Delicate Links: Ephemerality in Web-Based Evidence in Electronic Theses and Dissertations

Sarah Potvin, Tina Budzise-Weaver, and Kathy Christie Anders

abstract: This study suggests the need for best practices, specialized tools and standards, and targeted outreach related to Web-based content cited in theses. It analyzes citation practices in a corpus of master's theses in performance studies published at Texas A&M University from 2012 to 2020. Finding that only a slim majority of Web-based material cited in the theses remains fully functional within a decade of citation, this study confirms that "Web at large" content poses the greatest risk of irretrievable loss. Additionally, it considers actions by student authors that make theses vulnerable to evidentiary loss or change. A deeper understanding of the fragility of Web-based content and the potential for mitigation can inform needed interventions by librarians and other partners in graduate research.

Introduction

Citations are small, but mighty. As a convention of academic writing, they indicate a depth of research, an adherence to ethical norms, and an awareness of the author's participation in a scholarly conversation. The standard explanation of "why it is important to cite your sources," aimed at student authors and repeated across many institutional LibGuides, outlines four functions of citations:

1. Demonstrating an awareness of the literature
2. Giving credit and acknowledging ideas
3. Avoiding plagiarism
4. Guiding readers to locate sources and confirm the authors' representation of cited work.[1]

portal: Libraries and the Academy, Vol. 24, No. 3 (2024), pp. 519–552.

Many resources that graduate students cite in their papers, articles, and theses—both print and digital—can be expected to remain unchanged, fixed, and static. Libraries and other repositories, as well as publishers and other organizations, have invested considerable expertise, care, and resources to ensure ongoing access to print publications as well as to digital journals, electronic books, and materials deposited into research and data repositories. Additionally, these institutions coordinate the preservation of these materials.

But what about less stable "Web at large" materials, those texts, videos, and documents that live on the Web but are not formally stewarded with an eye toward fixity and preservation? The Web, as Jill Lepore describes it, "dwells in a never-ending present. It is—elementally—ethereal, ephemeral, unstable, and unreliable."[2] Visitors to a website may find that the site is missing, replaced by an error message. In this type of link rot, the loss is obvious both to human users and machines, which register http response codes corresponding to the failure. But there are less obvious and thus more insidious scenarios of loss and change, known as content drift, when a website has been altered or updated. As Lepore explains this problem, "It's impossible to tell that what you're seeing isn't what you went to look for: the overwriting, erasure, or moving of the original is invisible."[3]

Graduate students read and cite Web-based materials that change, move, and disappear. Like other scholars, they rely on these materials for their research. Content can also remain in place but inaccessible, locked behind subscription paywalls or custom logins. In other cases, the core content remains the same, while the site itself shifts to accommodate the user's search history and demographics, displaying targeted advertisements or other forms of personalization. The loss of ephemeral online materials is complex and confounding. These changes undermine the fourth function of citations: readers following links in notes and works cited may be unable to locate the referenced materials or to confirm the authors' representation of those materials. Indeed, the authors themselves may not be able to retrieve the Web-based materials that were central to their argument or analysis. This, in turn, challenges the validity of the scholarship and the scholarly record.

. . . readers following links in notes and works cited may be unable to locate the referenced materials or to confirm the authors' representation of those materials.

A substantial arc of digital preservation concerns has centered around the thesis, the article, and the book as intact objects. Often, digital preservation considers the many challenges involved in ensuring the persistence and fixity of these objects and related materials, such as supplementary data files hosted in repositories. But, as Jonathan Zittrain, Kendra Albert, and Lawrence Lessig have observed, loss of access to Web resources cited in a publication creates vulnerability *within* the scholarly work that extends to the scholarly record; this vulnerability "threatens the integrity of the resulting scholarship."[4]

The analysis in this article focuses on a set of electronic master's theses from the Performance Studies Department at Texas A&M University in College Station. Examining a small set of theses, in an interdisciplinary department with complex approaches toward evidence, allows for a closer attention to the loss of evidence than would a larger-scale, partially automated process. The researchers deployed a grounded theory approach, developing theory and coding based on the observation and analysis of data. They endeavored to ascertain student authors' reliance on Web-based resources and to closely examine the types of losses occurring with those references.[5] In addition to the data-forward approach to quantifying and characterizing the functionality of Web-based resources by cross-referencing the live Web, the researchers looked carefully at each thesis and considered methods used by the authors to embed, describe, and otherwise capture ephemeral Web-based performances or other materials. The researchers were guided by two sets of related research questions:

- Are Web-based materials linked as evidence and cited by the authors of theses in performance studies findable and verifiable? What are the barriers to accessing and verifying Web-based materials cited in theses?
- How can an analysis of lost evidence in master's theses guide interventions, including social and technical mitigation?

Ultimately, this research characterizes a preservation problem that the researchers seek to mitigate through sociotechnical interventions. Graduate students would benefit from educational interventions by academic librarians and other partners in graduate research, data management, collection management, and outreach at all stages of thesis research, writing, and submission. For example, resources and workshops could be developed to inform thesis authors about the relative fixity and persistence of print publications, publisher-hosted Web-based content such as journal articles and electronic books, library-hosted content such as theses and research data, and Web at large content. Librarians could introduce processes and tools that would enable authors to preserve Mementos, or archived snapshots, of Web-based content central to their work. Specialized tools could be incorporated into the writing and submission process to prompt or automate the capture of Mementos. At the same time, librarians and developers can consider how to better incorporate Web preservation elements into tools that graduate students use in thesis and dissertation drafting and submission. Further assessment will be necessary to identify optimal points of intervention.

