Coming Back Differently: An Exploratory Case Study of Near Death Experiences of Webpages

Lesley Frew
Department of Computer Science
Old Dominion University
Norfolk, Virginia, USA
lfrew001@odu.edu

Michael L. Nelson
Department of Computer Science
Old Dominion University
Norfolk, Virginia, USA
mln@cs.odu.edu

Michele C. Weigle
Department of Computer Science
Old Dominion University
Norfolk, Virginia, USA
mweigle@cs.odu.edu

Abstract

In this case study, we use web archives to analyze 8,824 webpages that were taken offline and subsequently put back online, thus experiencing a "near death experience." We enumerate the stages of a webpage's near death experience, including the change from a successful HTTP status code to non-successful and back, the intermediate stage with markers such as an under construction banner, and an analysis of how the pages came back differently.

CCS Concepts

• **Information systems → Users and interactive retrieval**; **Digital libraries and archives**.

Keywords

versioned document collections, web archives, government documents

The web is ephemeral. According to a recent study, the median lifespan of a webpage is 2.3 years [15]. However, referring to a webpage as having a *lifespan* implies that a webpage begins, then eventually and abruptly ends. For many webpages this is true, but there are also webpages that temporarily stop working, but then resume functioning again. Upon resolving, the webpage could be different than before it stopped functioning. This phenomenon highlights a violation of the *ceteris paribus* assumption that is common in online research [7].

In this case study, we analyze 8,824 webpages on the United States Centers for Disease Control (CDC) website (`cdc.gov`) that were taken offline and subsequently put back online in early 2025, thus experiencing a "near death experience." For example, Figure 1 shows a CDC webpage taken offline, and Figure 3 shows a page that has returned with a notification banner that it will be undergoing changes.

1 Identifying Webpages with Near Death Experiences

We made a domain match CDX query[1] on February 17, 2025 for cdc.gov webpages captured in 2025, which resulted in 6,505,819 URL matches, canonicalized by SURT. We then filtered these matches for pages with an HTTP 200 status code followed by at least one non-200 status code and then followed by a 200 status code. This resulted in 25,000 matches. Many of these matches were likely embedded resources, such as images, so we further filtered the matches for items ending in `htm, html,` or a slash. This resulted in 8,824 candidates for webpages that experienced a near death experience. Pages with internal redirects, such as pages without a slash redirecting to pages with a slash, appear in the CDX as a candidate but did not truly have a near death experience.

For each of these webpages, we used the Memento protocol [14] to request an archived version of the webpage from the Wayback Machine from near January 1, 2025 and March 1, 2025. We also collected a version of the webpage on the live web in June 2025. CDC webpages contain a meta property with their last updated date, and we separated the webpages into two groups: pages with announced updates via the meta tag as of June 2025 and pages without announced updates. There were 276 pages with announced updates. We then further analyzed these pages manually and discovered additional silent updates to the pages in February 2025, discussed further in Section 2.3.

2 Near Death Experience Stages

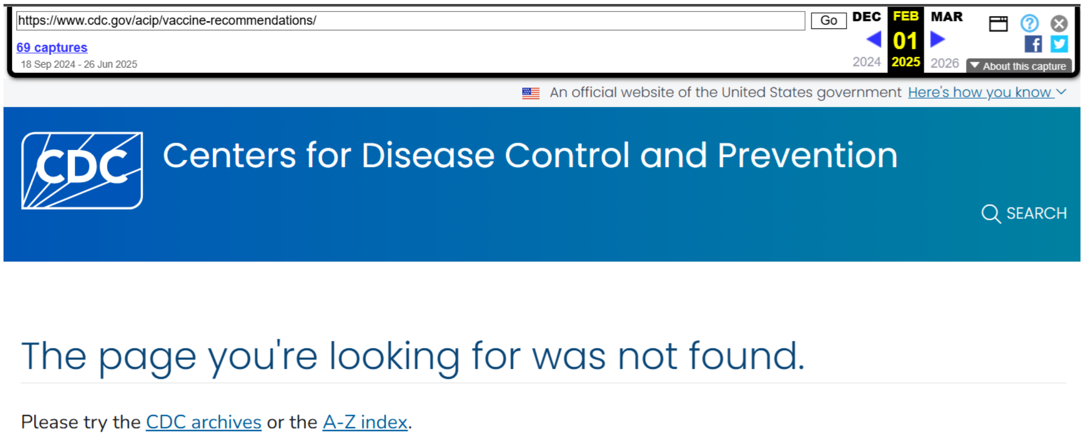
We find that pages go through three stages in their near death experience. First, they experience a period of time where they have a non-200 HTTP status code, which is the *clinical death stage*. Next, the pages resolve again but contain a marker that they are imminently changing or in danger, which is the *liminal stage*. Finally, the pages come back, possibly with semantically different content.

