URL Link Rot: Implications on the Scholarly Publishing

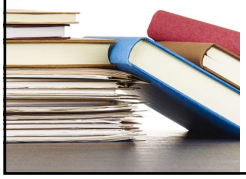
Dr B T Sampath Kumar
Professor
Department of Library and Information Science
Tumkur University, Tumakuru
www.sampathkumar.info

Internet: The 8th Wonder of the earth

- Network of networks.
- Information superhighway.
- Ocean of information.
- It is a global computer network providing a variety of information and communication facilities.
  - E-mail
  - Chatting
  - Video conferencing
  - Social networking sites

The Web

- Today, the web is most popular among researchers, students and faculty members for
  - Searching information
  - Downloading articles
  - Sharing research results
- The present day researchers have been increasingly citing Web sources (URLs) in their scholarly publications.



Use of URLs as citations

- Use of URLs as citations have become common in:
  - journal articles
  - conference papers
  - theses
  - dissertations
  - students projects/assignments.
- A plethora of literature in available on the use of web sources as citations in scholarly journal articles.

Why do we use URLs as citations?

- The popularity of Web sources among academic and scientific community.
- Most of the articles/books available online.
- Easy to find and access information.
- Available 24X7 and for free of cost.
- We have all the latest information at his fingertips.

Link Rot: The Problem

- In spite of its popularity, the URLs of websites have also few disadvantages/problems.
- Some URLs simply removed as their website domain names expire.
- In some other cases the web pages are removed from the Web by the webmaster.
- Some websites are designed with new file structures.
- Some websites may move to new online locations.

Problem with URLs

- The content may be removed from the Web because the information is no longer relevant or correct.
- Typographical errors associated while citing a URL.
- The server is made offline for maintenance.
- It causes a temporary inaccessibility of the web resource.

Problem with URLs

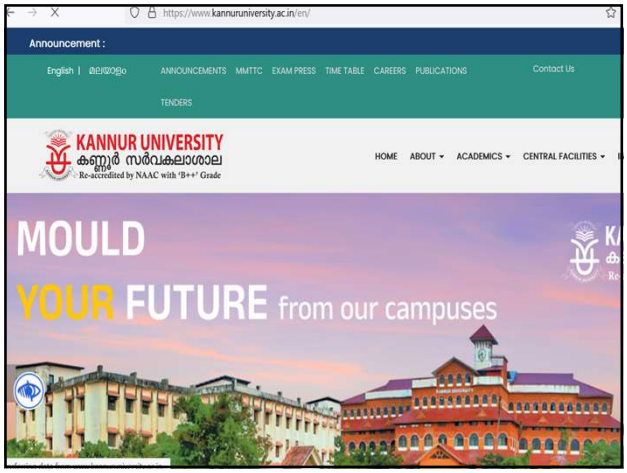
- Network problem at client side or at the server side.
- This disappearance of websites over a period of time is referred as web decay.

Link rot / Web decay

- Link rot/Web decay is the state or process of rotting or decomposition.
- Link rot or Web decay refers to a process whereby the URLs of the web source disappear over a period of time.

Accessible /Active URLs

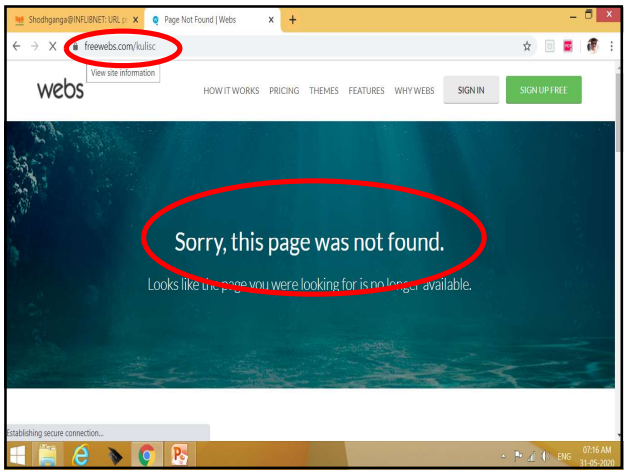
- URL of the web site is valid, and it can be accessible on the Internet using a web browser
- http://newhorizonindia.edu/



Cont..

