Web Decay Analysis and Digital Archiving of Websites of Technical Institutions: a view from Wayback Machine

Tonjam Govind Singh

Research Scholar, Department of Library and Information Science, Manipur University

Dr. K. Sangeeta Devi

Assistant Professor, Department of Library and Information Science, Manipur University

Abstract

Web Decay analysis is essential for evaluating website management and conducting regular link audits. This research is centred on evaluating the efficacy of the Wayback Machine in preserving websites belonging to Technical Institutions in the Northeastern Region of India (NEI). The specific objectives encompass comprehending archiving frequency, conducting quantitative and qualitative analyses of web links to compare features, and ranking technical institutions based on their utilisation of the Wayback machine. In this investigation, all technical institutions' websites in North East India were virtually surveyed using web browsers. A total of thirteen (13) technical institutions in the Northeastern Region of India were included in the study. The data obtained from the Wayback Machine (Internet Archive, n.d) pertain to the URLs of the technical institutions in North East India from 2013 to 2023. The analysis was performed using MS Excel 2019 and SPSS software to statistically compare the institutions' websites, revealing that the Wayback machine automatically archives all technical institution websites in NEI indiscriminately. It has proven to be a powerful tool for retrieving domain-specific information and understanding the paths to website demise. The study underscores the utility of the Wayback Machine in examining archival content, formats, and links and highlights the inconsistencies in capturing different files. It is revealed that there is a significant variation in the frequency of archiving the websites among technical institutions in NEI. The findings emphasise the need for an independent mechanism to archive websites, ensuring the retrieval of dead links for future reference.

Keywords: Digital archiving, Digital resources and services, North East India, Technical institutions, Wayback machine, Web archiving

1. Introduction

The internet's evolution and enhanced connectivity have created a global community, facilitated by mobile technologies and high-speed wireless Internet. In India, public internet access began in 1995, reaching 833 million users by July 2022. Websites play a crucial role in information delivery, necessitating streamlined design and content for user engagement. The World Wide Web, commonly referred to as the Web, serves as a vital gateway for accessing information. While exploring websites, readers often encounter what is colloquially known as a 'dead link'. These broken links occur for

various reasons, such as pointing to non-existent or inaccurate web pages. This phenomenon, termed 'Web Decay' or 'link rot', occurs when hyperlinks on individual websites or across the internet lead to permanently unavailable web pages, servers, or other resources (Kumar & Prithviraj, 2014). One prevalent cause of broken links is the disappearance of the target web page. When a previously accessible page no longer exists, users encounter a frustrating '404 error,' indicating that the web server responded but the specific content could not be found. Another type of dead link arises when the server hosting the target page ceases to function or relocates to a new domain name. Research reveals that the average lifespan of a random web page is approximately two years. When a dead web link appears within a scientific article, it carries broader implications. Fortunately, several tools are available to address these issues (Bar-Yossef, Broder, Kumar, & Tomkins, 2004). The Wayback Machine, maintained by the Internet Archive, is a crucial tool for accessing archived web content, offering a journey through digital history with over 370 billion webpages dating back to 1996.

This study focuses on the Wayback Machine's role in archiving websites of prestigious educational institutions in North East India such as Indian Institute of Technology (IIT), National Institute of Technology (NIT), Indian Institute of Management (IIM), Central Institute of Technology (CIT), Don Bosco Institute of Management (DBIM), and Manipur Institute of Technology (MIT) from 2013 to 2023.