2.1 Clinical Death Stage

In the clinical death stage, webpages change from resolving with a successful 200 HTTP status code to any non-successful HTTP status code, as shown in Table 1. Web archives can capture pages with non-successful HTTP status codes, compared to timeout or resolution errors, which are not HTTP events. Figure 1 shows an archived version of a webpage with a 404 status code. In this dataset, pages were only offline for a brief period of time due to a judicial order[5]. In other datasets and contexts, the amount of time a webpage is in this stage would vary.

Table 1 shows the status codes of the pages in this dataset. In order to filter out false positive redirects, we filtered for pages that had a prefix of "http" to remove http to https redirects, and also removed non-200 pages with a different but canonical URL than the 200 match (such as ending in a slash or not), which removed 3,000 false positive redirects from the results. The status codes shown in Table 1 include only those with at least 20 pages in the dataset. In addition to the 404 Not Found status shown in Figure 1, we also found 403 Forbidden and 3xx Redirect statuses as shown in Figure 2.

[1] https://github.com/internetarchive/wayback/tree/master/wayback-cdx-server

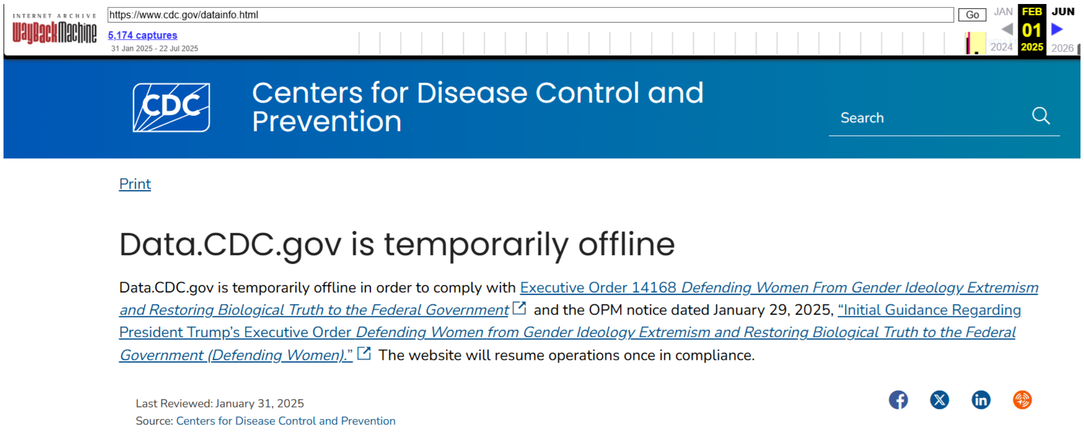

https://www.cdc.gov/acip/vaccine-recommendations/
Go
69 captures
18 Sep 2024 - 26 Jun 2025
DEC
FEB
MAR
01
2024
2025
2026
About this capture
An official website of the United States government
Here's how you know
CDC
SEARCH

Centers for Disease Control and Prevention

The page you're looking for was not found.

Please try the CDC archives or the A-Z index.

Figure 1: Many CDC webpages have at least one non-successful HTTP status code in February 2025. This webpage's 404 status code corresponds to the clinical death stage of the webpage's near death experience.