. . . librarians and developers can consider how to better incorporate Web preservation elements into tools that graduate students use in thesis and dissertation drafting and submission.

Literature Review

In the formulation of Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard Tobin, "reference rot" denotes "the combination of two problems involved in using URI [Uniform Resource Identifier] references, both of which relate to the dynamic and ephemeral nature of the web."[6] These two problems are defined as:

1. Link rot: "The resource identified by a URI may cease to exist and hence a URI reference to that resource will no longer provide access to referenced content."
2. Content drift: "The resource identified by a URI may change over time and hence, the content at the end of the URI may evolve, even to such an extent that it ceases to be representative of the content that was originally referenced."[7]

Multiple studies have looked at link rot and content drift across corpora, developing methodologies to address targeted assessment and intervention goals.[8] Typically, these studies quantify the incidence of link rot through automated means, then extract a random sample for manual assessment of content drift. While early studies only occasionally distinguished between link rot and content drift, more recent studies demonstrate a growing awareness of and attention to the related but distinctive issues encompassed by "reference rot." Many such studies utilize sophisticated techniques to compare Mementos captured in Web archives.[9]

Concerns about the persistence and reliability of documents and sites on the World Wide Web have been frequently raised in the library and information studies literature since the Web's debut in 1991. In a 2001 article reporting on a four-year longitudinal study of Web documents, Wallace Koehler explores two types of behaviors: (1) demise and (2) change other than demise. Koehler deploys life cycle classifications to characterize sites as dead, comatose, intermittent, aging, and young. Observing a "great deal of as of yet unclassified subtlety to the Web," Koehler calls for a closer examination of "the purpose, function, and use of Web pages and sites."[10]

Concerns about the persistence and reliability of documents and sites on the World Wide Web have been frequently raised in the library and information studies literature since the Web's debut in 1991.

As scholars increasingly depended on electronic sources and publication venues, research focused on the reliability of Web-based materials. Multiple studies examined the incidence of Internet references and link rot across a set of discipline-specific journals or publications. Steve Lawrence and his coauthors focused on the Web-based resources cited in computer science journals, conference proceedings, and technical reports.[11] Like Koehler, they identified complexity in Web-based resources, documenting not only the likelihood of degradation over time but also the possibility of revival or improvements to fixity. For example, they observed that some lost Web references are "attributable to the initial rapid growth and evolution of the Web . . . Web pioneers ran their own servers on personal machines, and the links they created were lost with the disconnection of machines, relocation of servers, or change of server names."[12] They expect greater stability in time, as sites become institutionally managed and controlled: "Today, however, there are more widespread conventions for setting up servers . . . The easy availability of domain names has accelerated the movement of software and other projects from personal repositories to more stable, dedicated sites maintained by universities and corporations."[13]

In a 2003 article in *Science*, Robert Dellavalle and his coauthors assessed the accessibility of Internet references in articles in *Science*, *New England Journal of Medicine*, and

Journal of the American Medical Association. They coded whether a URL was active or inactive and if it was recoverable via Google or the Internet Archive. Dellavalle and his team found that 2.6 percent of all references in these articles were Internet references, with such references increasing over time. The URLs cited became frequently, and increasingly, inaccessible.

Examining citations to online sources in articles published in communication journals, Daniela Dimitrova and Michael Bugeja explore the "relationship between citation characteristics and their stability," with a goal of predicting stability and persistence of digital resources. They argue that "the erosion of Internet footnotes—the phenomenon of inaccessible online footnotes—undermines the standards of scholarship and the methods of research, primarily because it destabilizes fixed language and original source. This reverses the modern emphasis on the verifiable truths of subject matter or ideas, without which we empower unsubstantiated claims."[14] Dimitrova and Bugeja limit their coding to link rot, with coders reporting URL error messages, and pay little or no attention to content drift. This methodology captures the persistence of the URL itself, rather than the referenced resource. Edmund Russell and Jennifer Kane examine the persistence of Internet references in two top history journals. Their findings echo others: high rates of unreliability or inactivity, increasing over time, with some references findable via digital archives.[15]

Recent large-scale studies analyze shifting Web references in cross-disciplinary corpora using automated techniques. These studies have further operationalized distinctions between missing and changed Internet sources. Robert Sanderson, Mark Phillips, and Herbert Van de Sompel examine 160,000 Web references in hundreds of thousands of scholarly papers posted to two repositories—the University of North Texas Digital Library and the arXiv preprint server. As they note, previous studies relied on manual checking to discover missing resources. Sanderson, Phillips, and Van de Sompel use Memento TimeMaps, machine-readable lists of time-specific copies of archived original resources. With TimeMaps, they automate large-scale queries of nine prominent Web archives and compare resolution and archival rates for repositories and disciplines.[16] They identify a need for repository managers to "become more involved in the preservation of the scholarly communication record, beyond their own deposited content or other managed resources," through a process that would extract Internet references and automate their capture in Web archives.[17]

. . . Klein and his coauthors argue that current approaches to capturing resources cited in scholarly publications are inadequate, relying on "incidental archiving" by authors, publishers, and libraries.

