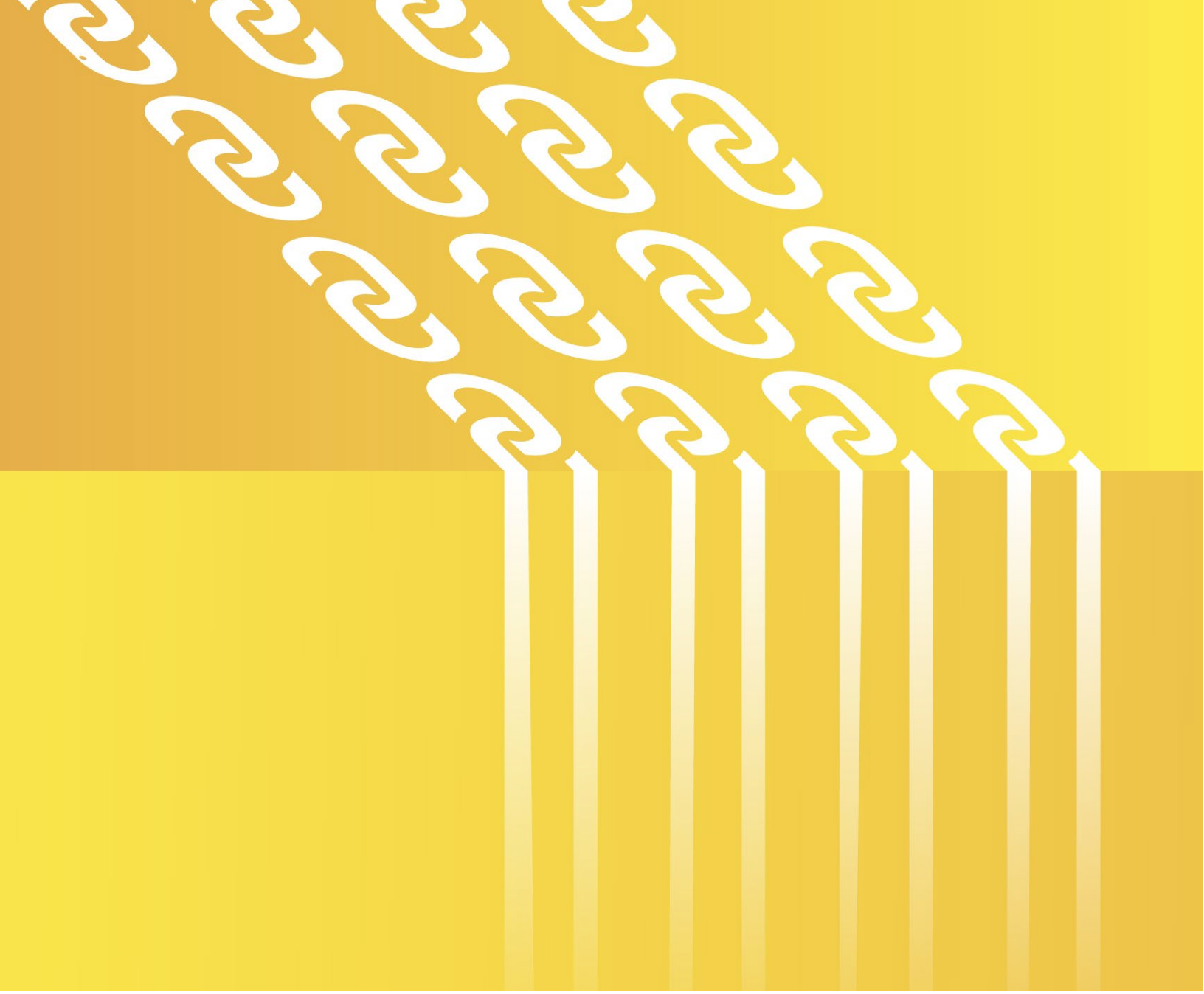
BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

Research Publication No. 2021-2
April 2021

The Paper of Record Meets an Ephemeral Web: An Examination of Linkrot and Content Drift within *The New York Times*

Jonathan Zittrain
John Bowers
Clare Stanton

23 Everett Street • Second Floor • Cambridge, Massachusetts 02138
+1 617.495.7547 • +1 617.495.7641 (fax) • http://cyber.law.harvard.edu/ •
cyber@law.harvard.edu



THE PAPER OF RECORD MEETS AN EPHEMERAL WEB:
AN EXAMINATION OF LINKROT AND CONTENT DRIFT WITHIN THE *NEW YORK TIMES*

JOHN BOWERS • CLARE STANTON • JONATHAN ZITTRAIN

SPRING 2021

LIL.LAW.HARVARD.EDU

Library Innovation Lab

The potential of web-based journalism to expand and enrich the ability of reporters to share the news is only growing, but reliance on hyperlinks can have consequences for the historical record. Readers experience "link rot" when clicking a URL only to receive a "404: Page Not Found" error, or an unexpected redirect.