2. Review of related literature

Several recent studies have delved into the content analysis of library websites, a well-documented area. Ramalingam (2013) found that Indian Institute of Technologies (IITs) surpassed others in its web page offerings. Pfanzelter (2020) noted that the Internet Archive's Wayback Machine remains the primary access point for obtaining historical screenshots from websites and serves as a valuable mechanism for defending intellectual property rights. Loan et al. (2023) conducted a comparative study on the efficiency of Chrome, Google, and WayBack Machine in retrieving dead links, concluding that the Wayback Machine was more efficient and effective in preserving web sources. Arora et al. (2016) highlighted challenges in scaling up the acquisition of website data from the Wayback Machine, emphasising the need for automated and manual efforts to create high-quality datasets. Additionally, Kanhabua (2016) proposed an entity-oriented search system using Bing to retrieve ranked live web results linked to the Wayback Machine, while Manohar, Banu and Punithavathani (2022) emphasised the importance of web ranking for measuring the quality and reliability of web pages.

The Internet Archive's Wayback Machine is widely utilised to estimate 'death dates' and track citations per year for Scientific Data Analysis Resources (SDAR) (Wren, et al., 2017). Studies have shown that while URL decay is a persistent phenomenon, more recognised SDARs tend to persist longer. Additionally, the Wayback Machine serves as a powerful tool for retrieving information on domain history and understanding website death paths (Kumar, 2023). It has proven beneficial in recovering vanishing online citations (Kumar, Kumar, & Prithviraj, 2015), offering researchers, authors, publishers, and editorial staff a valuable resource. Moreover, compared to search engines like Google, the Internet Archive is found to be more efficient in recovering vanished URLs (Tajedini et al., 2011). Recommendations include proactive measures to prevent citation decay and

enhance web accessibility, underscoring the importance of improved guidelines and policies for authors, editors, publishers, librarians, and web designers (Pearce & Charlton, 2009).

3. Significance of the study

In today's dynamic Information and communication technology landscape, websites of technical institutions hold immense importance. However, encountering dead links while browsing is common, often due to the disappearance of target web pages. The Wayback Machine emerges as a crucial tool to address such challenges, offering invaluable assistance to researchers, librarians, and those interested in the evolution of the web. This study focuses on analysing the role of the Wayback Machine in archiving the websites of thirteen (13) technical institutions in the north-eastern region of India, recognising the necessity to understand the archival process amidst the continuous evolution of URLs and website contents.

4. Objectives

The main objectives of the study are:

i. to outline a general overview of the technical institution's webpage

ii. to study the application of the Wayback Machine in the websites of different institutions in NEI

iii. to understand the frequency of archiving the website of the technical institution in NE India

iv. to compare the features of the archiving of the websites of the technical institutions in NEI during 2013 to 2023

v. to evaluate the quantity and quality of web links archived in the selected technical institutions

vi. to rank the technical institution in NEI with respect to the usage of Wayback Machine in solving the problem of dead links

vii. to make suggestions for improving the libraries' webpages/websites.

5. Scope and methodology of the study

The present study includes the web decay analysis and digital archiving of the websites of thirteen (13) technical institutions of North East India. They are National Institute of Technologies (NITs) in north-eastern India, Indian Institute of Technology (IIT) Guwahati, Indian Institute of Management Shillong(IIMS), Central Institute of Technology Kokhrajhar (CITK), Manipur Technical University (MTU) Manipur, and Don Bosco Institute of Management Guwahati (DBIMG). Out of these, ten (10) are Institutes of National Importance, and three (3) are All India Council for Technical Education (AICTE)-approved institutions. The websites of all the technical institutions of North East India have been surveyed virtually using Microsoft Edge web browser under Microsoft Windows 11 operating system. The data used for the study are also taken through the Wayback Machine (Internet Archive) concerning the URLs of the technical institution in North East India under study. In this regard, the data collected during 2013 - 2023 is analysed using MS Excel 2019 and SPSS software for statistical comparison of the Institution's websites.

6. Analysis of data and interpretation

The data analysis and interpretation of the study is discussed in the following sections.