Other studies have distinguished between the types of Web-based content cited in scholarly references. Martin Klein and his coauthors analyze citations in 3.5 million science, technology, and medicine articles published between 1997 and 2012. They seek to determine "the extent to which the web at large context that surrounds journal articles can be revisited some time after their publication." By focusing on "Web at large resources," Klein and his team exclude references to Web-based journal articles, which are typically scaffolded through URIs such as digital object identifiers (DOIs) that enable greater persistence and accessibility. Echoing both Koehler's call for distinguishing the variety of Web-based resources and his deployment of health-based classification metaphors, Klein and his coauthors introduce a typology for reference rot at the article level. Articles without Web at large references are considered "immune." Those that reference Web at large resources which are active and represented in Web archives are called "healthy." Articles that cite URIs which are lost, drifted, or not represented in Web archives are considered "infected" with reference rot.[18]

Studies have also suggested potential solutions to the problem of inaccessible Web references. In a 2005 article, Bugeja and Dimitrova take into account the practical needs of disciplinary educators and practitioners. These experts, they say, are "not only are expected to use the medium as scholars but are responsible in large part for its diffusion into academe and beyond."[19] In the absence of "universal standards . . . ensuring accuracy and access," Bugeja and Dimitrova reinforce a reliance on print. They recommend that scholars (1) cite print journals and books when possible; and (2) when referencing Internet resources, print "two hard copies of Web citations (one for the author's files and one for the editor's files, to be sent upon request). We also recommend storing digital copies of source documents."[20] They note, "We also considered an option that might have copyright implications—posting sections of cited materials in one author-generated URL per paper or study."[21] In effect, the current possibilities of Web archives and making links more robust formalize these recommendations, enabling digital copies to be "stored" in online archives where they can be referenced.[22] Russell and Kane call on "professional societies, journals, and presses to create and adopt professional standards for the use of Internet documents, including means for preserving materials in a way that ensures their accessibility into the indefinite future."[23] With attention to the sociotechnical nature of a solution, Lawrence and his eight coauthors provide detailed instructions for authors, including citation practices and the use of preprint repositories and software archives. They also delineate roles for professional societies and funding agencies in developing standards, sponsoring repositories, and requiring their usage.[24] Like Sanderson, Phillips, and Van de Sompel, Klein and his coauthors argue that current approaches to capturing resources cited in scholarly publications are inadequate, relying on "incidental archiving" by authors, publishers, and libraries.[25] Instead, they promote the work of the Hiberlink project, which aims to preserve Web resources referenced in scholarly papers. They call for the development of "pro-active archiving approaches intended to seamlessly integrate into the life cycle of an article and to require less explicit intervention by authors."[26]

Studies have analyzed and considered the incidence of link rot in academic-adjacent fields, such as journalism and the law. An influential 2013 study by Zittrain, Albert, and Lessig documents widespread reference rot in both law review journals and Supreme Court opinions. The study introduced a new Web archiving service, Perma.cc, to combat the problem.[27] In their 2021 study of link rot in the *New York Times* online, John Bowers, Clare Stanton, and Jonathan Zittrain find that 25 percent of deep links from www.nytimes.com were "completely inaccessible." They observe differential rot associated with older materials or those with .gov or .edu domains.[28] Further, 13 percent of a random, manually reviewed sample of intact URLs demonstrated significant content drift. Journalists, they advise, should develop frameworks to inform their usage of URLs in online news articles, buttressed by technical tools and partnerships with library and information professionals.[29]

Only a handful of studies have considered reference rot in theses and dissertations.[30] Mia Massicotte and Kathleen Botter study link rot and reference rot in Concordia University's collection of electronic theses and dissertations. They find, in alignment with other studies, that "linkrot manifests itself quickly after publication and increases over time."[31] Massicotte and Botter extracted and programmatically checked for link rot in 664 doctoral dissertations submitted over five years. They then used random sampling and hand coded a subset (10 percent) of dissertations for content drift. To determine the extent of content drift, they consulted Mementos, archived snapshots of prior webpages maintained in the Internet Archive Wayback Machine, a service that digitally archives the Web. Their analysis was segmented into disciplines, finding distinctions in link rot for dissertations from arts, business, engineering, fine arts, and science across the five years of publication. Massicotte and Botter detail some of the challenges of programmatic link rot assessment, including the potential for false positives when URLs return active status codes. A Tumblr post that has been removed, for example, will display a customized 404-page-not-found message from Tumblr. When programmatically analyzed, this Tumblr 404 appears to be an active, functioning URL. Seventy-seven percent of tested links in their set returned active response codes.[32]

Tackling link rot in electronic theses and dissertations might benefit from the relative control that universities exercise over their production and dissemination. For example, as previously described, Russell and Kane suggest that "professional societies, journals, and presses" must coordinate to address the link rot problem in the history discipline.[33] This coordination proves an insurmountable barrier. For theses, however, coordination might be achieved within the confines of an institution, drawing on intramural expertise and resources. Massicotte and Botter make a compelling argument for the importance of librarians' taking charge of this problem for theses, writing: "[Librarians] collectively bear greater responsibility for this body of scholarly work, and need to move forward from a position of benign neglect to one of informed curation and pro-active preservation."[34] This article builds on Massicotte and Botter's methodology. It also takes up their call for academic librarians to recognize the "new obligations and curatorial functions" necessitated by the move toward electronic theses and to engage in proactive stewardship.[35]