Missing URLs/ Link rot

- A missing URL is one that returns with an HTTP error message when its URL is entered in Web browser/W3C Link Checker.
- Vanished websites/decayed/missing/rotted.
- Ex: www.freewebs.com/kulisc



Implications of Link Rot on Scholarly Publishing

- The researchers will not trust the web sources.
- The researchers will not the get the original information.
- Web sources will not be used as references in the scholarly communication.
- More impact on the open access.

How to check the availability of web site?

- Directly enter the URL of the source in a web browser.

  or
- Use of link checker.

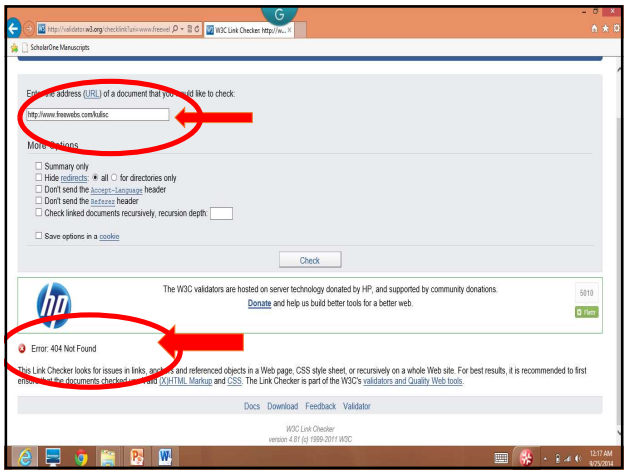
URL link checkers

These tool are used to verify for broken URL's on a website

Some of the URL link checkers are:

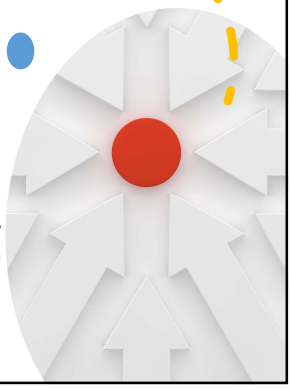
- W3C Link Checker
- Online Website Link Checker
- Dead link checker
- Web accessibility checker

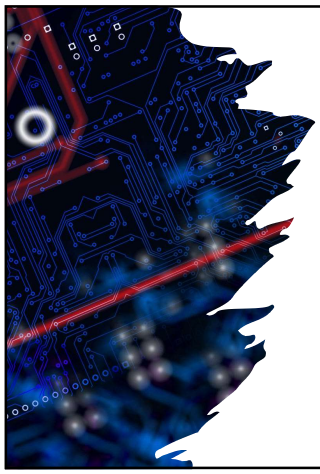
W3C link checker



HTTP error codes

- **HTTP 301 (Moved permanently):** This code indicates that the requested resource has been sent to a new URL.
- **HTTP-302** (Moved Temporarily): This code indicates that a new temporary location has been assigned to a resource.





HTTP error codes

- **HTTP-400 (**Bad Request): The request contains bad syntax, so that the server could not understand the request
- **HTTP-401** (Unauthorized): Few web resources require a username and password to process

HTTP error codes

- **HTTP-403** (Forbidden): This code indicates that the failure of the operation for the request due to the un-readability of a file or directory.
- **HTTP-404** (Page Not Found): This code indicates that the server has not found anything matching with the requested URL.

HTTP error codes

- **HTTP-415** (Unsupported Media Type): When the request is made for a particular media resource and the server finds it inappropriate
- **HTTP-503** (Service Unavailable): This code indicates that the server is temporarily unable to handle the request made by the client.
- Overload on the server causes the unavailability of the server's response.