Centers for Disease Control and Prevention

Data.CDC.gov is temporarily offline

Data.CDC.gov is temporarily offline in order to comply with Executive Order 14168 *Defending Women From Gender Ideology Extremism and Restoring Biological Truth to the Federal Government* and the OPM notice dated January 29, 2025, "Initial Guidance Regarding President Trump's Executive Order *Defending Women from Gender Ideology Extremism and Restoring Biological Truth to the Federal Government (Defending Women)*." The website will resume operations once in compliance.

Last Reviewed: January 31, 2025
Source: Centers for Disease Control and Prevention

Figure 2: Data.CDC.gov webpages redirected to this error sink, which corresponds to the clinical death stage of the webpage's near death experience.

Status Code	All CDX matches	HTML matches only
301 Moved Permanently	9791	4496
302 Found	1153	2
403 Forbidden	8133	793
404 Not Found	3063	1593

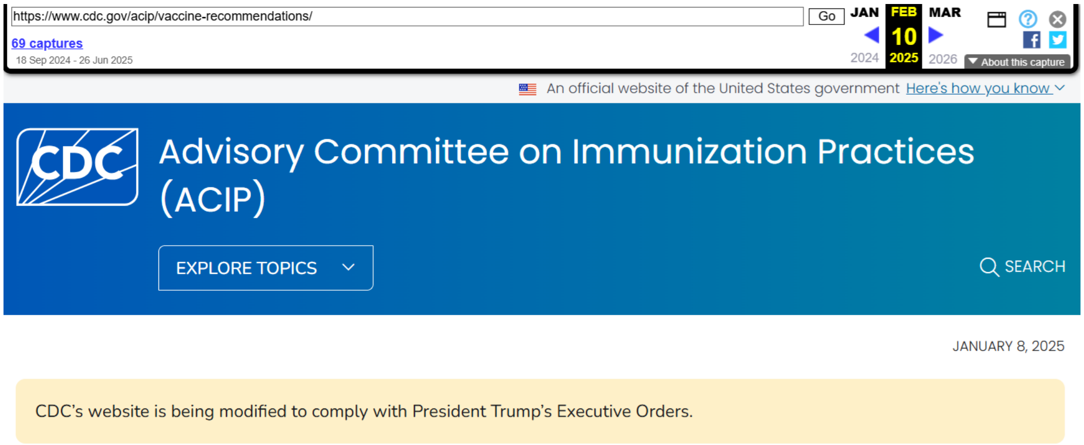
Table 1: Status codes of CDC near death webpages, for status codes with at least 20 pages. Redirects 3xx are the most common status code, followed by Forbidden and Not Found.

We further analyzed the redirects. data.cdc.gov had 736 redirects all to the same page as shown in Figure 2, also known as an *error sink* [6]. The National Healthcare Safety Network (NHSN) had 1,730 pages redirecting to another error sink. The Morbidity and Mortality Weekly Report site was temporarily moved resulting in 8,021 redirects. These three sites comprise nearly all of the redirects. The error sinks themselves had a status of 200. The presence of so many error sinks highlights the need to explore all status codes when investigating near death webpage experiences, even seemingly transient status codes like redirects.

In this dataset, the clinical death stage corresponds to intentional take down. In other datasets, the clinical death stage may correspond to access prevention for bots or geographical regions [1] or transient server unavailability [8].

2.2 Liminal Stage

Similar to how humans who experience near death experiences are referred to as being in a liminal state during a coma [9], webpages that experience a near death experience can also exist in a liminal state for a period of time. During this state, these webpages are on the boundary between life and death. The webpages have begun resolving with a 200 HTTP status code again during the liminal stage. Figure 3 shows an archived version of the Advisory Committee on Immunization Practices Vaccine Recommendations


https://www.cdc.gov/acip/vaccine-recommendations/
Go
JAN
FEB
MAR
10
2024
2025
2026
69 captures
18 Sep 2024 - 26 Jun 2025
About this capture
An official website of the United States government Here's how you know
CDC
EXPLORE TOPICS
SEARCH

Advisory Committee on Immunization Practices (ACIP)

JANUARY 8, 2025

CDC's website is being modified to comply with President Trump's Executive Orders.