6.1 *Technical institutions under study*

Table 1: Technical institutions under study **N = 13**

General Information of the Institutions				
Name of the institution	**State**	**Year of establishment**	**Ownership**	**Name of the library**
NITAP	Arunachal Pradesh	2010	Central	Central Library
NITSIL	Assam	2002	Central	Central Library
CITK	Assam	2006	Central	Central Library
IITG	Assam	(1994)	Central	Lakshminath Bezbaroa Central Library
DBIMG	Assam	2009	Private	Julhash Ali Library
NITMA	Manipur	2010	Central	Central Library
MTU	Manipur	2016	State	Manipur Technical University Library
NITM	Meghalaya	2010	Central	Central Library
IIMS	Meghalaya	2007	Central	Knowledge Centre
NITMZ	Mizoram	2009	Central	Central Library
NITN	Nagaland	2010	Central	Central Library
NITS	Sikkim	2011	Central	Knowledge and Information Centre
NITA	Tripura	2006	Central	Knowledge Centre

(Source: Websites of the technical institutions under study.)

Note: NITAP (NIT Arunachal Pradesh) NITSIL (NIT Silchar), NITMA (NIT Manipur), NITM (NIT Meghalaya), NITMZ (NIT Mizoram), NITN (NIT Nagaland), NITS (NIT Sikkim), NITA (NIT Agartala).

Table 1 summarises key aspects of the technical institutions in North East India under examination. The table shows a range of establishment years, with IITG being among the oldest (established in 1994) and MTU being the most recent (established in 2016). Assam hosts the highest number of institutions (four), followed by Manipur and Meghalaya, each with two. Other states such as Arunachal Pradesh, Mizoram, Nagaland, Sikkim, and Tripura have one NIT each. Ownership-wise, the majority of the institutions (11 out of 13) are centrally sponsored, with one state-owned and one private institution. Interestingly, seven institutions consider their library as the central library, while others have distinct names for their libraries. The Wayback Machine search indicates automatic archiving of all technical institution websites.

6.2 *Frequency of archiving*

Table 2 displays the count of web captures by the Wayback Machine of technical institutions (TI) of North East India (NEI) under distinct domain names. The present analysis reveals that a total of about 16,519 items from the websites were saved by all the technical institutions during the period from 2013 to 2023.

Table 2: Frequency of archiving the websites of the technical institutions of NEI N=13

Year	Institute Name	URL of the Institution	No. of Times Saved the Web	No. of Web Capturing
2013 -23	NITAP	https://www.nitap.ac.in/	351	7028
2013 - 23	NITSIL	https://www.nits.ac.in/	598	41278
2013 - 23	CITK	https://www.cit.ac.in/	1907	17934
2013 - 23	IITG	https://www.iitg.ac.in/	4781	313095
2013 - 23	DBIMG	https://dbim.ac.in/	146	1427
2013 - 23	NITMA	https://www.nitmanipur.ac.in/	1645	19317
2013 - 23	MTU	http://mtu.ac.in/	1212	15973
2013 - 23	NITM	https://www.nitm.ac.in/	469	12123
2016 - 23	IIMS	https://www.iimshillong.ac.in/	2284	20893
2013 - 23	NITMZ	https://www.nitmz.ac.in/	623	14860
2013 - 23	NITN	https://www.nitnagaland.ac.in/	501	11201
2013 - 23	NITS	https://nitsikkim.ac.in/	913	14779
2013 - 23	NITA	https://www.nita.ac.in/	480	12670
		Total	16519	567658

(Source: Websites of the technical institutions under study and Wayback Machine)

It is found that IITG scored the highest in capturing the webpage with 3,13,095 times, followed by NITSIL (41,278), IIMS (20,893), NITMA (19,317), CITK (17,934), MTU (15,973), NITMZ (14,860), NITS (14,779), NITA (12,670), NITM (12,123), NITN (11,201), NITAP (7,028), and DBIMG the least at 1,427 score, under a unique domain name from 2013 to 2023. Wayback Machine effectively retrieved dead path URLs for all institutions, demonstrating its effectiveness in capturing web data for NEI's technical institutions.

6.3 *Frequency of web capturing in technical institution*

Table 3 and figure 1 present the duration (in days) of the current websites of the technical institutions as captured by the Wayback Machine from 2013 to 2023.

