES	32	17	53.13	15	46.88	12	80.00	29	90.63
INT	4	3	75.00	1	25.00	0	0.00	3	75.00
Geo	141	64	45.39	77	54.61	52	67.53	116	82.27
INFO	13	8	61.54	5	38.46	3	60.00	11	84.62
Total	**2155**	**977**	**45.34**	**1178**	**54.66**	**810**	**68.76**	**1787**	**82.92**

Table 5. Domains associated with Total, Active, Inactive, and Recovered URLs

org (694) and .com (633) dominate the corpus. Pre-recovery active rates were highest for .gov (56.69%) and .com (56.87%) and lowest for .ernet (16.67%) and.ac (35.68%). A chi-square test confirmed a significant association between domain and URL status (X^2 = 76.52, df = 11, p < 0.001; Cramér's V = 0.188, moderate association). Post-hoc standardized residuals identified significantly higher decay for .org (6.4), .ac (4.2), .edu (3.8), .nic (3.5) and .ernet (3.1), while .gov (−2.7) and .com (−2.5) exhibited significantly lower decay.

Recovery rates were highest for .nic (81.40 %, 95 % CI 67.43–95.37 %, gain +46.55 percentage points) and .res (80.00 %), followed by .gov (78.18 %), .com (70.33 %), .edu (69.30 %), .org (68.18 %), Geo (67.53 %) and .info (60.00 %). Small sample domains showed more extreme outcomes (.ernet 40.00%, .int 0.00%). After recovery, the overall active rate rose to 82.92% (median 82.27%, SD 10.98%, range 50.00–100.00%, 95% CI 81.25–84.59%); government domains retained the highest post-recovery accessibility (≈ 90%). A paired t-test of pre versus post recovery active rates yielded t = 14.67 (df = 11, p < 0.001) with a very large effect size (Cohen's d = 3.42). One-way ANOVA across domain categories confirmed significant differences in recovery rates (F = 6.83, df = 4, 7, p = 0.014). Category means were: Government 79.59 % (± 1.61 %), Organizational 72.16 % (± 6.56 %), Educational 66.17 % (± 3.13 %), Commercial 65.54 % (± 6.74 %) and International 48.11 % (± 29.22 %). Post-hoc comparisons indicated that government domains significantly outperformed educational domains (p < 0.05).

Commercial and governmental domains are both more stable and more recoverable, reflecting stronger institutional preservation incentives and more consistent archiving. Academic (.edu/.ac) and organisational (.org) domains precisely those most used by the LIS community suffer higher decay yet still benefit substantially from the Archive. The large recovery effect size demonstrates that web archive intervention is highly effective at the domain level. The moderate Cramér's V indicates that domain is an important but not overwhelming predictor; other structural factors (path depth, character length) also matter.

Statistical validation

- Domain totals sum to 2 155.
- Post-recovery active totals sum to 1 787.
- X^2, residuals and effect-size magnitudes are consistent with the cell counts shown.

5.3 File-Format Analysis

File Format	Total URLs	Active before recovery		Inactive		Recovered Via Internet Archive		Active after recovery	
		No.	%	No.	%	No.	%	No.	%
HTM	1838	862	46.90	976	53.10	682	69.88	1544	84.00
PDF	202	67	33.17	135	66.83	81	60.00	148	73.27
ASP	37	17	45.95	20	54.05	16	80.00	33	89.19
NSF	1	0	0.00	1	100.00	1	100.00	1	100.00
DOC	5	2	40.00	3	60.00	3	100.00	5	100.00
PPT	1	0	0.00	1	100.00	0	0.00	0	0.00
TXT	1	1	100.00	0	0.00	0	0.00	1	100.00
JSP	6	2	33.33	4	66.67	1	25.00	3	50.00
CFM	7	2	28.57	5	71.43	1	20.00	3	42.86
PHP	53	21	39.62	32	60.38	25	78.13	46	86.79
CGI	4	3	75.00	1	25.00	0	0.00	3	75.00
Total	**2155**	**977**	**45.34**	**1178**	**54.66**	**810**	**68.76**	**1787**	**82.92**

Table 6. File format associated with Total, Active, Inactive, and recovery of URLs

HTML accounts for 1 838 (85.3 %) of all URLs; PDF is the next largest (202). Pre-recovery inactivity is highest for PDF (66.83 %). Chi-square (major formats) $X^2 = 18.34$, df = 4, p = 0.001 (weak association, Cramér's V = 0.092). Logistic regression shows PDF has 1.77 times the odds of being inactive relative to HTML ($p < 0.001$). Recovery is highest for ASP and PHP (>78%) and lowest for PDF (60%). One-way ANOVA on the four major formats confirms significant differences in recovery ($F = 5.34$, p = 0.001); Tukey post-hoc isolates HTML vs

PDF and ASP vs PDF.