Figure 3: CDC webpages displayed a banner in February 2025 while they were edited to comply with new executive orders. This banner corresponds to the liminal stage of the webpage's near death experience.

webpage displaying a banner referencing pending changes. These webpages are highly fragile and are at risk of changing or being entirely permanently deleted. The explicit fragility of this webpage resulted in public efforts to preserve it, as evidenced by captures in February 2025 from Save Page Now and Archive Team. In other datasets and contexts, additional markers of individual webpages or entire websites in this stage could include being labeled "under construction",[2] or including a note on the site's main landing page [12].

2.3 Coming Back Differently

In this dataset, we found three categories of pages: pages that experienced ongoing changes, pages completely unchanged after coming back, and pages that falsely claim to be unchanged. Of the 8,224 pages, 7,257 of them were completely unchanged after coming back. Without additional captures in web archives showing the non-200 status code and the liminal state banner, we would not know these pages had undergone a near death experience at all.

Up to 1,569 pages contained unannounced, silent updates, with 957 of these pages containing the meta tag with the date last updated. Our calculation uses a predefined government website boilerplate removal algorithm [10], though customizing this algorithm would allow us to further differentiate between pages with updated main content and pages with updated sidebar content that was not properly filtered out by the existing algorithm. Figure 4 shows an example of a silent update. The last updated date of this page for both versions is January 8, 2025, but there was a section of the page deleted between January 24 and February 10, 2025. Of these 957 pages with incorrect meta tags, 638 contained text replacements, 243 contained only deletions, and 76 contained only additions. We computed these changes using set calculations on tokens [2]. The median number of words deleted was 14 and the median number of words added was 4. CDC.gov does not use the HTTP Last-Modified header, so the meta tag is the best way to examine this property in this instance. Only 276 of these pages had announced updates as of June 2025. With one-eighth of pages in this dataset experiencing these silent updates, we conclude that the last updated date is untrustworthy.

There were 276 pages that experienced ongoing changes, with announced changes by June 2025. These pages also experienced a non-200 and non-redirect status in their clinical death stage. Of these, a minimum of 33 pages contained unannounced, silent updates as of March 2025. The reason why this is a lower bound is that webpages that were updated in between January 1 and the presidential inauguration on January 20th would return as announced updates under this setup, as the pair compared was from January 1 and March 1.

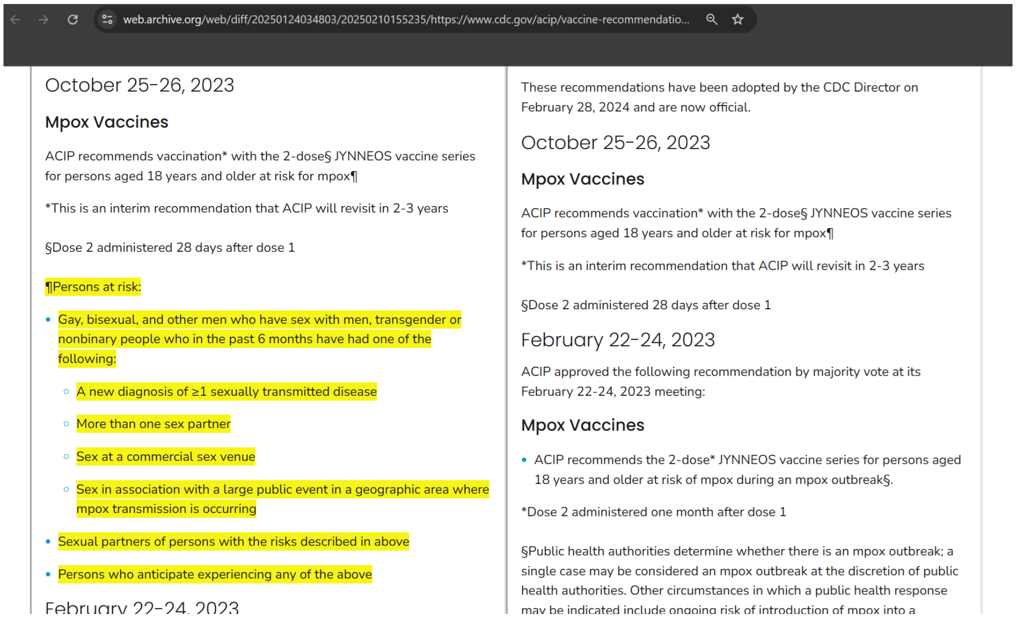
We analyzed the additions manually, but found no semantically meaningful additions. Rather, we found these pages experienced changes outside of their defined content, in areas such as the header or footer, or added internal table-of-contents style navigation to the page. We used Nost et al.'s 2020 boilerplate removal for government webpages [10], which demonstrably needs further updates for 2025.