This paper is solely the work of the authors. It is based off of a dataset created in partnership with the New York Times digital team and members of the Harvard Law School Library Innovation Lab's Perma.cc team. Perma.cc is an open source project aiming to combat link rot. This work is modeled after prior research conducted by one of the co-authors of this paper into the prevalence of link rot in the legal field. This report is part of the Berkman Klein Center for Internet & Society research publication series. The views expressed in this publication are those of the authors alone and do not reflect those of Harvard University or the Berkman Klein Center for Internet & Society at Harvard University.

HARVARD LAW SCHOOL LIBRARY

BERKMAN KLEIN CENTER FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

Library Innovation Lab

Hyperlinks give reporters and editors the freedom to bring anything that is publicly accessible on the web within a single click. Readers can turn to these links for detailed sources, related commentary, live Twitter skirmishes, and other vital contextualization of what, in print, was either in an article or nowhere to be found. But journalism's harnessing of the messy dynamism of the web is a double-edged sword. For all of its richness and interactivity, what's found on the web has a tendency to change, relocate, or disappear entirely. In tracking down a book or magazine pointed to by traditional citations, a reader can check with a local library, place an order with a bookseller, or query an online database. A hyperlink encountered online permits, counterintuitively, substantially less flexibility and redundancy. Pieces of web content accessible via a link, more formally known as a "uniform resource locator," or URL, exist as singularities controlled and maintained by a particular host. When hosts delete the content located at a given URL, whether intentionally or unintentionally, click throughs are met with a "file not found" page or an unreachable website. This often-irreversible decay of web content and the URLs that point to it is commonly known as "linkrot."

Linkrot has a subtler but no less insidious partner in "content drift," often-unannounced changes – retractions, additions, replacement – to the content available at a particular URL. Such changes can make the content at the end of a URL misleading – or dramatically divergent from the original linker's intentions, as when a clever troll purchased control over a domain to which Supreme Court Justice Alito linked in a published opinion, replacing the relevant material with a snarky reflection on "the transience of linked information in the internet age."[1] The Internet Archive and other services like it have sought to preserve captures of pages before they rot, but such preservation efforts are never comprehensive, and, thanks to the time that often passes between captures, rarely a complete answer to content drift.

The ephemerality of the web isn't just a problem for journalists. Any area of work that is reliant on the written record – a growing share of which is constituted by web content – is vulnerable to linkrot and content drift. In 2014, for example, a study[2] co-conducted by one of the co-authors of this essay found that nearly half of all hyperlinks included as references in United States Supreme Court opinions pointed to content which had disappeared from the web or changed significantly since publication. But given the role of journalism in shaping and substantiating public discourse, its adaptation to these vulnerabilities deserves a special measure of attention.[3]

We've undertaken a project to gain insight into the extent and characteristics of journalistic linkrot through a comprehensive examination of hyperlinks included in *New York Times* articles from the launch of the *Times* website in 1996 through mid-2019, developed on the basis of a dataset provided to us by the *Times*. We focus on the *Times* not because it is a deeply influential publication

1 https://blogs.harvard.edu/perma/2015/08/04/justice-alito-and-link-rot/, archived at https://perma.cc/55B9-482D

2 Jonathan Zittrain, Kendra Albert & Lawrence Lessig, *Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations*, 127 Harvard Law Review F. 176 (2014) *available at* https://harvardlawreview.org/2014/03/perma-scoping-and-addressing-the-problem-of-link-and-reference-rot-in-legal-citations, *archived at* https://perma.cc/D29D-MV4L

3 Studies examining the patterns and frequency of linkrot have been being conducted since the early 2000s focusing on various fields. In the legal sphere, Rumsey, as well as Davis examined web resources cited in the law reviews, while Liebler and Liebert took on US Supreme Court opinions. Markwell and Brooks conducted work relating to educational materials in biochemistry, and Koehler, Hennesey, and Xijin Ge took on a generalized view of the web. More recently, Jones et al. revisited a cross disciplinary dataset from 2014 examined by Klein et al. which focused on the sciences. Massicotte and Botter took on both content drift and link rot found in all ETDs added to an institutional repository managed by an academic library between 2011 and 2015. See methods appendix for full citations of previous work.

whose archives are often turned to for the purpose of helping to reconstruct the historical record around an event or an era. Rather, the substantial linkrot and content drift we find here across the *New York Times* corpus accurately reflects the inherent difficulties of long-term linking to pieces of a volatile web. In addition to exploring the patterns observable in this linkrot and content drift, we suggest potential steps that the *Times* and other publications might take to mitigate its impact both proactively and retroactively. As our examination of linkrot in the legal context resulted in the integration of the Library Innovation Lab's Perma[4] system into the canonical Bluebook legal citation guidelines, so too do we hope that this study might prompt a new conversation around journalistic best practices in relation to the web.

