404 not found: A study of hyperlinks on Irish academic library websites

AUTHOR(S)

Colm O'Connor, Aoife Doherty

An Leabharlann
29 – 1
March 2020

404 NOT FOUND: a study of hyperlinks on Irish academic library websites

Colm O'Connor and Aoife Doherty

Abstract

Link rot refers to inaccessibility of resources on the World Wide Web due to decayed links. Using a link checking tool the authors examined links from ten Irish academic library websites using the LibGuides content management system. Four percent of links were affected by link rot, rendering them inaccessible. It is recommended that librarians be mindful of link rot when creating content and periodically check their webpages for broken links.

Keywords: *Link Rot, Academic Library websites, Ireland*

Introduction and Literature Review

This paper examines the incidence of link rot in ten Irish academic library websites at a given point in time. The term "link rot", first coined by Jakob Nielsen (1998), refers to the inaccessibility of resources on the World Wide Web due to decayed links. Decayed links may be due to several causes including non-existent web pages, restructuring of a website's URL (Uniform Resource Locator) or the omission of a redirect when moving a web page.

The instability of resources on the World Wide Web is a great source of frustration for web users, making finding the original source of information difficult or, in some instances, impossible. Notwithstanding the issue of permanency of web resources, web citations are prevalent in the scholarly literature owing to their ease of access and speed at which they can be made available online. Reference rot (that is link rot within references) in scholarly publications across a variety of disciplines has been the subject of study of many research papers. Notable findings include those of a large scale study conducted by Klein et al (2014) discovering that 20-70% of references in multiple science, technology and medicine (STM) journals are affected by link rot[1] and of Burnhill et al (2015) revealing that 20% of articles in the same field 'contain at least one reference to a web resource that is rotten'.

Research carried out by Kaufmann and Campana (2019) found that 16% of links to resources chosen for the online delivery of a social studies curriculum suffered from link rot, calling into question the efficacy of online resources as a learning tool. White (2019) makes the point that this issue could easily apply to the many links to educational resources contained in library LibGuides, an argument supported by the results of our study.

Methodology

Ten Irish academic library websites were examined for link rot. All ten websites used Libguides software; LibGuides is a content management and information sharing system designed specifically for libraries. Each website was assessed for link rot using the link checking tool 'Dr. Link Check'. This tool is easy to use and requires no technical or programming knowledge. Inserting the URL of the library webpage being investigated into a search box returns a results list containing the total number of links checked, the number of broken links and their location.

Additional information given includes the type of errors found (eg.404 not found, Timeout, 403 forbidden etc.), a breakdown of link types (eg. <a href>, <img src> etc) and link scheme type (http:, https: etc.). To double-check the accuracy of the reported results 5% of the links identified as broken by Dr. Link Check were tested manually confirming the effectiveness of the tool. The entirety of some sites can be checked by inserting the URL of the LibGuides homepage into the link checking tool; others require the URL of each individual LibGuide to be entered.

Results

In total 47,308 hyperlinks were examined across ten websites. 1,954 of these links were broken, a total of 4.10%. The percentage of broken links on the websites examined ranged from 0.91% (23 broken links out of 2,541) to 9.8% (432 broken links out of 4,411). See table 1 for details.