Internet Archive

- **Internet Archive** is a non-profit library of millions of free books, movies, software, music, websites, and more

Web Archives	Country	Year	Main scope of archive content
Australia's Web Archive	Australia	1996	National
Government of Canada Web Archive (GCWA)	Canada	2005	National governmental
Internet Archive (Wayback machine)	USA	1996	International & service provider
Internet memory Foundation	France, Netherlands	2004	International & service provider
Japan Web Archiving Project	Japan	2004	National

Wayback machine

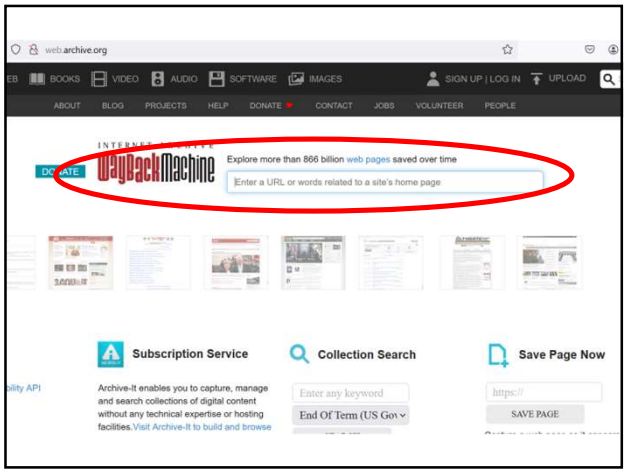
- It is a free online resource that was created in 1996
- It helps to build a digital library of webpages
- It offers permanent and free access to researchers, historians, scholars, and the general public

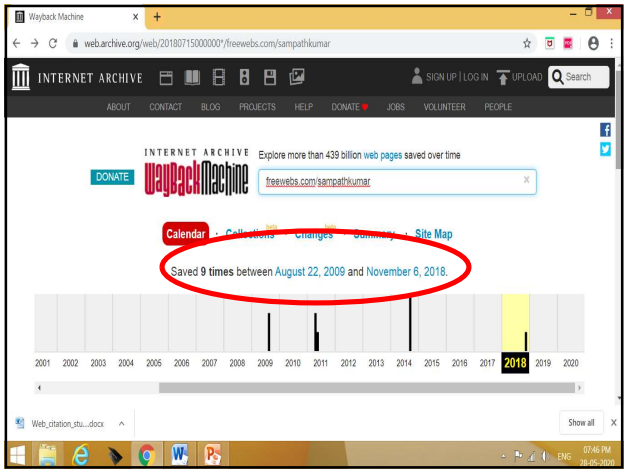
Wayback machine

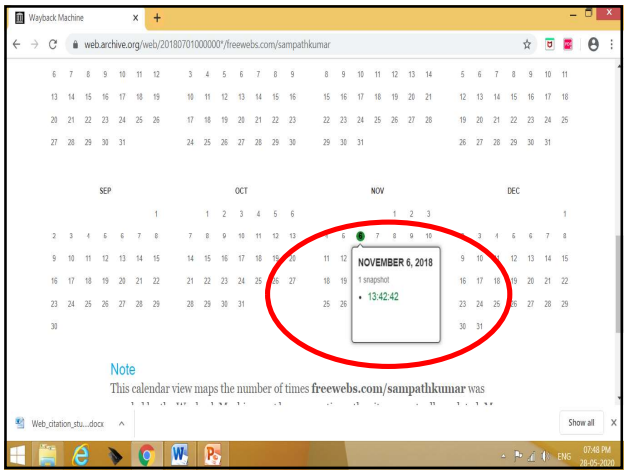
It contains:

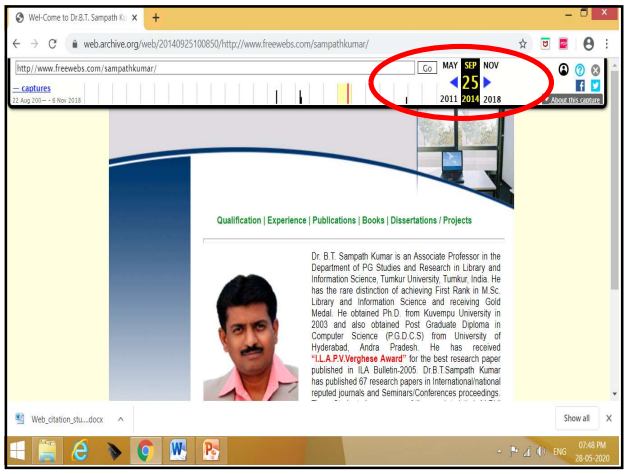
- 866 billion web pages
- 44 million books and texts
- 15 million audio recordings
- 10.6 million videos
- 4.8 million images
- 1 million software programs

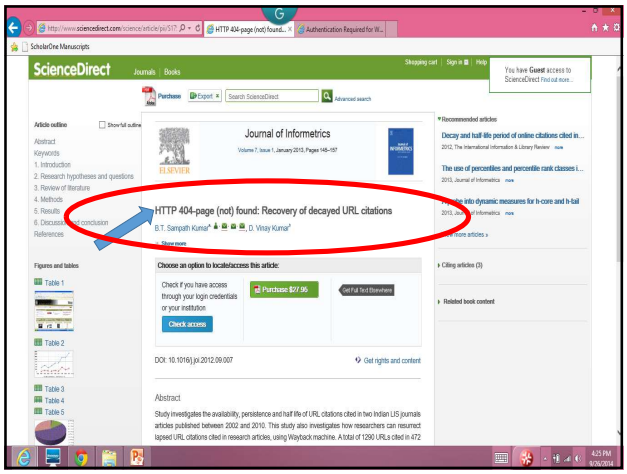
(I million =10 lakh, 1 billion =100 crores)