We found that of the **553,693 articles** that included URLs on nytimes.com between its launch in 1996 and mid-2019, there were a total of **2,283,445 hyperlinks** pointing to content outside of nytimes.com. **28%** of these were "shallow links" such as example.com. **72%** were "deep links" including a path to a specific page, such as example.com/article.

We focused our analysis on deep links, as they were the large majority of the sample, and lead to *specific* material that the article author hopes to point readers to. Of those, **25%** of all links were completely inaccessible, with linkrot becoming more common over time – **6%** of links from 2018 had rotted, as compared to **43%** of links from 2008 and **72%** of links from 1998. **53%** of all articles that contained deep links had at least one rotted link.

On top of that, some reachable links were not pointing to the information journalists had intended. **An additional 13%** of "healthy" links from a human-reviewed sample of 4,500 had drifted significantly since publication, with content drift becoming more common over time – **4%** of reachable links published in articles from 2019 had drifted, as compared to **25%** of reachable links from articles published in 2009.

Methodology

The dataset of links on which we built our analysis was assembled by the *New York Times* digital team, with each entry in the dataset corresponding to a URL included in a *Times* article between 2006 and April of 2019. After a distillation to remove duplicates and links to *Times* articles and social media accounts, the dataset consisted of 2,283,445 entries representing 1,627,612 unique URLs across the 553,693 articles containing URLs.

Structurally speaking, these links fell into two categories mentioned above – "shallow" links and "deep" links, which serve substantially different purposes. Shallow links are generally used to point readers towards a website in general terms – an article in the Business section might make a generic mention of IBM including a live link to www.ibm.com. Deep links, on the other hand, often point to specific articles, documents, or pieces of evidence – that same Business section article might, for example, include a deep link to a page with earnings data for a specific quarter. When deeplinks rot, specific material upon which the author relied in writing and substantiating the article is often rendered inaccessible; this is less the case for shallow links.[5] As such, we narrow our

4 https://perma.cc/

5 There are obviously exceptions to the mappings between link type and purpose outlined here. Sometimes, for example, authors cite specific pieces of content or information displayed on a homepage, or include a deep link to a page we would expect to be highly dynamic.

analysis of linkrot on deep links, which, per the figures given above, make up about 72% of the links in the corpus. Shallow links were omitted from the sample.

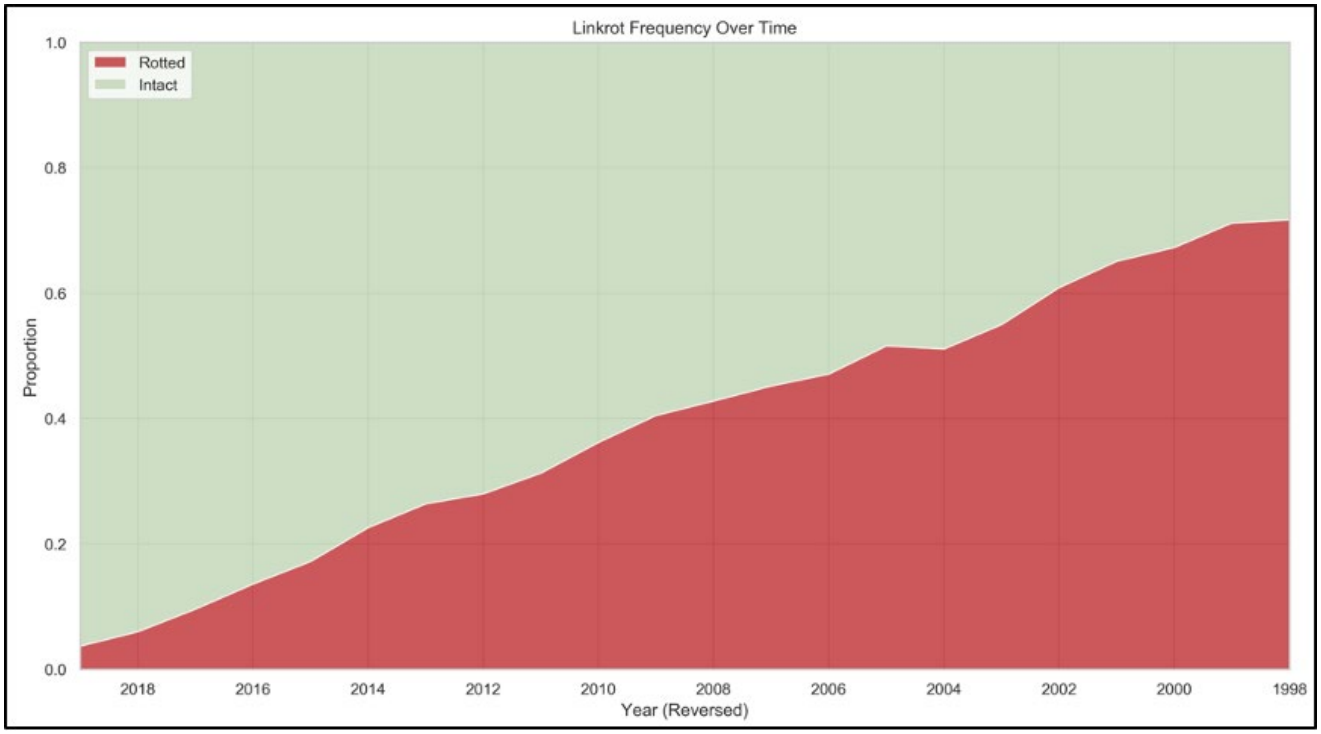
We measured linkrot by writing a script to visit each of the unique "deep" URLs in the dataset and log HTTP response codes, redirects, and server timeouts. On the basis of this analysis, we labeled each link as being "rotted" (removed or unreachable) or "intact" (returning a valid page)