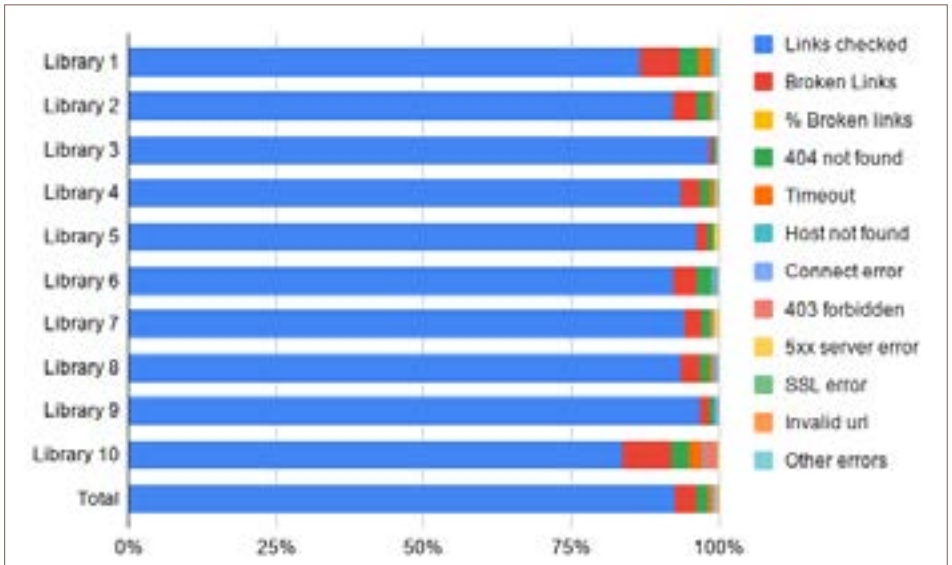
The most common type of error found was the familiar '404 not found' which

1 The study looked at multiple journals.

comprised 47.6% of all errors. Others were 'Timeout' (19.1%), '5XX Server Error' (9.1%), '403 Forbidden' (8.6%) and 'Host Not Found' (6.5%). Less frequent were 'Connect Error' (2.3%), 'SSL Error' (1.5%), 'Invalid URL' (0.8%) and 'Send/Receive Error' (0.1%). The remainder was made up of other errors.

Discussion, Solutions & Recommendations

Our findings demonstrate that link rot is a significant issue for Irish academic library websites. Of the links examined, one in every 25 was inaccessible. If this problem is not addressed this number is likely to get worse; Milligan (2019) notes that the average lifespan of a webpage is thought to be between 44-100 days. Several studies have shown that the incidence of link rot increases over time (Zeng et al 2019, Bansal & Parmar 2020, Zhou et al 2015).



Website	Links checked	Broken links	% broken links	404 not found	Timeout	Host not found	Connect error	403 forbidden	5xx server error	SSL error	Invalid url	Send/receive error	Other errors
Library 1	5676	449	7.91%	213	154	33	10	8	8	6	2	x	20
Library 2	4095	179	4.38%	106	18	5	2	7	5	3	1	1	25
Library 3	2341	23	0.98%	11	x	1	x	2	x	3	x	x	6
Library 4	4721	161	3.40%	84	23	33	1	3	13	3	1		3
Library 5	8002	149	1.90%	87	8	2		2	33	3	2		2
Library 6	2800	117	4.20%	73	6	18	3	5	6	1	x	1	6
Library 7	7576	249	3.18%	124	14	21	5	9	64	6	4		9
Library 8	3944	135	3.40%	66	27	17	3	13	3	1	x	x	14
Library 9	2592	43	1.66%	20	3	5	13		1	1	x	x	1
Library 10	4411	432	9.80%	157	117	12	9	115	15	2	x	x	5
Total	47308	1954	4.10%	931	121	177	45	169	177	36	15	2	85

Our findings demonstrate that link rot is a significant issue for Irish academic library websites. Of the links examined, one in every 25 was inaccessible. If this problem is not addressed this number is likely to get worse; Milligan (2019) notes that the average lifespan of a webpage is thought to be between 44-100 days. Several studies have shown that the incidence of link rot increases over time (Zeng et al 2019, Bansal & Parmar 2020, Zhou et al 2015).

This study checked a relatively high number of links compared to previous studies examining link rot. According to Sanderson et al (2011) "sample size has often been very small- …previously the number of URLs checked…was also limited by the manual nature of such checks". The authors overcame this problem by use of a link checking tool.