3 Discussion

The applications stemming from web pages with near-death experiences involve two areas: monitoring and presentation. Web archives have seen a rise in many more webpages being captured via citizen archiving and monitoring groups [4] such as EDGI [10]. These monitoring organizations need a way to flag pages in the liminal stage for increased crawling coverage, to make the best use of their finite resources. Already, Save Page Now uses different crawling technology than is used for wide crawls in order to preserve these intentionally saved pages with higher fidelity [11]. While pages that temporarily go offline, perhaps because of a transient server issue, and then return identically will be marked as such in a web archive as a warc/revisit, pages in the liminal stage will have a new hash code, which opens the opportunity for automatic identification.

The second application area of interest is in the presentation of the changes on these pages. Past work focused on presenting pages to best capitalize on cognitive phenomena such as pre-attentive

[2] http://www.textfiles.com/underconstruction/


web.archive.org/web/diff/20250124034803/20250210155235/https://www.cdc.gov/acip/vaccine-recommendatio...

October 25-26, 2023

Mpox Vaccines

ACIP recommends vaccination* with the 2-dose§ JYNNEOS vaccine series for persons aged 18 years and older at risk for mpox¶

*This is an interim recommendation that ACIP will revisit in 2-3 years

§Dose 2 administered 28 days after dose 1

¶Persons at risk:

- Gay, bisexual, and other men who have sex with men, transgender or nonbinary people who in the past 6 months have had one of the following:
  - A new diagnosis of ≥1 sexually transmitted disease
  - More than one sex partner
  - Sex at a commercial sex venue
  - Sex in association with a large public event in a geographic area where mpox transmission is occurring
- Sexual partners of persons with the risks described in above
- Persons who anticipate experiencing any of the above

February 22–24, 2023

These recommendations have been adopted by the CDC Director on February 28, 2024 and are now official.

October 25-26, 2023

Mpox Vaccines

ACIP recommends vaccination* with the 2-dose§ JYNNEOS vaccine series for persons aged 18 years and older at risk for mpox¶

*This is an interim recommendation that ACIP will revisit in 2-3 years

§Dose 2 administered 28 days after dose 1

February 22-24, 2023

ACIP approved the following recommendation by majority vote at its February 22-24, 2023 meeting:

Mpox Vaccines

- ACIP recommends the 2-dose* JYNNEOS vaccine series for persons aged 18 years and older at risk of mpox during an mpox outbreak§.

*Dose 2 administered one month after dose 1

§Public health authorities determine whether there is an mpox outbreak; a single case may be considered an mpox outbreak at the discretion of public health authorities. Other circumstances in which a public health response may be indicated include ongoing risk of introduction of mpox into a

Figure 4: This CDC webpage experienced changes between January 24 and February 10, as shown in the Wayback Machine Changes Tool, but the last updated meta tag date for both pages is January 8.

processing [3], but future work will need to identify how to highlight things that should have changed but did not, such as the last modified date area on the page.