Of course, returning a valid page isn't the same thing as returning the page as seen by the author who originally included the link in an article. To identify the prevalence of content drift, we conducted a human review of 4,500 URLs sampled at random from the URLs that our script had labelled as intact. For the purposes of this review, we defined a link that had fallen victim to content drift as a URL used in a *Times* article that did not point to the relevant information the original article was referring to at the date the article was published. On the basis of this analysis, reviewers marked each URL in the sample as being "intact" or "drifted."

The two labeled datasets emerging from these stages of collection form the basis for the results presented in this paper. We will refer to the latter as "the content drift sample" throughout.

Results – Linkrot

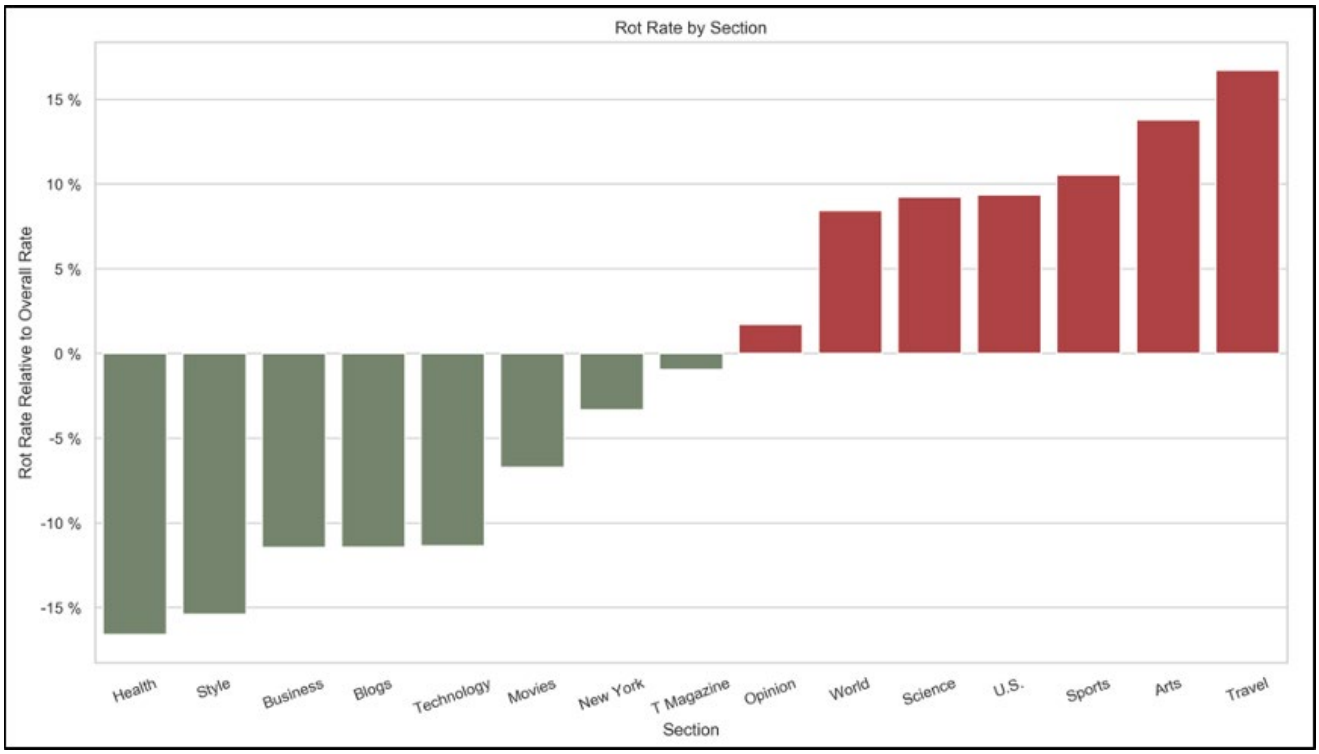
It is clear from this analysis that links included in *Times* articles are rotting at a consistent rate, and that a substantial proportion of the link corpus has already been rendered unreachable. Overall, we observed that about a quarter of deep links included in *Times* articles between 1996 and mid-2019 had been made inaccessible by linkrot. Unsurprisingly, URLs included in older articles were far more likely to have rotted than those included in more recently published articles, and the trend over time is roughly linear. Over half – about 53.5% – of all articles with a deep link included in the cleaned dataset had at least one rotten link.