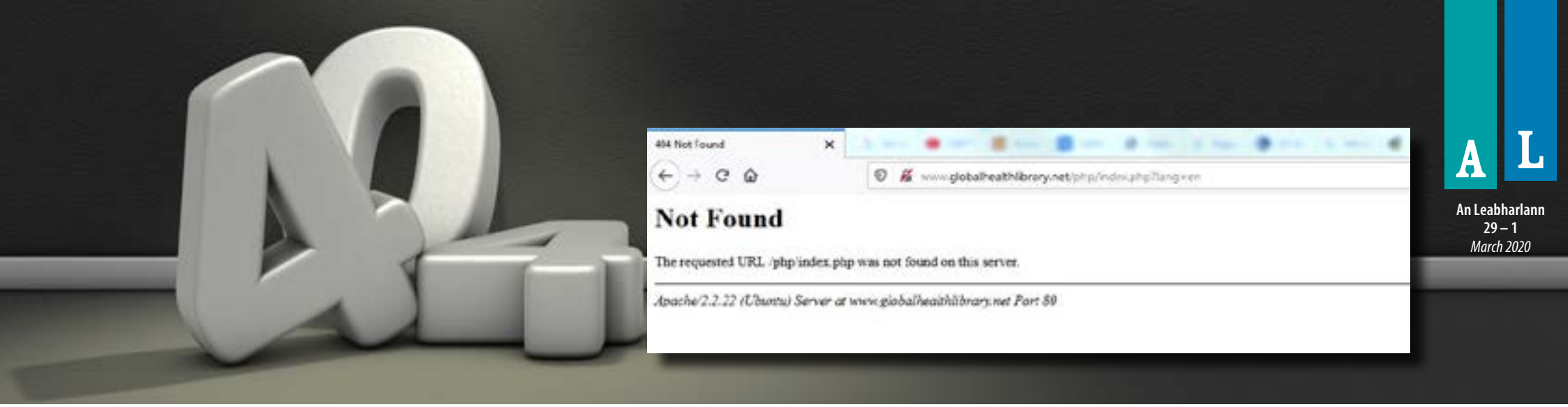
As aforementioned, this study recorded link rot at a given moment in time. While carrying out their own research into link rot, Tyler and McNeil (2003) found that a number of webpages recorded as 'dead' were discovered to be 'undead' upon revisiting the websites after a period of six weeks. Future studies may benefit from rechecking the original results to obtain a more exact picture of the issue.

What can librarians do to combat this problem? Fortunately there are a number of ways to deal with the issue of link rot. These can be divided into prevention and cure.

Firstly prevention. When creating content for their websites librarians should be mindful of link rot. Where possible, librarians should link to content which has a persistent identifier such as a DOI rather than a URL. DOIs are primarily used for scholarly journal articles rather than web pages however and the fact there is a cost to register a DOI means that most content will never have one. Another solution is to use Perma.cc. This is a service supported by a wide range of organisations including university libraries which takes a snapshot of a URL's content and returns a permanent link (known as a permalink) which is then used in place of the original link. Tim Berners Lee, the founder of the World Wide web believed that 'it is the duty of a webmaster to allocate URLs which you will be able to stand by in 2 years, in 20 years, in 200 years. This needs thought, and organisation, and commitment' (Berners-Lee, 1998).

Website content creators should avoid linking directly to documents such as PDFs. As noted by Kille (2015) these tend to be less stable than normal web pages - 'landing pages are generally more stable than PDFs. Because the latter

An Leabharlann
29 – 1
March 2020



are documents, they tend to be renamed or move around on websites. They can also be updated, potentially invalidating the reason for your original link, yet this won't necessarily be indicated to you or your users.'

Looking at the cures, one method of ensuring links on websites are kept up- to -date is by using a link checking tool. There are a number of such tools available, examples include Sitechecker, Interrobot, and the one used in this study, Dr. Link Check, as well as numerous others. All offer different features and most have a free version which usually allow a certain number of broken links to be checked before a purchase is required. Many of these tools have an automatic link checker which can be scheduled to run periodically; the broken links found can then be updated by a librarian. It is recommended that libraries examine the different features available through the various link checking tools and choose the one best suited to their individual needs. Checking for, and correcting, link rot should be considered a natural extension of usual website maintenance.

It should be noted that while link checking tools will detect broken links, they will not identify web pages that have been affected by content drift. Content drift refers to the evolution of an identified resource over time into one that bears no resemblance to the content originally referenced (Klein 2014). Content drift needs to be checked for manually.