Methodology

The researchers examined 27 master's theses from 2012 through 2020 that were publicly available through the Texas A&M University DSpace repository.[36] Of the 27 theses, 5 did not contain URIs or uniform resource locators (URLs) and were excluded from further analysis. The researchers identified and tested 484 discrete linked resources in 22 theses submitted between 2012 and 2020. All footnotes or bibliographic citations with URIs were copied and pasted into a spreadsheet. The URIs were then reviewed. The majority of URIs were rendered as URLs, with the exception of some DOIs; the researchers transformed DOIs represented as URIs into URLs. URLs that were typed into the thesis incorrectly, demonstrating evident human error, were corrected.[37] Duplicate citations and links were also removed for efficiency when checking broken links.

Table 1.
URLs included in coding

Analyzed theses	Quantity
URLs extracted from theses	678
URLs removed because of duplication	194
URLs coded	484

The researchers used a qualitative content analysis approach to analyze the URLs in the performance studies theses. Margrit Schreier describes qualitative content analysis as systematically analyzing qualitative data to capture relevant meaning and to create a coding frame consisting of categories. The coding frame can be developed from inductive or deductive analysis of the data. Once the coding frame emerges, pilot testing is necessary to determine any modification to the categories before coding all data and interpreting the results.[38] The researchers acted as three independent coders. A coding manual was developed, and norming exercises were conducted to test the categories before coding and to ensure each coder understood the fields in the spreadsheet. Kimberly Neuendorf describes human coders as "useful" because, she says, they generally have knowledge of the data and methodology.[39] The researchers looked at inductive and deductive methods to examine the data and create meaningful categories of analysis.

In this study, the researchers focus on coding as human readers, rather than using purely computational approaches. A large-scale computational assessment of link rot would distinguish between http status codes; human readers must interpretively assess the functionality of the live Web representation of the cited resource.[40] This study seeks to capture the experience of encountering reference rot as a human reader.

In addition to coding the URLs extracted from the theses, the researchers examined each thesis as a text, attentive to the authors' distinctive approaches to describing, capturing, and referencing ephemeral performances. As detailed in the discussion, performance studies is an interdisciplinary field invested in both reproducibility and ephemerality. Researchers have evolved varied approaches to conceptualizing texts, images, theater, and other media as performances rather than static objects. These approaches are suggestive in the realm of Web-based evidence that moves, changes, and performs for every viewer, generating personalized ads and tallying page views. The types of ephemeral texts and information present in performance studies are important not only to that field but also to many other disciplines. For example, a YouTube video could be a significant data point for political scientists, journalists, and historians. Informed by Donna Lanclos's "argument for open-ended exploratory, anthropologically informed, qualitative work" in library and information studies, this study brings an inquisitive methodology to bear on the question of reference rot assessment.[41]

To start forming initial reactions to the URL dataset, the researchers reviewed past literature on reference rot, which deductively influenced the creation of initial categories. But upon review of the URLs, the researchers took a grounded theory approach to identify emerging themes to better define the final categories used in the overall analysis. Grounded theory involves the "systematic discovery of the theory from data."[42] The researchers found that as functionality issues arose, the emerging data influenced the categories, thereby modifying the final interpretation and addition of categories for coding. The 484 URL links were divided among the researchers, each of whom served as a coder, with overlap of independent coding to help synthesize results later. Researchers coded the data from January 19, 2021, through March 10, 2021. They did not sign into a university network to avoid IP (Internet Protocol) authentication on campus that obscured paywalled resources. Once all links were analyzed, the three authors convened to finalize all coding. If coding differed, the researchers collaboratively examined the URL and discussed the coding rationale, ultimately coming to an agreement of the analysis.

The researchers first classified each resource type based on the domains where materials were hosted, such as .gov, .com, .org, or .edu. They then identified the URL as an article, book, newspaper, blog, social media, or other resource in a separate column (see Appendix A for coding spreadsheet categories). The URLs in the spreadsheet were double-checked against the original thesis to ensure that no errors had been introduced in extraction. Observing seven links that appeared to have typos introduced by thesis authors, the researchers added a column of corrected URLs.

The coders went through incremental testing. After assessing whether the link resolved, the coders inspected the referenced website and coded for:

- URL resolving to the resource described in the citation
- Indications of content drift, including no observed content drift, evidence of core (central) content drift, and evidence of minor content drift
- URL redirecting to a different landing page
- Observing error in creation of URL link by author, beyond simple typographical errors
- Resource blocked by restriction, paywall, or other access barrier
- URL linking to an error landing site, 404, time-out, or missing resource page.