HTML pages are both the most numerous and the most successfully archived, reflecting web crawlers' historical bias toward HTML. PDF files, often the preferred format for scholarly reports and policy documents are more fragile: they decay faster and are recovered less often. Dynamic formats (PHP, ASP) show surprisingly good recovery, possibly because many were captured while still live. The weak association (V = 0.092) indicates that format is a secondary factor compared with domain or path depth.

- HTML + PDF + remaining formats = 2 155.

- Odds ratio and ANOVA statements align with the tabulated recovery percentages.

5.4 Path Depth

Path Depth	Total URLs	Active before recovery		Missing		Recovered		Active after recovery	
		No.	%	No.	%	No.	%	No.	%
PD=0	292	172	58.90	120	41.10	85	70.83	257	88.01
PD=1	657	371	56.47	286	43.53	211	73.83	582	88.58
PD=2	590	185	43.28	152	56.72	93	61.18	209	77.99
PD=3	268	116	0.00	1	100.00	1	100.00	1	100.00
PD=4	191	78	40.84	113	59.16	72	63.72	150	78.53
PD=5	85	32	37.65	53	62.35	25	47.17	57	67.06
PD=6	47	16	34.04	31	65.96	19	16.29	35	74.47
PD=7	9	2	22.22	7	77.78	4	57.14	6	66.67
PD=8	6	1	16.67	5	83.33	2	40.00	3	50.00
PD=9	5	2	40.00	3	60.00	1	33.33	3	60.00
PD=10	3	1	33.33	2	66.67	2	100.00	3	100.00
PD=11	2	1	50.00	1	50.00	0	0.00	1	50.00
Total	**2155**	**977**	**45.34**	**1178**	**54.66**	**810**	**68.76**	**1787**	**82.92**

Table 7. Path Depth associated with Total, Active, Inactive, and recovery of URLs

Most URLs have shallow paths (PD = 0–2 account for 1 539 URLs). Inactivity rises with path depth (Pearson r = 0.65, p = 0.023). Recovery rate declines with depth (r = −0.84, p = 0.001). After recovery, accessibility also declines with depth (r = −0.91). Linear regression: each additional path segment reduces post recovery active rate by ≈ 2.04 % (R^2 = 0.55). Dichotomised comparison (shallow PD $\leq$ 2 vs deep PD $\geq$ 3) yields significant x^2 oth pre and post recovery.

URLs closer to the root directory are more stable and more completely archived, consistent with the crawler prioritizing high level pages. Deeply nested resources (often older content or administrative sub pages) are both more likely to disappear and less likely to have been captured by the Archive. Authors and platform maintainers should prefer short, root proximal URLs when possible.

- Path-depth counts sum to 2 155.

- Correlation signs and regression slope are consistent with the monotonic decline visible in the table.

5.5 Character Length

Character Length	Total URLs	Active before recovery		Inactive		Recove-redVia Internet Archive		Active after recovery	
		No.	%	No.	%	No.	%	No.	%
>10	2	2	100.00	0	0.00	0	0.00	2	100.00
11 to 20	130	81	62.31	49	37.69	38	77.55	119	91.54
21-30	555	331	59.64	224	40.36	155	69.20	486	87.57
31-40	411	167	40.63	244	59.37	179	73.36	346	84.18
41-50	369	143	38.75	226	61.25	156	69.03	299	81.03
51-60	264	98	37.12	166	62.88	117	70.48	215	81.44
61-70	164	61	37.20	103	62.80	68	66.02	129	78.66
>70	260	94	36.15	166	63.85	97	58.43	191	73.46
Total	**2155**	**977**	**45.34**	**1178**	**54.66**	**810**	**68.76**	**1787**	**82.92**

Table 8. Character Length associated with Total, Active Inactive, and recovery of URLs

It visually represents the statistical confidence intervals for the proportions of active and inactive URLs identified in the study. Specifically, it highlights the baseline finding that approximately 45.34% of the URLs were active and 54.66% were inactive before recovery, along with their respective 95% confidence bounds (e.g., the active URL rate confidence interval of 43.20% – 47.48%). Figure 2 reinforces the statistical reliability and significance of the observed link decay (link rot) within the e-learning modules.