Many Irish libraries, including all of those considered in this study, use the LibGuides software to build and edit their website. LibGuides includes an inbuilt link checking tool which many librarians may believe is sufficient to discover all instances of link rot. However, it is important to be aware that this tool has a number of limitations. As Randtke and Burell (2012) noted: 'Links in bulleted lists will be checked...other links will not'. This means that any links inserted within a paragraph (these comprise the vast majority of links in the authors study) or links inserted within an image will not be checked.

Conclusion

This study has shown that link rot is a problem in Irish academic libraries. When creating content for webpages, librarians should be mindful of the issue and link to webpages which are likely to be more stable (such as those with a .gov domain), avoid linking to documents such as PDFs and use permanent links such as those created by perma.cc. Librarians should also routinely check their webpages for link rot using a link checking tool and should incorporate link checking into their standard website maintenance.

Colm O'Connor MA MLIS , Assistant Librarian (Clinical Support)
and Aoife Doherty, BA, Library Assistant, Royal College of Surgeons in Ireland*

* Currently Post-graduate student Information and Library Management, Ulster University

References

- Bansal, S. & Parmar, S. (2020). "Decay of URLs Citation: A Case Study of Current Science". *Library Philosophy and Practice*, pp.1-9. Available from: https://digitalcommons.unl.edu/libphilprac/3582 [Accessed 3 January 2020].
- Berners-Lee, T. (1998). "Cool URIs don't change". Available from: https://www.w3.org/Provider/Style/URI [Accessed 22 December 2019]. Archived at: https://perma.cc/6RP8-S247
- Burnhill, P., Mewissen, M. & Wincewicz, R. (2015). Reference rot in scholarly statement: threat and remedy. *Insights: the UKSG Journal,* 28(2), pp.55–61. DOI: http://doi.org/10.1629/uksg.237
- Kaufman, R. & Campana, A. (2019). "OER: Lessons from the Field". *Insights: the UKSG Journal*, 32(1): pp.1-14. DOI: http://doi.org/10.1629/uksg.464
- Kille, L. W. (2015). The Growing Problem of Internet "Link Rot" and Best Practices for Media and Online Publishers". *Journalist's Resource, Harvard Kennedy School.* Available at: https://journalistsresource.org/studies/society/internet/website-linking-best-practices-media-online-publishers/ [Accessed 13 December 2019]. Archived at https://perma.cc/88TS-N3FR
- Klein M., Van de Sompel, H., Sanderson, R., Shankar, H., Balakireva, L., Zhou, K., et al. (2014). Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. *PloS One*, 9(12): e115253.
- Milligan, I. (2019). *History in the Age of Abundance?: How the Web Is Transforming Historical Research.* Montreal: McGill-Queen's University Press.
- Nielsen, J. (1998). Fighting Linkrot. *Nielson Norman Group.* Available from: https://www.nngroup.com/articles/fighting-linkrot/ [Accessed 20 December 2019].Archived at https://perma.cc/G663-RLCG
- Randtke, W. & Burrell, M.D. (2012). Tools for Reducing and Managing Link Rot in LibGuides. *Code4lib*,17 Available from: https://journal.code4lib.org/articles/7019 [Accessed 28 December 2020]. Archived at: https://perma.cc/WD5Y-RS37
- Sanderson R, Phillips M, Van de Sompel H (2011). Analyzing the Persistence of Referenced Web Resources with Memento. Technical Report arXiv: 1105.3459
- Tyler, D.C. & McNeil, B. (2003). Librarians and Link Rot: A Comparative Analysis with Some Methodological Considers. Available from: https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1067&context=libraryscience [Accessed 28 December 2019]. Archived at: https://perma.cc/DZJ9-BNDW
- White, J.M. (2019). Link Rot, Reference Rot and Link Resolvers. *In:* Varnum, K.J., ed. *New Top Technologies Every Librarian Needs to Know: A LITA Guide.* Chicago: ALA.#
- Zhou, K., Grover, C., Klein, M. & Tobin, R (2015). 'No More 404s: Predicting Referenced Link Rot in Scholarly Articles for Pro-Active Archiving', *Proceedings of the 15th ACM/IEEE-CE on Joint Conference on Digital Libraries,* 2015-June. https://doi.org/10.1145/2756406.