These categories grew out of the data, with one category informing the next as the researchers moved through testing. Given the study's goal of informing the design of targeted interventions to reduce reference rot in electronic theses, the methodology was aimed toward identifying potential factors associated with or deployed effectively against reference rot. To invoke the medical immunity metaphor suggested by Klein and his coauthors, the researchers were attuned to observing patterns in why some theses were "healthy" (referencing relatively stable or well-documented Web at large resources), while others were "infected" with reference

. . . some theses were "healthy" (referencing relatively stable or well-documented Web at large resources), while others were "infected" with reference rot (citing URIs that were lost or had drifted, undermining the validity of their evidence).

rot (citing URIs that were lost or had drifted, undermining the validity of their evidence).[43] Additional categories were developed to understand if missing or drifted content could be found, with a goal of addressing the first research question ("Are Web-based materials linked as evidence and cited by performance studies thesis authors findable and verifiable?"). Seeking any patterns or characteristics of URL health that would address the second research question, aimed at designing relevant interventions, the researchers added coding categories for the use of permalinks, DOIs, or other persistent identifiers; the language of the referenced resource; and additional coder observations.

The researchers asked if the live websites accessed via links in the theses matched the websites that thesis authors had consulted and cited. In addition to examining missing websites, the researchers coded for content drift and explored the instances where it occurred across the corpus of theses.

Klein and his coauthors describe content drift as the change that takes place when the content has ceased to represent the original resource referenced.[44] Deploying a computational approach, Shawn Jones, Herbert Van de Sompel, Harihar Shankar, Martin Klein, Richard Tobin, and Claire Grover quantify content drift in their study of more than a million URIs extracted from journal articles that referenced Web at large resources. The six authors rely on "well-established similarity measures to compare textual content" between live websites and Mementos captured in any of 19 Web archives.[45] Massicotte and Botter classify content drift into nuanced, subjective incidences of "major" and "minor," based on a comparison between live websites and archived Mementos captured in the Wayback Machine. Their category of minor content drift applies to "pages that differed somewhat from a memento in visual appearance, such as header and footer differences, changes in background theme or style, or changes in navigation or search functionality which did not represent a high degree of impairment."[46] Adapting Massicotte and Botter's approach to categorizing content drift as major or minor, this study defines content drift as "core" or "minor."

Table 2 details the researchers' coding categories for links, expanding upon categories in previous studies, including Massicotte and Botter. These categories are tailored to this study's unique approach of having human coders observe incidents and assess functionality, as opposed to using a computational or automated analysis.

Results

This study's assessment of whether Web-based materials referenced by performance studies theses at Texas A&M were findable and verifiable required attention to link rot, content drift, and additional measures. In the most fundamental analysis of the persistence of referenced weblinks, the researchers looked at whether the links resolved and if they were functional. In more refined analyses, this study looked at URL breakages over time, by type, and by retrievability of the source through expert searching.

Resolution and Functionality

The researchers first investigated whether the links provided in thesis citations resolved. In the performance studies corpus, 128, or 26.4 percent, of 484 links did not resolve or

Table 2.

Types of loss and change categorized by researchers

Observed incident	Observed effect on functionality (coding category)	Description
Broken link/link that does not resolve	Not resolved	A live website does not materialize via the referenced URL. Coding further specified 404, time-out, error in link, not designed to resolve, page not found, expired domain, or other. Additionally, the researchers coded custom 404 messages and Web resources that were entirely missing (such as removed YouTube videos) as broken/not resolved.
Lost resource	Not resolved or diminished, functionality depending on error	Declared in the event that a resource described in a citation but not findable at the provided URL cannot be located by two professional librarians. Each professional librarian dedicated at least 15 minutes to searching via both Google and, if applicable, library indexes. Researchers looked for missing resources using the Wayback Machine associated with the Internet Archive but did not systematically examine other Web archives.
Core content drift	Diminished functionality	An observable, significant change, such as major revisions, missing or changed content. Web resource is tangibly different from what the student referenced.
Author error	Diminished functionality	The use of URLs that were problematic when first referenced, including links to the wrong YouTube video or links to nonspecific landing pages such as jstor.org. Not included in this category are minor, evident typographical errors in URLs, such as transposed letters, which researchers could readily identify and fix.
Erroneous redirect	Diminished functionality	An area of complexity in this analysis, redirects typically signal maintenance by website custodians, who have architected jumps away from URLs that are no longer active. However, in some cases redirects fail to point to

Table 2, continued.

Observed incident	Observed effect on functionality (coding category)	Description
		functional resources and instead point away from the intended site. For this study, the researchers flagged redirects as instances of diminished functionality where they led away from the intended resources, calling them "erroneous redirects." Researchers did not flag redirects that led to the intended resource.
Significantly restricted access	Diminished functionality	A number of resources were not immediately accessible to coders, requiring log-in, subscription, or invitation-only membership. The category of "significantly restricted access," signaling diminished functionality, was reserved for resources that the coders could not access; these were password or log-in restricted sites that required either invitation or administrative approval to access. Researchers reasoned that resources that required users to set up log-in credentials or pay subscriptions or fees, while involving barriers, did not meet the standard of significantly restricted access.
Perfectly functional	Perfectly functional	Perfectly functional links were defined by the researchers as meeting three criteria: (1) the student has linked the appropriate resource; (2) the link resolves to the correct page as indicated by the citation; (3) and the linked Web resource shows no evidence of content drift in the central content of the page, such as changes to the core text. In these instances, the researchers deduced that the resource they observed on the live Web likely matched the resource referenced by the author of the citing thesis.
Minor content drift	Perfectly functional	Researchers observed dynamic, interactive aspects of sites, such as comment sections, advertisements, likes, and minor insignificant changes that do not affect the "core" intent of the resource.

resolved to missing pages. This means that when a coder opened the link in a browser, it showed an error message such as a 404, a time-out, or a missing core resource. A little under three-quarters, or 73.6 percent, of the links resolved. This analysis alone does not provide an accurate representation of whether the material cited appears as the author intended or whether the live website represented via URL is an accurate, valid representation of the originally referenced work.