Our analysis revealed that certain sections of the *Times* were displaying much higher rates of rotted URLs. Links in the Sports section, for example, demonstrate a rot rate of about 36%, as opposed to 13% for The Upshot. Given that links rot over time, however, these figures are confounded by differentials in the average age of links across sections. For example, the average age of a link in The Upshot is 1450 days, as opposed to 3196 days in the Sports section. These chronological distributions have little to do with differences in the sorts of links sections use, and how they use them.

To see how much these chronological differences alone account for variation in rot rate across sections, we developed a metric, Relative Rot Rate, to quantify this variation and provide a chronologically normalized indication of whether a section has suffered proportionally more or less linkrot than the *New York Times* overall. A section's RRR reflects the sign and magnitude of its divergence from the site-wide linkrot baseline.

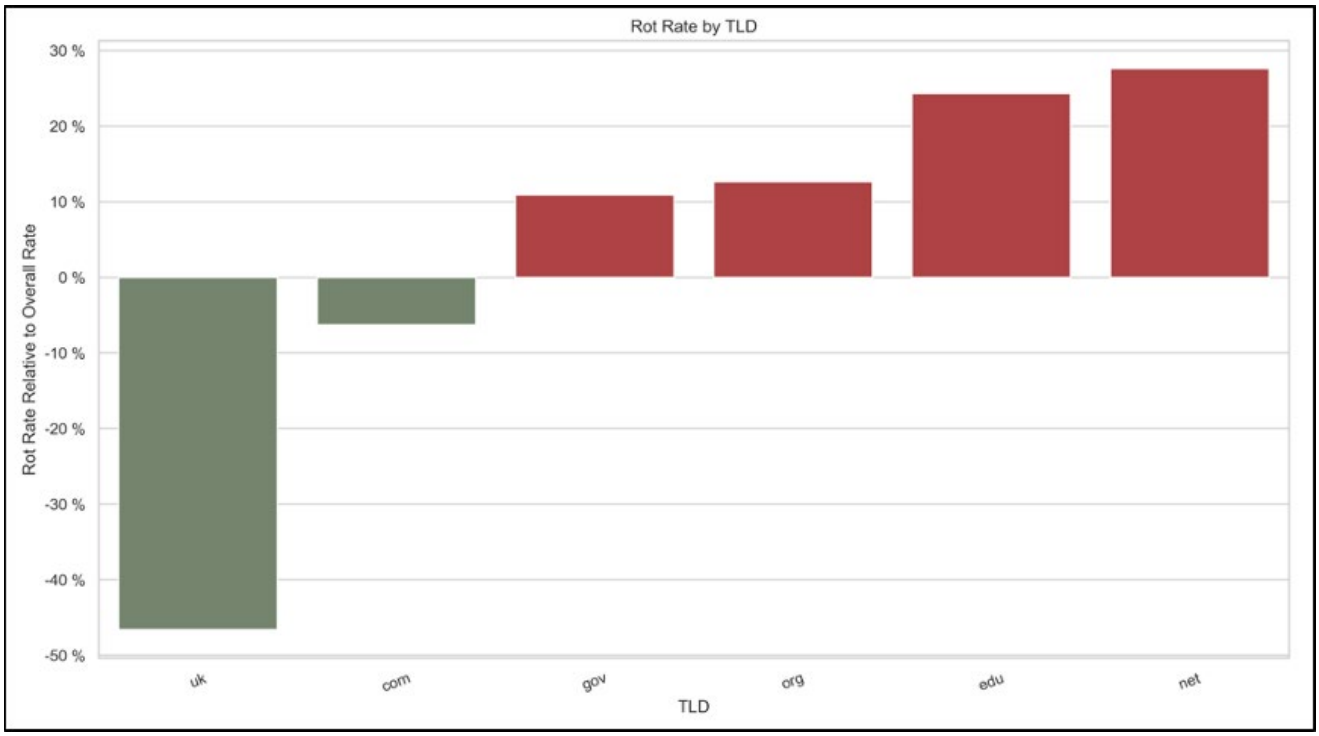
Using the RRR, we still found significant variation across the site's sections. Of the 15 sections with the most articles, the Health section had the lowest RRR figures, falling about 17% below the baseline linkrot frequency. The Travel section had the highest rot rate, with more than 17% of links appearing in the sections' articles having rotted.