4 Future Work

Future work will include analyzing additional markers of stages of webpage near death experiences in larger longitudinal data sets, such as ClueWeb [13], Common Crawl[3], or the Not Your Parents' Web dataset[6, 15]. We aim to showcase additional websites that have undergone near death experiences for different amounts of time, and how the stages differ because of that. We also plan to further investigate the motivations for silent updates, both on this CDC dataset and other more heterogeneous datasets.

5 Conclusions

In this work, we presented a case study of CDC webpages that experienced a near death experience: they went offline for a period of time, as captured in web archives, then returned with a banner signaling impending changes. We analyzed the eventual changes to the webpages, and showed that the last modified date is not trustworthy due to a significant amount of silent, unannounced updates on these pages. We use this case study as simply one example of this common phenomenon of near death experiences of webpages, outlining the stages of this experience to enable future study on other datasets.

References

[1] Anat Ben-David and Adam Amram. 2018. The Internet Archive and the socio-technical construction of historical facts. *Internet Histories* 2, 1-2 (2018), 179–201.

[2] Lesley Frew. 2024. *Surfacing text changes in archived webpages.* Master's thesis. Old Dominion University.

[3] Lesley Frew, Michael L Nelson, and Michele C Weigle. 2023. Making changes in webpages discoverable: A change-text search interface for web archives. In *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 71–81.

[4] Lesley Frew, Michael L Nelson, and Michele C Weigle. 2025. Temporally Extending Existing Web Archive Collections for Longitudinal Analysis. *arXiv preprint arXiv:2505.24091* (2025).

[5] Lauren Gardner and Kyle Cheney. 2025. Judge orders Trump admin to restore removed health agency webpages. https://www.politico.com/news/2025/02/11/health-agency-webpage-removal-lawsuit-00203582.

[6] Kritika Garg, Sawood Alam, Michele C. Weigle, Michael L. Nelson, and Dietrich Ayala. 2025. Not Here, Go There: Analyzing Redirection Patterns on the Web.. In *Proceedings of the 17th ACM Web Science Conference*. doi:10.1145/3717867.3717925

[7] David Karpf. 2012. Social science research methods in Internet time. *Information, communication & society* 15, 5 (2012), 639–661.

[8] Andy Lawrence and Lenny Simon. 2021. Annual outage analysis 2021. *Uptime Institute Intelligence, UII-46 v1. 1PM* (2021).

[9] Limor Meoded Danon. 2016. Between My Body and My "Dead Body" Narratives of Coma. *Qualitative health research* 26, 2 (2016), 227–240.

[10] Eric Nost, Gretchen Gehrke, Grace Poudrier, Aaron Lemelin, Marcy Beck, Sara Wylie, on behalf of the Environmental Data, and Governance Initiative. 2021. Visualizing changes to US federal environmental agency websites, 2016–2020. *PLOS ONE* 16, 2 (02 2021), 1–27. doi:10.1371/journal.pone.0246450

[11] Jessica Ogden, Edward Summers, and Shawn Walker. 2024. Know (ing) Infrastructure: The Wayback Machine as object and instrument of digital research. *Convergence* 30, 1 (2024), 167–189.

[12] Magdalena Olszanowski. 2020. *girl. is. a. four. letter. word The Collective Practices of Amateur Self-Imag (in) ing and Personal Website Production 1996 to 2001.* Ph. D. Dissertation. Concordia University Montreal, Quebec, Canada.

[13] Arnold Overwijk, Chenyan Xiong, and Jamie Callan. 2022. ClueWeb22: 10 Billion Web Documents with Rich Information. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 3360–3362. doi:10.1145/3477495.3536321

[14] Herbert Van de Sompel, Michael Nelson, and Robert Sanderson. 2013. RFC 7089 - HTTP framework for time-based access to resource states–Memento. https://tools.ietf.org/html/rfc7089

[15] Michele C. Weigle. 2024. Some URLs Are Immortal, Most Are Ephemeral. https://ws-dl.blogspot.com/2024/09/2024-09-20-some-urls-are-immortal-most.html.

[3] https://commoncrawl.org/