To examine this question further, the researchers looked at the set of resolved links to see if they were functional. Perfectly functional links were defined as meeting three criteria: (1) the student has linked the appropriate resource; (2) the link resolves to the correct page as indicated by the citation; (3) and the linked Web resource shows no evidence of content drift in the central content of the page, such as changes to the core text. As an example, a citation that references a news article with a link to https://www.nytimes.com would not meet the criteria of perfect functionality. While this link resolves to a live website, it does not bring up the article referenced by the thesis author but, rather, the container website of the *New York Times*. Additionally, the core text of the site has changed since the moment of the student's reference (and, indeed, changes often in a single day). Of the total set, approximately 21.7 percent of the links that resolved were not perfectly functional. Just over half, 51.9 percent, of the total set of links both resolved and met the criteria of perfect functionality (see Figure 1).

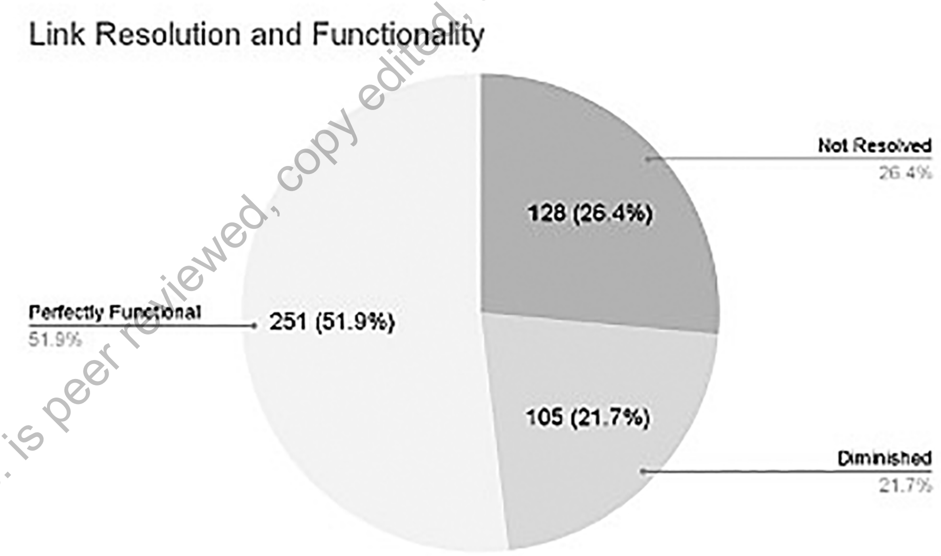

Figure 1. Percentages of link resolution and functionality.

The most common reasons for reduced link functionality were core content drift; author error in creating a link, such as linking to a home page instead of a specific resource page; and significantly restricted access to the site, such as password-protected social media and blog sites.

Several common reasons for links having diminished functionality emerged in the analysis. The most common reasons for reduced link functionality were core content drift; author error in creating a link, such as linking to a home page instead of a specific resource page; and significantly restricted access to the site, such as password-protected social media and blog sites. In these instances, the coders could not use the URL included in the citation to gain access to the referenced resources. Finally, redirects also caused links to be nonfunctional if they failed to point to moved content. Some links had more than one reason for being nonfunctional and so were coded for multiple errors. Figure 2 shows the breakdown of links coded as having diminished functionality.

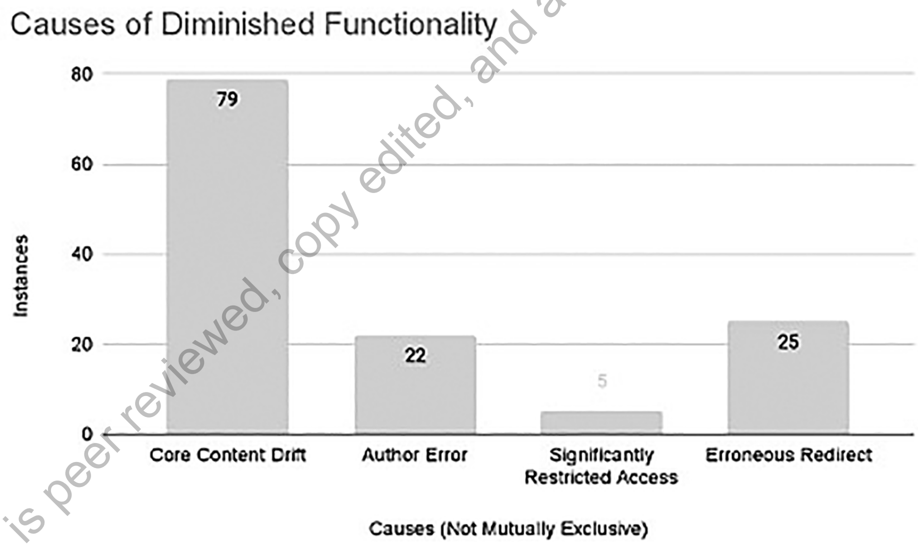

Figure 2. Causes of diminished link functionality.