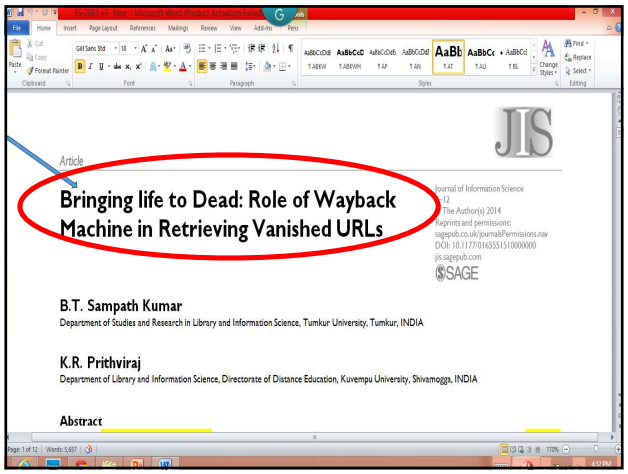


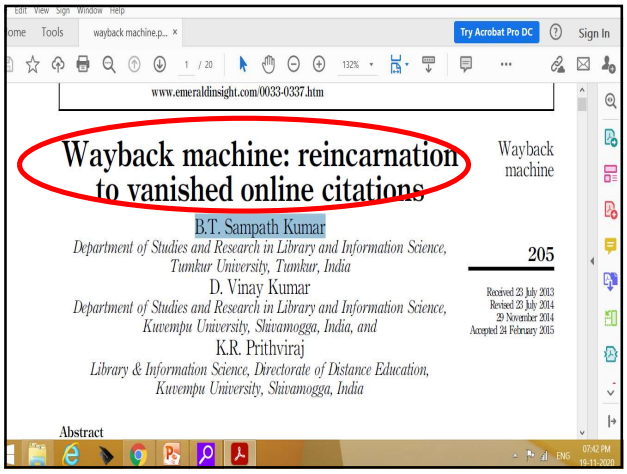


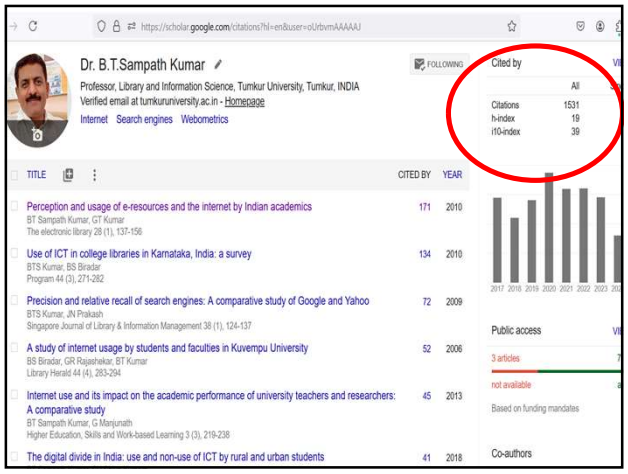


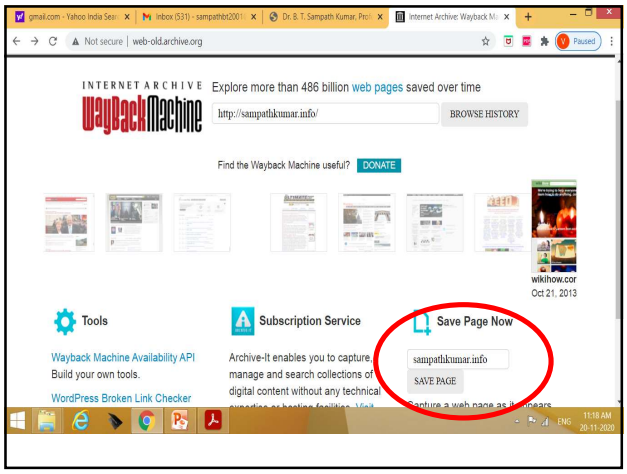


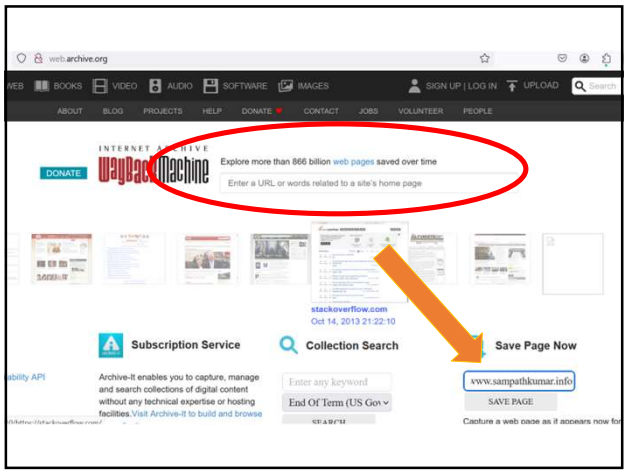

ScienceDirect
Journal of Informetrics
HTTP 404-page (not) found: Recovery of decayed URL citations
B.T. Sampath Kumar
D. Vinay Kumar
Abstract

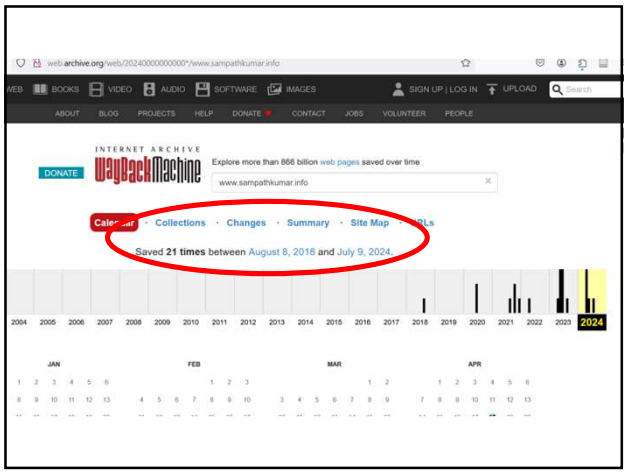

JIS
Article
Bringing life to Dead: Role of Wayback Machine in Retrieving Vanished URLs
B.T. Sampath Kumar
Department of Studies and Research in Library and Information Science, Tumkur University, Tumkur, INDIA
K.R. Prithviraj
Department of Library and Information Science, Directorate of Distance Education, Kuvempu University, Shivamogga, INDIA
Abstract