Explaining the variation in linkrot across sections after controlling for article age is an exercise in approximation and, indeed, speculation, because most sections contain multitudes in terms of article content, style, and link usage. There is nevertheless much to be gleaned from analyzing the relative frequencies with which different types of links are used across different sections, and how those types might be relatively more or less susceptible to linkrot.

For example, a section that reports heavily on government affairs or education might be

disadvantaged by the fact that deep links to top-level domains[6] like .gov, and .edu show higher relative rot rates. This phenomenon is initially counterintuitive, as both governments and academic institutions are well regarded as enduring entities. In some ways however, this is unsurprising as these URLs are volatile by design: whitehouse.gov will always have the same URL but will fundamentally change in both content and structure with each new administration. Similarly, universities and academic institutions are controlled by a vast network of stakeholders who by nature have a high turnover rate. It is precisely because their domains are fixed that their deep links are fragile.[7]



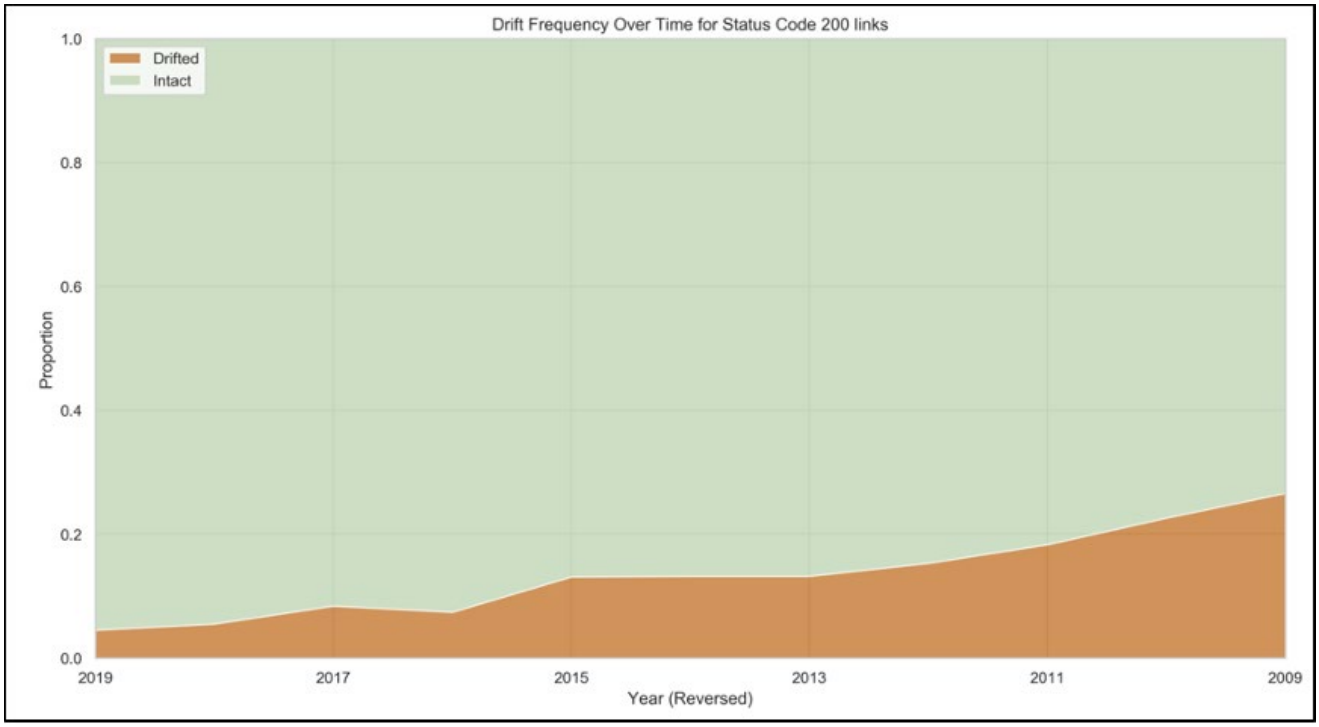
Results – Content Drift

Of the 4,500 "intact" links reviewed by our team, a total of 13% had changed significantly from the time of publication. Like the rate of linkrot, the rate of content drift increases linearly over time. We only observed 4% of reachable links published in 2019 to have drifted, as opposed to about 25% of reachable links from ten years prior.

We noted two distinct types of drift, each with different implications. First, a number of sites had drifted because the domain containing the linked material had changed hands and been repurposed. Although our sample only revealed about 1% of drifted links to be manipulated in this way, they are also the examples of content drift with the most immediate negative impact. Stories like Justice Alito's are versions of this type of content drift.

6 Every link URL comes with a "top-level domain" like .COM, .NET, or .ORG, many of which are intended to reflect the nature of the organization hosting the site.