Diminished functionality appears as a multidimensional problem. The researchers observed gradations of the problem, attributable to core content drift, author error, restricted access, erroneous redirects, and occasionally an overlapping combination of these issues. Core content drift, where some essential part of the work being cited had changed over time, was by far the most common cause of diminished functionality. Examples of core content drift include references to continuously updating pages, such as Wikipedia entries, and sites that obviously had new content not related to the citation. Some weblinks showed multiple reasons for diminished functionality. In some cases, for example, the author had made an error in generating the weblink; when the corrected weblink was entered, the page showed evidence of core content drift. Rather than assign only one reason for diminished functionality, the researchers coded all applicable categories. Thus, the instances reported in Figure 2 are not mutually exclusive.

Core content drift, where some essential part of the work being cited had changed over time, was by far the most common cause of diminished functionality.

Restricted access proved to be a particularly complex area of diminished functionality. Researchers identified 37 URLs that led to pages which were not publicly accessible. These included articles in scholarly journals, the *New York Times*, and the *Washington Post*, among other venues limited to subscribers; content hosted on sites like Tumblr and Facebook that required log-ins to access; and content that required membership in private, invitation-only groups. Researchers noted that links to sources restricted because of paywalls (27 links) and accounts required (5 links) presented a barrier to access but did not necessarily diminish the functionality of the links themselves. Five links, however, presented a more significant barrier by linking to a private group (three links), to blocked searching (one link), and to a location-restriction site (one link). The researcher coded those five links in the "significantly restricted access" category.

Minor Content Drift

While not a cause of links being nonfunctional, it is significant that coders observed evidence of minor content drift in 244 out of 484 links (50.4 percent). In this study, minor content drift is characterized by changes to the webpage that are paratextual interactive features, such as view counts and comments sections. Some pages also had advertising inserts. Given that these functions are designed to be dynamic, it is to be expected that they will change after the point at which the thesis authors cited them. Minor content drift was pervasive in the set of resolved links, appearing 68.5 percent of the time (see Figure 3). While minor content drift might not be a concern in many disciplines, there are larger implications for the field of performance studies, as well as for any disciplines that find significance in the interactions between a text and the reader.

Breaks over Time

In addition to considering the total set of links and their functionality, the researchers looked at aspects of the links such as their age and the genre of material being cited. The researchers observed a direct relationship between the age of a thesis and whether links included in its references resolved. Figure 4 shows the percentage of weblinks not resolved by the year the theses were filed. A best-fit line shows a clear decrease in the percentage of links that did not resolve from 2012 to 2020.

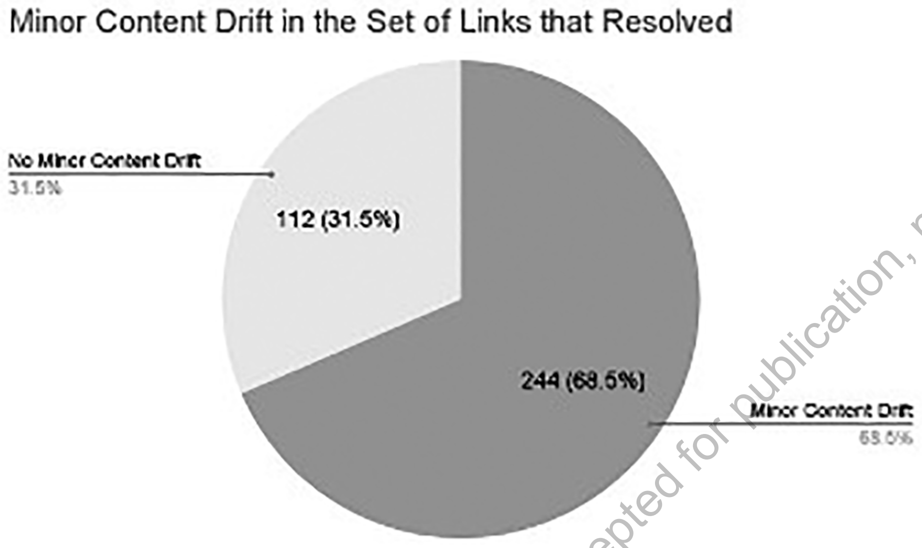

Figure 3. Minor content drift in the set of links that resolved.