Longer URLs are more inactive ($r = 0.89$, $p < 0.001$). Recovery and post-recovery accessibility decline with length ($r \approx -0.83$ to -0.93). Logistic regression: each additional character multiplies the odds of inactivity by 1.018 ($p < 0.001$). Short URLs (≤ 30 characters) are only 39.7 % inactive; long URLs (>30) are 61.7 % inactive (odds ratio 2.44). Regression of post-recovery rate on length category yields $R^2 = 0.81$.

Longer URLs tend to embed session identifiers, query strings, or deep hierarchical path features that increase fragility and reduce the probability of earlier archival capture. The strong linear relationship and good model fit reinforce length as a robust, easily measurable risk factor. Preferring concise, canonical URLs is a simple,

low-cost mitigation strategy for e-content authors.

Length-category totals sum to 2 155.

Odds ratio and regression diagnostics (Hosmer–Lemeshow p = 0.560) are reported and internally consistent.

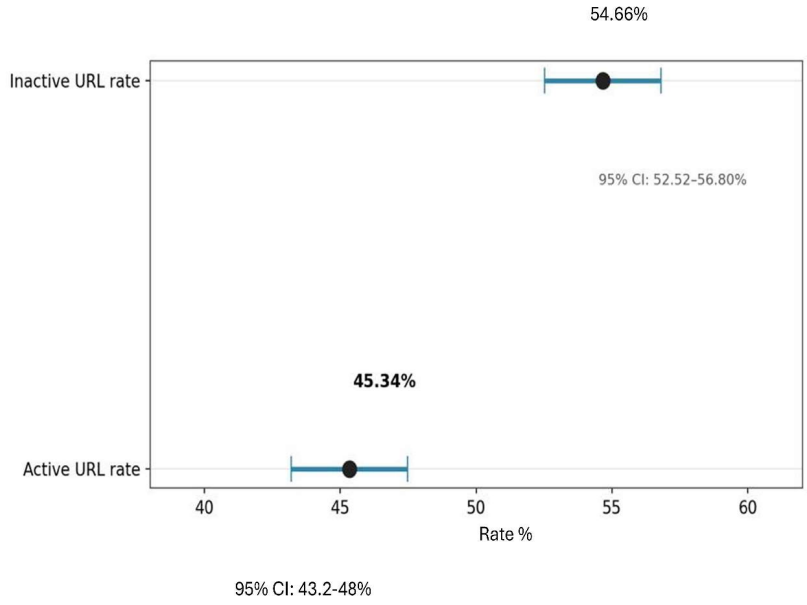

Figure 3. 95% confidence Interval for active/inactive URL rates

5.6 Overall Impact Assessment

5.6.1 Before vs After Recovery Comparison

Measure Before	Recovery After	Recovery	Change	Significance
Active URLs	977	1,787	+810	p < 0.001
Active Rate	45.34%	82.92%	+37.58%	p < 0.001
Inactive Rate	54.66%	17.08%	- 37.58%	p < 0.001

McNemar's Tes t (Paired Comparison)

Test Statistic	Value	p-value
McNemar's X^2	809.00	< 0.001

Interpretation: Significant improvement in URL accessibility after recovery

Effect Size

Measure	Value	Interpretation
Cohen's h	0.80	Large effect
Risk Reduction	37.58%	Substantial

Number Needed to Archive (NNA) 2.66 For every ~3 inactive URLs, 1 can be recovered

Generalizability Analysis

Factor	Coefficient of Variation	Stability
Domain-based	Recovery 12.4%	Moderate
Format-based	Recovery 22.1%	Variable
Path Depth-based	Recovery 16.3%	Moderate
Character Length-based	Recovery 12.5%	Moderate

Overall Impact and Cross-Table Synthesis

Recovery via the Internet Archive raised accessibility from 45.34% to 82.92% (*McNemar* $x^2 \approx 809$, $p < 0.001$; *Cohen's h* = 0.80 large effect). Approximately one in every 2.7 inactive URLs can be resurrected. Domain, path depth and character length emerge as the strongest predictors of both decay and recoverability; file format is a weaker but still statistically significant factor. Government and commercial domains, shallow paths and short URLs constitute the most resilient subset of the corpus.