www.emeraldinsight.com/0033-0337.htm
Wayback machine: reincarnation to vanished online citations
Wayback machine
B.T. Sampath Kumar
Department of Studies and Research in Library and Information Science, Tumkur University, Tumkur, India
D. Vinay Kumar
Department of Studies and Research in Library and Information Science, Kuvempu University, Shivamogga, India, and
K.R. Prithviraj
Library & Information Science, Directorate of Distance Education, Kuvempu University, Shivamogga, India
205
Abstract


Dr. B.T.Sampath Kumar
Professor, Library and Information Science, Tumkur University, Tumkur, INDIA
Cited by
Citations 1531
h-index 19
i10-index 39
Perception and usage of e-resources and the internet by Indian academics
Use of ICT in college libraries in Karnataka, India: a survey
Precision and relative recall of search engines: A comparative study of Google and Yahoo
A study of internet usage by students and faculties in Kuvempu University
Internet use and its impact on the academic performance of university teachers and researchers: A comparative study
The digital divide in India: use and non-use of ICT by rural and urban students
Public access
Co-authors


INTERNET ARCHIVE
WaybackMachine
Explore more than 486 billion web pages saved over time
http://sampathkumar.info/
BROWSE HISTORY
Find the Wayback Machine useful?
DONATE
Tools
Wayback Machine Availability API
Build your own tools.
WordPress Broken Link Checker
Subscription Service
Archive-It enables you to capture, manage and search collections of digital content without any technical
Save Page Now
sampathkumar.info
SAVE PAGE


INTERNET ARCHIVE
WaybackMachine
DONATE
Explore more than 866 billion web pages saved over time
Enter a URL or words related to a site's home page
Subscription Service
Collection Search
Save Page Now
Archive-It enables you to capture, manage and search collections of digital content without any technical expertise or hosting facilities.
www.sampathkumar.info
SAVE PAGE



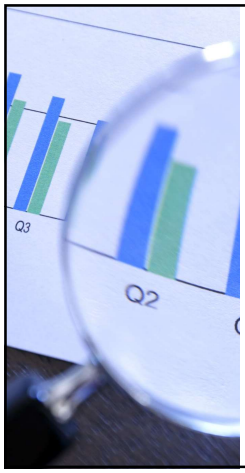
Pandora

- It is a national web archive for the preservation of Australia's online publications.
- Established by the National Library of Australia in 1996.
- It has been built in collaboration with Australian state libraries and cultural collecting organisations

What is the lesson?

- The citations to web content should include full bibliographic information.
- Authors should test web sources for their availability..
- Web master should not to use lengthy URLs while creating web sources.

- The authors need to cite the scholarly information from the authentic web documents.
- The need to check the accessibility status of URLs before citing them in their scholarly works.
- If a URL citations is not accessible, then, the author may use any Internet Archive to recover the web source.



Cont..

- It is suggested that editors or editorial staff of journals need to check the URL citations (URLs) cited in the articles submitted to them.
- The URL check should confirm the accessibility and this shall be one of the criteria for the acceptance of the article.

Role of LIS professionals

- Need to conduct orientation program among the faculty/researcher students on:
  - Citing web sources in scholarly literature.
  - Decay of web sources and its implications on scholarly content
  - Link checkers
  - Internet Archives
  - Recovery of decayed web sources
  - Archiving the web sources in Web archives

Conclusion

- Internet archive plays a significant role in archiving web site.
- However it covers only a portion of the web.
- It will not frequently monitor the changes in the content of the web page.
- We need to Save web page to archive manually.

NAMASTE

Feed back to
94483 20187
sampathkumar.info