Table 3: Number of day's web capturing during 2013 - 2023 **N=13**

Year	NITAP	NITSIL	CITK	IITG	DBIMG	NITMA	MTU	NITM	IIMS	NITMZ	NITN	NITS	NITA
2013	-	21	3	27	2	3	60	5	-	11	14	15	3
2014	2	27	7	24	6	10	19	8	-	20	14	19	8
2015	2	41	11	41	6	13	11	8	-	18	13	16	10
2016	6	78	40	60	13	36	4	25	32	48	25	30	27
2017	17	91	56	85	27	66	28	37	226	75	42	51	43
2018	6	126	48	126	34	69	21	16	211	69	18	70	34
2019	9	103	31	190	23	52	29	20	23	52	21	49	27
2020	6	35	72	358	7	24	17	21	16	18	16	23	21
2021	8	28	75	358	9	14	14	26	23	20	22	26	29
2022	14	21	302	352	11	296	297	39	32	26	21	24	33
2023	140	25	285	313	3	283	214	151	142	143	145	145	41
Total	210	596	930	1934	141	866	714	356	705	500	351	468	276
Average	19	54	84	175	12	78	64	32	64	45	31	42	25

(Source: Wayback Machine)

Some of the technical institutions like IIT Guwahati, CIT Kokrajhar, NIT, and Manipur consistently archived their institution's websites during the study. It is observed that there was a significant increase in the year 2022 and 2023. The duration of web capturing of the remaining technical institutions in NEI is highlighted in table 3. The present analysis reveals a lack of uniformity in Wayback machine archiving for all the technical institutions in NEI, except for IIMG.

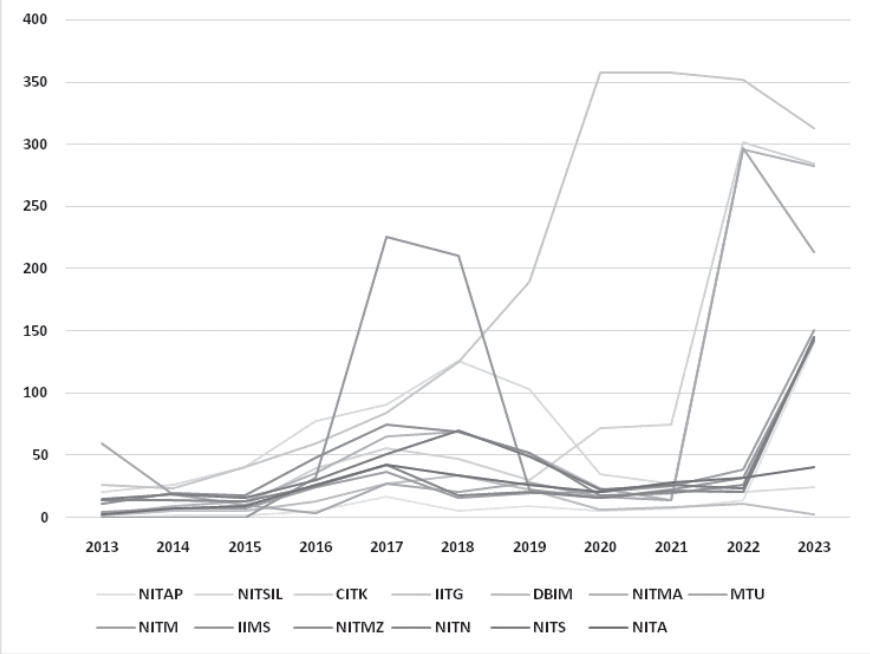

Figure 1: Trend in capturing website

6.4. *Frequency of capturing URLs in the website of technical institutions in NEI*

Table 4: Number of URLs captured in the websites of the technical institutions in NEI