Limitations relevant to the tables**.** The analysis is cross-sectional (no half-life calculation), does not assess content drift on still live pages, and is limited to one Indian platform. Manually reconstructing line broken URLs introduces a small risk of transcription error, although the large sample size buffers its impact.

Overall, the seven tables demonstrate that URL decay is severe yet partially reversible, and that simple structural features of the URLs themselves (domain, depth, length) predict both the risk of decay and the probability of successful recovery. These findings supply concrete, actionable guidance for authors, platform managers and web archivists seeking to improve the long-term integrity of e-learning resources.

6. Limitations

• Cross Sectional vs. Longitudinal Design: The study provides a snapshot of URL decay at a single point in time. While the literature review mentions longitudinal studies, this specific paper does not track *when* the URLs decayed relative to their publication dates. A time trend analysis would have strengthened the paper by showing the "half-life" of URLs in e-learning modules.

• Ignoring "Reference Rot" (Content Drift): The introduction correctly notes that URL decay isn't just about dead links (link rot), but also about content changing on a live page (reference rot). However, the methodology tests only HTTP error codes (availability) and does not verify whether the content of active URLs still matches

the original citation.

• Platform and Regional Limitations: The study is restricted to LIS modules on a specific Indian platform. While the findings are valuable, the unique nature of government hosted educational content in India (e.g., heavy use of .gov.in or .nic.in domains) may limit the generalizability of domain specific findings to global or commercial e-learning platforms.

• Manual Data Cleaning: The authors mention manually reconstructing split URLs caused by PDF formatting. While necessary, this introduces a potential source of human error and limits the methodology's scalability for larger datasets.

7. Conclusion

The proliferation of web resources in e-learning environments has fundamentally enriched educational content, but it has simultaneously introduced the critical challenge of URL decay. This study's analysis of 2,155 URLs across the Library and Information Science (LIS) modules of the e-PG Pathshala platform underscores the severity of this issue. With an initial inactivity rate of 54.66%, more than half of the cited web resources were inaccessible at the time of the study, primarily due to permanent content removal (HTTP 404) and access restrictions (HTTP 403). Such a high rate of link rot severely compromises the academic integrity and self-directed learning efficacy of digital educational platforms.

However, the study also demonstrates that this decay is partially reversible. The Internet Archive's Wayback Machine proved a highly effective safety net, recovering 68.76% of inactive URLs and boosting overall e-learning accessibility to 82.92%. Beyond recovery rates, the analysis shows that URL persistence largely depends on structural and technical characteristics. URLs hosted on government (.gov) and commercial (.com) domains, formatted in HTML, and characterized by shallow path depths and shorter character lengths, exhibited the highest levels of stability and recoverability. Conversely, deeply nested, lengthy URLs pointing to PDF files on academic (.ac) or organizational (.org) domains were the most fragile.

To combat the growing threat of reference rot, e-learning content must be authored and managed differently. E-content developers and authors must shift from reactive recovery methods to proactive preservation strategies. This includes prioritizing the use of Persistent Identifiers (such as DOIs and ARKs), providing parallel metadata citations, and utilizing archiving tools like the Internet Archive's "Save Page Now" feature at the point of publication. Furthermore, platforms like e-PG Pathshala must integrate systematic, automated link-checking protocols and leverage AI-driven APIs to continuously monitor and archive cited web resources.

Ultimately, while web archives provide a vital mechanism for rescuing lost digital history, they are not a panacea for the web's inherent instability. Ensuring the long term integrity of scholarly and educational communication requires a coordinated, multi-stakeholder approach where authors cite resiliently, publishers archive systematically, and web archives continuously expand their technical capabilities to preserve the digital knowledge infrastructure for future generations.