7 Another irony: Both educational institutions and government entities have mandates for historical repositories of their materials, and content produced by them has long been seen as necessary for preservation. This practice appears to have lessened the focus on maintaining older material on the live web, as workflows existed long before the internet to maintain records offline in pre existing repositories.



More common and less immediately obvious, however, were web pages that had been significantly updated since they were originally included in the article. Such updates are a useful practice for those visiting most web sites – easy access to of-the-moment information is one of the Web's key offerings. Left entirely static, many web pages would become useless in short order. However, in the context of a news article's link to a page, updates often erase important evidence and context.

For example, an article published in the Metro News section about candidates for election to a Congressional seat representing Staten Island, published in 2008, references a member of New York City Council, and links to what was originally a website about him. Today, clicking on the same link would lead you to the current city council member's website. This type of content drift accounted for 99% of the cases we observed in our subsample.

A Path Forward

The existence of linkrot and content drift at this scale across the *New York Times* corpus is not a sign of neglect, but rather a stark reflection of the state of modern online citation and reference. The practice of creating connections between the contents of an article and the wider internet is one that enhances journalistic world-making. That it is being compromised by the fundamental volatility of the web points to the need for new practices, workflows, and technologies across the field of journalism. For news organizations such as The Times, external linking is a mark of journalistic transparency and a public service in making it easy for readers to reach the original material. But it also creates dependencies on third-party content beyond the journalists' control, and these findings suggest that the challenge of link rot has yet to be understood or addressed across the field.

Retroactive options – mitigating existing rot and drift – are limited, but nonetheless important to consider. The Internet Archive hosts an impressive though far from comprehensive assortment of

snapshots of websites, and can be used as a means of patching incidents of link rot and content drift. Publications could work to improve the visibility of the Internet Archive and other services like it as a tool for readers, or even automatically replace broken links with links to archives automatically, as the Wikipedia community has done.[8]

Looking forward, more fundamental measures should be taken. Journalists have adopted some proactive solutions, such as screenshotting and storing static images of websites. But it doesn't solve for the reader who comes across an inaccessible or confusing link.

New frameworks for considering the purpose of a given link will further help bolster the intertwined processes of journalism and research. Before linking, for instance, journalists should decide whether they want a dynamic link to a volatile web—risking rot or content drift, but enabling further exploration of a topic—or a frozen piece of archival material, fixed to represent exactly what the author would have seen at the time of writing. Newsrooms—and the people who support them—should build technical tools to streamline this more sophisticated linking process, giving writers maximum control over how their pieces interact with other web content.

Newsrooms ought to consider adopting tools to suit their workflows and make link preservation a seamless part of the journalistic process. Partnerships between library and information professionals and digital newsrooms would be fruitful for creating these strategies. Previously, such partnerships have produced domain-tailored solutions, like those offered to the legal field by the Harvard Law School Library's Perma.cc project (which the authors of the report work on or have worked on). The skills of information professionals should be paired with the specific concerns of digital journalism to surface particular needs and areas for development. For example, explorations into more automated detection of link rot and content drift would open doors for newsrooms to balance the need for external linking with archival considerations while maintaining their large-scale publishing needs.

Digital journalism has grown significantly over the past decade, taking an essential place in the historical record. Linkrot is already blighting that record—and it's not going away on its own.

8 https://blog.archive.org/2018/10/01/more-than-9-million-broken-links-on-wikipedia-are-now-rescued/, archived at https://perma.cc/49MR-UM7A

Methods Appendix

Distillation of the data set

For each URL, we had access to a range of relevant metadata including the headline and unique web address of the article in which the URL was included, the date and time of that article's publication, and the section and desk under which it was published. When a given URL appeared in multiple articles, an entry was provided for each appearance. To focus our analysis on links pointing to external content outside of the *Times*' direct control, we removed all links to *Times* articles and social media accounts. Additionally, we removed duplicate entries, malformed URLs, and URLs corresponding to protocols other than HTTP or HTTPS (like mailto or FTP).

Link Classification

When a request for a given URL returned a status code indicating unreachability (e.x. 403 Forbidden, 404 Not Found, 500 Internal Server Error) or our script could not establish a connection with the server, we marked that URL as having rotted. If the request resulted in a redirect to a different URL, we marked it as a redirect – unless the redirection was from a "deep link" (like www.example.com/pages/test.html) to a base URL (in this case www.example.com), in which case we marked the original URL as having rotted. If a request returned a non-erroneous status code (generally 200 OK) and did not redirect, we marked it as intact. To minimize false positives, we treated non-homepage redirects as intact. We did not check for so-called "soft 404s," whereby certain websites erroneously return a 200 status code even in cases where a resource is inaccessible. As such, the figures presented in this paper should be treated as lower bounds on the rate of linkrot.