Year	NITAP	NITSIL	CITK	IITG	DBIM	NITMA	MTU	NITM	IIMS	NITMZ	NITN	NITS	NITA
2013	-	773	150	4208	22	184	1582	166	-	121	106	103	2
2014	12	1290	154	6371	19	368	844	264	-	249	192	278	367
2015	67	1112	426	5725	20	185	19	24	-	464	128	403	402
2016	361	1195	225	4122	6	642	1	-	112	584	235	559	292
2017	459	1362	720	6487	54	853	389	2	104	742	955	903	644
2018	483	1597	225	11073	59	621	393	115	87	765	230	391	1091
2019	278	2194	414	12512	-	478	734	335	306	418	121	728	387
2020	4	1237	540	10512	34	761	538	641	1156	229	30	297	301
2021	237	1973	791	11837	37	936	577	875	2098	596	619	847	918
2022	445	1717	1084	19560	212	2393	2696	2999	2639	1219	1027	428	670
2023	1520	750	322	5134	39	463	486	1156	1186	225	735	238	291
Total	3866	15200	5051	97541	502	7884	8259	6577	7688	5612	4378	5175	5365
Average	351	1381	459	8867	45	716	750	597	698	510	398	470	487

Single factor ANOVA						
Source of Variation	**SS**	**df**	**MS**	**F**	**P-value**	**F crit**
Between Institution	710919410.4	12	59243284.2	29.7159611	1.44757E-31	1.827349397
Within Institution	259174755.3	130	1993651.964			
Total	970094165.7	142				

It is observed from table 4 that there is a variation in capturing URLs (links) by the different technical institutions of NEI. Only a few institutions, like IIT Guwahati, NIT Silchar, Manipur Technical University (MTU), and NIT Manipur, have the maximum frequency of capturing URL links. The study revealed an uneven trend in capturing links in the websites of technical institutions in NEI from 2013 to 2023. By using the single factor ANOVA test, the present analysis also revealed that the F value of 29.71 is greater than the F critical value of 1.82, which means that the test is significant. The p-value of 1.44 is greater than the 0.05 significance level with 12 degrees of freedom. It shows a significant difference in the URLs captured in different technical institutions under study at a 0.05 significance level.

6.5 *Type of media captured and its frequency of capturing.*

The media file format and Media Type (MT) are categorised as Txt (Text), Img (Image), App (Application), Font, and AV (Audio and Video). The Wayback Machine archives the websites of technical institutions multiple times between 2013 and 2023, capturing both the media files and original URLs of these websites. Within the Wayback Machine, there are two types of URLs: "URLs" and "New URLs." During the archival process, "New URLs" represent the most recent links captured during the specified time frame, while "URLs" refer to previously saved links from before the designated period. Table 5 has highlighted an overview of the types of media captured in the websites of technical institutions in NEI from 2013 to 2023. Txt is predominantly captured in all technical institution websites except IIMS and NIT Arunachal Pradesh. Moreover, IIT, Guwahati is the only technical institution that scores maximum in capturing all types of media file formats (about 345491). NIT, Silchar and MTU follow it. IIMS, Shillong is the second highest in capturing the image file format. However, it is observed that the fonts and audio-visual file formats are not captured in almost all the technical institutions.

Table 5: Number of different file formats captured in the website of the technical institutions in NEI N=13

MT	NITAP	NITSIL	CITK	IITG	DBIM	NITMA	MTU	NITM	IIMS	NITMZ	NITN	NITS	NITA
Txt	1705	24353	11928	254130	1515	8805	11712	5311	6207	7058	6847	9934	7848
Img	3410	6914	2920	62547	211	3231	5040	2439	10900	2870	1220	1780	3011
App	1502	16463	2997	28751	292	7120	5726	3733	3699	5063	2977	3170	1798
Font	156	0	0	21	1	0	0	246	0	0	307	18	0
AV	0	5	9	138	0	0	0	2	87	0	0	0	0
Total	6773	47735	17854	345491	2019	19156	22478	11731	20893	14991	11351	14902	12657
Avg	615	4339	1623	31408	183	1741	2043	1066	1899	1362	1031	1354	1150