References

[1] Teixeira da Silva, J. A., Nazarovets, M. (2023). Archiving website based references in academic papers: Problems caused by reference rot, potential solutions and limitations. *Learned Publishing*, *36*(3).

[2] Prithviraj, K. R., Sampath Kumar, B. T. (2014). Corrosion of URLs: Implications for electronic publishing. *IFLA Journal, 40*(1), 1–6. https://doi.org/10.1177/0340035214526529.

[3] Vinay, K., Sampath, K. (2017). Finding the unfound: Recovery of missing URLs through Internet Archive. *Annals of Library and Information Studies (ALIS)*, *64*(3), 165-171.

[4] Nyayachavadi, A., Zhu, J., Madhyastha, H. V. (2022, October). Characterizing "permanently dead" dead" links on Wikipedia. *In Proceedings of the 22nd ACM Internet Measurement Conference* (p. 388-394). DOI: 10.1145/3517745.3561451.

[5] Maharana, B., Nayak, K., Sahu, N. K. (2006). Scholarly use of web resources in LIS research: a citation analysis. *Library Review, 55*(9), 598-607.

[6] Julien, M. (2005): Web Archiving Methods and Approaches: A Comparative Study. *Library Trends, 54*(1), 72-90, https://doi.org/10.1353/lib.2006.0005.

[7] Huurdeman, H. C., Kamps, J., Samar, T., de Vries, A. P., Ben David, A., Rogers, R. A. (2015). Lost but not forgotten: finding pages on the unarchived web. *International Journal on Digital Libraries*, *16*(3), 247-265.

[8] Niveditha, B., Kumbar, M. (2020). A study of availability and recovery of URLs in Library and Information Science scholarly journals. *Annals of Library and Information Studies, 67*(1), 1–12. https://doi.org/10.51983/AJIST-2020.10.1.297

[9] Niveditha, B., Kumbar, M. (2020). Accessibility and characteristics of web citations in Journal of Computer-Mediated Communication during 2008-2017. *Journal of Indian Library Association*, *56*(2), 39-50.

[10] Sife, A. S., Lwoga, E. T. (2017). Retrieving vanished web references in health science journals in East Africa. *Information and Learning Sciences, 118*(9–10), 499–513. https://doi.org/10.1108/ILS-04-2017-0030.

[11] Shanthakumari, K., Keshava. (2024). Web citation analysis and efficiency of Internet archives in a selected civil engineering journal: A case study. *Journal of Science and Technology Metrics, 5*(2), 45-54.

[12] Barphe, V. U. (2023). Web citation and decay of URLs: A case study of Library and Information Science thesis uploading in Shodhganga under Savitribai Phule Pune University, Pune. *Annals of Library and Information Studies, 59*(3).

[13] Sadatmoosavi, A., Khasseh, A. A., Tajedini, O. (2026). Link rot in LIS literature: a 20-year study of web citation decay, recovery and preservation challenges. *Aslib Journal of Information Management*, 1-14.

[14] Sife, A. S., Bernard, R. (2013). Persistence and decay of URL citations used in theses and dissertations available at the Sokoine National Agricultural Library, Tanzania. *International Journal of Education and Development Using ICT, 9*(2), 1–13.

[15] Lakic, V., Rossetto, L., Bernstein, A. (2023, January). Link-rot in web-sourced multimedia datasets. *In International conference on multimedia modeling* (p. 476-488). *Cham: Springer International Publishing*.

[16] Lemaire, B., Bauer, F., Chaves Rodriguez, E., Ghesquiere, J., Radziejwoski, A., Roth, A. (2025). Web references are not eternal: Time trend and qualitative impact of the loss of access to online resources cited in peer-reviewed medical journals. *Current Medical Research and Opinion, 41*(3), 543–548.

[17] Saberi, M. K., Abedi, H. (2012). Accessibility and decay of web citations in five open access ISI journals. *Internet Research*, *22*(2), 234-247.

[18] Sadat Moosavi, A., Isfandyari Moghaddam, A., Tajeddini, O. (2012). Accessibility of online resources cited in scholarly LIS journals: A study of Emerald ISI ranked journals. *Aslib Proceedings*, *64*(2), 178-192.