Relative Rot Rate

First, for each year that a section published an article including a link, we multiplied the number of links in articles from that section by the site-wide rot rate for that year to get an "expected" number of rotten links for that year, then subtracted that number from the actual number of rotten links for that year. We summed these differences across all of the years that the section published at least one link. Next, we divided this sum of differences by the sum of the expected numbers of rotten links in the section for each year in which at least one link was published under the section. More formally, the RRR for a given section can be calculated as

$$RelativeRotRate = \frac{\sum_Y s_y - e(y)}{\sum_Y e(y)}$$

where

$$e(y) = t_y \frac{n_{sy}}{n_{ty}}$$

and Y is the set of years in which the section published at least one link, s_y is the number of links that rotted in the section for year y, $e(y)$ is the expected number of rotten links in the section for year y, t_y is the total site-wide number of rotten links for year y, n_{sy} is the total number of links included in the section in year y, and n_{ty} is the total number of links included site-wide in year y.

For ease of interpretation, we multiply all RRR figures by 100 to convert them into percentages. For example, an RRR of -30% indicates that the chronologically-adjusted frequency of linkrot in a given section is 30% lower than the frequency across the site overall, while an RRR of 130% indicates an adjusted frequency 130% higher than that of the site overall.

Human Review of Content Drift

The content drift set was generated randomly as a subset of URLs in the larger corpus that had been coded as intact. We focused our analysis on links within articles from the last 11 years – where link volume is much higher – so as to produce a statistically robust sample under our coding resource constraints and where one would expect less drift to have happened than for older links.

Our human reviewers – sourced from among the authors of this paper and the Harvard Law School Library staff – each received a subset of the URLs, along with the URL for the article each website appeared in, the unique linkID, and the date of the NYT publication.

Every included URL was individually opened in the reviewer's browser. URLs that opened to a seemingly functional page were reviewed alongside the article in which the URL was published to determine if the material of interest was in all likelihood still present. As with the distinction between deep and shallow links in our analysis of HTTP statuses, our reviewers treated "relevant information" as a reflection of the intention the author had in sending users to a page. It was often the case that a URL was used as a "further information" source rather than a specific citation.

Previous Work (Chronological):

Mary Rumsey, *Runaway Train: Problems of Permanence, Accessibility, and Stability in the Use of Web Sources in Law Review Citations,* 94 Law Libr. J. 27 (2002)

John Markwell & David W. Brooks, *"Link Rot" Limits the Usefulness of Web-based Educational Materials in Biochemistry and Molecular Biology,* 31 Biochemistry & Molecular Biology Educ. 69 (2003)
available at http://onlinelibrary.wiley.com/doi/10.1002/bmb.2003.494031010165/full, archived at http://perma.cc/N969-86A4.

Wallace Koehler, *A Longitudinal Study of Web Pages Continued: A Consideration of Document Persistence,* 9 Information Research, (2004)
available at http://informationr.net/ir/9-2/paper174.html, archived at http://perma.cc/8767-F7NG

Helane E. Davis, K*eeping Validity in Cite: Web Resources Cited in Select Washington Law Reviews 2001–03,* 98 Law Libr. J. 639 (2006)

Raizel Liebler & June Liebert, *Something Rotten in the State of Legal Citation: The Life Span of a United States Supreme Court Citation Containing an Internet Link (1996–2010),* 15 Yale J.L. & Tech. 273 (2013)

Jason Hennessey, and Steven Xijin Ge*, A cross disciplinary study of link decay and the effectiveness of mitigation techniques,* 14 BMC bioinformatics 14 (2013)
available at https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-

S14-S5, *archived at* https://perma.cc/8XTK-WCML

Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard Tobin, *Scholarly context not found: one in five articles suffers from reference rot.* 9 PLoS One 12 (2014)

available at doi:10.1371/journal.pone.0115253*, archived at* https://perma.cc/9GYG-X4NY

Shawn M. Jones, Herbert Van deSompe, Harihar Shankar, Martin Klein, Richard Tobin, and Claire Grover, *Scholarly Context Adrift: Three out of Four URI References Lead to Changed Content,* PLoSONE 11 (2016)
available at doi:10.1371/journal.pone.0167475*, archived at* https://perma.cc/KL4A-QVWL

Mia Massicotte and Kathleen Botter, *Reference rot in the repository: A case study of electronic theses and dissertations (ETDs) in an academic library,* 36 Information Technology and Libraries 1 (2017)
available at https://ejournals.bc.edu/index.php/ital/article/view/9598, *archived at* https://perma.cc/CP5W-TJHF

Acknowledgements

Thank you to our partners at the *New York Times*, particularly Seth Carlson and Matthew Ericson, for making this unprecedented dataset available to us. Additional thanks to those on our team at the Harvard Law School Library who were instrumental in completing this research. Their assistance with hand tracking and documenting content drift was essential to understanding the full extent of linkrot in this context. Of that group, thanks especially to Adam Ziegler, who additionally brought the data agreement process over the finish line.