(Source: Wayback Machine)

Table 6: Ranking of technical institutions' websites based on the average number of capturing by Wayback Machine during 2013 - 2023

Capturing Websites	NITAP	NITSIL	CITK	IITG	DBIM	NITMA	MTU	NITM	IIMS	NITMZ	NITN	NITS	NITA
Average no of Days in a Year (Score out of 100)	19 (2.6)	54 (7.4)	84 (11.5)	175 (24.1)	12 (1.6)	78 (10.7)	64 (8.8)	32 (6.2)	64 (8.8)	45 (6.2)	31 (4.2)	42 (5.7)	25 (3.4)
Average number of Links Captured in a Year (Score out of 100)	644 (2.2)	2533 (8.7)	841 (2.9)	16256 (56.3)	83 (0.2)	1314 (4.5)	1376 (4.7)	1096 (3.7)	1281 (4.4)	935 (3.2)	729 (2.5)	862 (2.9)	894 (3.0)
Average number of File Formats in a Year (Score out of 100)	615 (1.2)	4339 (8.7)	1623 (3.2)	31408 (63)	183 (0.3)	1741 (4.1)	2043 (2.1)	1066 (2.1)	1899 (3.8)	1362 (2.7)	1031 (2.0)	1354 (2.7)	1150 (2.3)
Total Score (out of 300)	6	24.9	17.7	143.5	2.3	18.8	17.6	10.3	17	12.1	8.8	11.4	8.8

6.6 *Ranking based on the number of archived links*

Table 6 displays the website ranking based on the number of archived links in the Wayback Machine. The ranking of technical institutions by content capture reveals IITG at the top (with 143.5), followed by NITSIL (with 24.9), NITMA (with 18.8), and others in descending order. The analysis underscores that the Wayback Machine is a potent tool for addressing web decay in various North East India (NEI) technical institutions.

7. Discussion

The findings of the study shed light on various aspects of web archiving and the digital footprint of technical institutions in Northeast India. Notably, the establishment years of these institutions, ranging from IIT Guwahati (1994) to Manipur Technical University (2016), illustrate the evolution of technical education in the region over time. The disparity in web capturing frequency and duration among institutions, exemplified by IIT Guwahati's high capturing frequency compared to DBIMG's significantly lower one, underscores the varied approaches to web presence maintenance. Additionally, the uneven trend in capturing links highlights the complexity of web content management across institutions, supported by the significant differences in URL capturing revealed by the ANOVA test. Furthermore, the study's insight into the diversity of website content and structure underscores the need for tailored archiving strategies. While IIT Guwahati emerges as a leader in web archiving, the ranking system demonstrates variations in institutions' digital preservation efforts, emphasising the need for standardized practices and increased awareness regarding web archiving's importance in preserving institutional history and digital heritage.

8. Recommendations

In light of the challenges faced by libraries and information centres in the digital era, the study strongly recommends the integration of all intellectual content, information, and knowledge sources into local Institutional Digital Repositories (IDRs) within the respective institutions. This approach reduces dependency on parent websites and ensures the preservation of valuable resources. The study emphasises the need for a distinct archiving procedure for library websites, separate from the main institutional webpage, to enhance maintenance and recovery efforts. Furthermore, the Wayback Machine is identified as a powerful tool for retrieving archival information and understanding inactive paths within institutional websites. However, the study underscores the lack of uniformity and significant variations in web capturing methodologies and file formats among institutions, recommending each institution adopt customised archiving mechanisms to address dead links and preserve historical content effectively.

9. Conclusion

In conclusion, while issues surrounding broken links and web decay have not received as much scrutiny as other web-related topics, their significance is growing. Webmasters can mitigate these issues by conducting regular audits and crafting content to withstand obsolescence. This study sheds light on web archiving practices in technical institutions of Northeast India over a decade, utilising the Wayback Machine for analysis. The findings highlight an increase in web capturing duration, improving website management. This study aims to inform the scientific community, identify research gaps, and guide future endeavours. It is expected to assist policymakers in managing digital resources effectively and promoting standardisation, contributing to ongoing webometric research trends and shaping future agendas.