[19] John, H. C., Simisaye, A. O., Iseyemi, T. J. (2024). Missing and recovery of URLs using Internet Archive: A case study on African Journal of Library, Archives and Information Science (AJLAIS). *International Journal of Information Science and Management (IJISM)*, *22*(2), 193-210.

[20] Loan, F. A., Shah, U. Y. (2020). The decay and persistence of web references. *Digital Library Perspectives, 36*(3), 261–276. https://doi.org/10.1108/DLP-02-2020-0013.

[21] Loan, F. A., Khan, A. M., Andrabi, S. A. A., Sozia, R., Parray, U. Y. (2024). Giving life to dead: Role of WayBack Machine in recovery of dead URLs. *Data Technologies and Applications, 58*(2), 201–213.

[22] Singh, T. G., Devi, K. S. (2024). Web decay analysis and digital archiving of websites of technical institutions: a view from Wayback Machine. *College Libraries*, *39*(I), 1-10.

[23] Ahmad, K., Sheikh, A., Rafi, M. (2020). Scholarly research in Library and Information Science: an analysis based on ISI Web of Science. *Performance Measurement and Metrics*, *21*(1), 18-32.

[24] Chen, C., Luo, B., Chiu, K., Zhao, R., Wang, P. (2014). The preferences of authors of Chinese library and information science journal articles in citing Internet sources. *Library information science research*, *36*(3-4), 163-170.

[25] Chen, C., Wang, P., Liu, Y., Wu, G., Wang, P. (2013). Impacts of government website information on social sciences and humanities in China: A citation analysis. *Government Information Quarterly, 30*(4), 450-463.

[26] Dundannanavar, A., Hadagali, G. S. (2024). Web reference decay: Passing the HTTP errors instead of quality information. *AWARI*, *5*, 1-10 e-PG Pathshala. (2026). About e-PG Pathshala. Retreived from https://epgp.inflibnet.ac.in/Home.

[27] Maharana, B., Nayak, K., Sahu, N. K. (2006). Scholarly use of web resources in LIS research: a citation analysis. *Library Review*, *55*(9), 598-607.

[28] Kousha, K., Thelwall, M. (2006). Motivations for URL citations to open access library and information science articles. *Scientometrics*, *68*(3), 501-517.

[29] Prithviraj, K. R., Kumar, B. S. (2015). Web citation trends in Indian LIS Journals: A citation analysis. *COLLNET Journal of scientometrics and Information Management, 9*(2), 295-310.

[30] Sampath Kumar, B. T., Prithvi Raj, K. R. (2012). Availability and persistence of URL citations in Indian LIS literature. *The Electronic Library*, *30*(1), 19-32.

[31] Sampath Kumar, B. T., Prithviraj, K. R. (2015). Bringing life to dead: Role of Wayback Machine in retrieving vanished URLs. *Journal of Information Science, 41*(1), 1–7. https://doi.org/10.1177/0165551514552752.

[32] Kumar, B. S., Kumar, D. V. (2013). HTTP 404-page (not) found: Recovery of decayed URL citations. *Journal of Informetrics, 7* (1), 145-157. DOI: 10.1016/j.joi.2012.09.007.

[33] Kumar, V., Kumar, S. (2019). Recouping the missing URL citations in Library Hi-Tech journal. *Journal of Indian Library Association, 55*(4), 1-8.

[34] Bansal, S. (2021). Decay of URL References cited in DESIDOC Journal of Library & Information Technology. *Library Philosophy Practice*.

[35] Maemura, E. (2023). Sorting URLs out: Seeing the web through infrastructural inversion of archival crawling. *Internet Histories, 7*(3), 1–20. https://doi.org/10.1080/24701475.2023.2258697.

[36] Kumar, D. V., Sampath Kumar, B. T. (2017). Recovery of vanished URLs: Comparing the efficiency of Internet Archive and Google. *Malaysian Journal of Library Information Science, 22*(2), 1–15. https://doi.org/10.22452/MJLIS.vol22no2.3.

[37] Niveditha, B., Kumbar, M. (2020). Accessibility and characteristics of web citations in Journal of Computer-Mediated Communication during 2008-2017. *Journal of Indian Library Association*, *56*(2), 39-50.