References

Arora, S. K., Li, Y., Youtie, J., & Shapira, P. (2015). Using the Wayback Machine to mine websites in the Social Sciences: a methodological resource. *Journal of the Association for Information Science and Technology,* 67(8), 1904-1915. doi:10.1002/asi.23503

Brewster K. (2005). Universal access to all knowledge: Brewster kahle: free download, borrow, and streaming. Retrieved from https://archive.org/details/SDForumBK Internet Archive. Retrieved from https://archive.org/web/

Jayanetti, H. R., Garg, K., Alam, S., Nelson, M. L., & Weigle, M. C. (2022). Robots still outnumber humans in web archives, but less than before. *Linking Theory and Practice of Digital Libraries,* 245-259. doi:10.1007/978-3-031-16802-4_19

Jonathan D., Wren, C.G., Cory B. & Giles, J.H. (2017). Use it or lose it: citations predict the continued online availability of published bioinformatics resources. *Nucleic Acids Research,* 45(7), 3627-3633. https://doi.org/10.1093/nar/gkx182

Kanhabua, N., Kemkes, P., Nejdl, W., Nguyen, T. N., Reis, F., & Tran, N. K. (2016). How to search the internet archive without indexing it. *Research and Advanced Technology for Digital Libraries,* 147-160. doi:10.1007/978-3-319-43997-6_12

Kumar, B.T.S., Vinay Kumar, D. & Prithviraj, K. R. (2015). Wayback Machine: reincarnation to vanished online citations. *Program: electronic library and information systems,* 49(2), 205-223. https://doi.org/10.108/PROG-07-2013-0039

Kumar, D. V., & Sampath, K. B. T. (2017). Recovery of vanished URLs: comparing the efficiency of Internet Archive and Google. *Malaysian Journal of Library and Information Science,* 22(2), 31-43. https://doi.org/10.22452/mjlis.vol22no2.3

Kumar, P. K. S. (2023) Web decay: analysis of the websites of the universities in Kerala using Wayback Machine. *Kelpro Bulletin,* 27(1), 27-36.

Loan, F. A., Khan, A. M., Andrabi, S. A. A., Sozia, S. R., & Parray, U. Y. (2023). Giving life to dead: role of Wayback Machine in recovery of dead URLs. *Data Technologies and Applications.* https://doi.org/10.1108/DTA-06-2022-0242

Manohar, E., Banu, A. E. & Punithavathani, S. D. (2022). Composite analysis of web pages in adaptive environment through Modified Salp Swarm algorithm to rank the web pages. *Journal of Ambient Intelligence and Humanized Computing,* 13, 2585-2600. https://doi.org/10.1007/s12652-021-03033-y

Pearce, D., & Charlton, B. G. (2009). Plagiarism of online material may be proven using the Internet Archive Wayback Machine (archive.org). *Medical Hypotheses,* 73(6) 875-875. https://doi.org/10.1016/j.mehy.2009.07.049

Pfanzelter, E. (2020). Narrating history/histories with data from the Wayback Machine using the example of Holocaust websites. *Zeitgeschichte,* 47(4) 491. https://doi.org/10.14220/zsch.2020.47.4.491

Ramalingam, J. (2012). Web page analysis of Indian Institute of Technologies' (IITs) websites: a webometric study. I*nternational Journal on digital libraries,* 3(1), 55-65.

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), 71-81. https://doi.org/10.1177/0165551514552752

Tajedini, O., Azimi, A., Sadatmoosavi, A., & Sharif-Moghaddam, H. (2011). Death of web citations: a serious alarm for authors, *Malaysian Journal of Library and Information Science,* 16(3) 17 - 29.