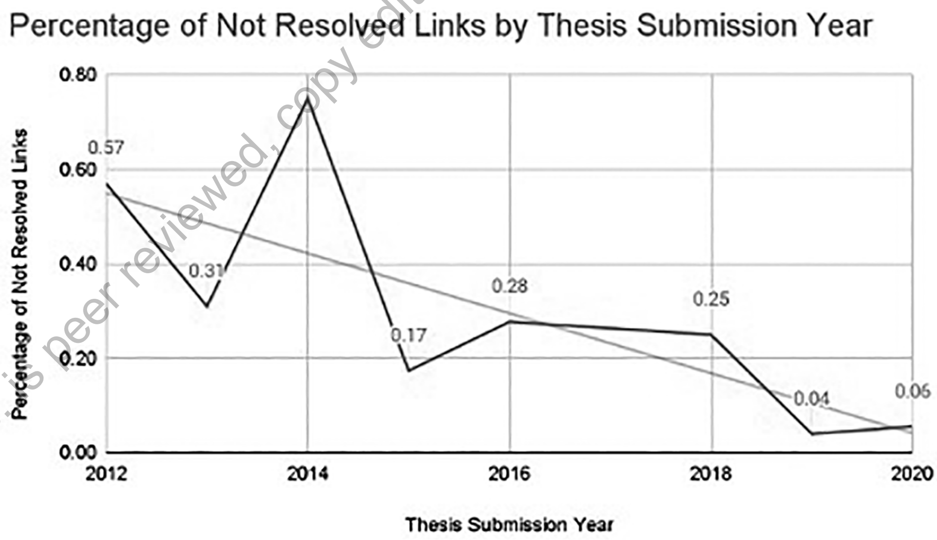

Figure 4. Percentage of links that filed to resolve, by thesis submission year.

These findings echo common assertions in the literature that reference rot will increase over time, with websites increasingly likely to become changed or lost.[47] However, it is striking how quickly the links break, with theses filed within three years of the researchers' analysis showing a 25 percent non-resolution rate.

. . . reference rot will increase over time, with websites increasingly likely to become changed or lost.

Breaks by Source Type

The researchers also analyzed whether the type of material referenced correlated with link resolution and functionality. Researchers coded links as referring to either scholarly papers or Web at large materials. As previously described, in this study, the term "scholarly papers" refers to formally hosted digital journals, books, data, and other research materials with more robust preservation and access infrastructure than that of the "Web at large."[48]

Table 3.

Resource type categories

Resource type	
Scholarly papers	**Web at large**
Articles	Special interest groups
Books	Blogs
Theses/Dissertations	News stories
	Organizations
	Conference presentations
	Governmental publications of any type
	Working papers
	Data files
	Videos
	Social media posts
	Wiki entries
	Other

In the analysis of the links provided in citations, approximately 7 percent (34 of 484) of links were to scholarly papers, while 93 percent (450 of 484) of links were to Web at large materials. The researchers had expected to find a higher perfectly functional rate for scholarly papers at the beginning of their analysis. However, a significant portion of link issues in scholarly papers comes from thesis authors' use of EZproxy links.

Table 4.
Weblink functionality by resource type

Weblink functionality	Scholarly papers	Web at large
Not resolved	20.6% (7)	26.9% (121)
Diminished functionality	17.6% (6)	22% (99)
Perfectly functional	61.8% (21)	51.1% (230)

Lost Sources

In many cases where links did not resolve or had diminished functionality, it was relatively easy for the researchers, all trained librarians with master's degrees in library and information science, to find the intended source, as described in the citation, by performing searches online or in library databases. The researchers found, however, that librarians could not locate the materials cited within 15 minutes in 80 of 484 links (16.5 percent). While it may be possible to track down materials in the longer term, such as by requesting broadcast video clips from international media companies, the process for doing so would be laborious and possibly unfruitful. Such sources are effectively lost to the average academic scholar. In other cases of loss, the materials seemed irretrievable. There were a number of broken social media posts where the materials in the citation were unlikely to be archived; these were extremely difficult, if not impossible, to retrieve, even with extensive and lengthy searching. For example, where authors had cited Tumblr posts that were now broken, there was not a time-effective way to discover what the posts had captured. In other cases, YouTube videos had been removed, and there was not enough information in the citation to know how one might find the videos again.

While it may be possible to track down materials in the longer term, such as by requesting broadcast video clips from international media companies, the process for doing so would be laborious and possibly unfruitful.

Web at large resources represented all the references that were considered irretrievably lost; no scholarly papers were identified as lost. This is perhaps unsurprising. Although authors may make citation errors or EZproxy links to scholarly papers may break, a considerable publishing and archiving apparatus stands behind scholarly publications. In Web at large resources, there are no guarantees of such mechanisms. Loss in Web at large resources becomes particularly problematic for authors when whole websites disappear. In one thesis, a student referenced different pages on a single online resource—the Television Without Pity online forum—11 times. Per Wikipedia, Television Without Pity was taken offline in 2014; archives of the site were posted online for a period, then disappeared again in 2021, and the site's capture in the Wayback Machine is incomplete.[49] The disappearance of Television Without Pity had an outsized effect on both the validity of that single thesis and the health of this study's overall corpus; references to this site represent 13.8 percent (11/80) of resources this study identified as "lost" and 2.3 percent (11/484) of the overall set.

Web at large resources represented all the references that were considered irretrievably lost; no scholarly papers were identified as lost.

Limitations

This study has a number of limitations. The researchers, in prioritizing close attention to a limited corpus, examined only a small number of master's-level theses from a single department at Texas A&M. Additionally, the researchers examined only citations that included URLs. Several citation styles do not require links and, indeed, advise authors not to include URLs because of their changeable nature. Students may have not included links for materials they assumed were also available in print, such as scholarly articles and books. The Web-based resources referenced in the corpus of theses are therefore (1) likely an undercount of the actual number of Web-based resources consulted by each thesis author; and (2) not analyzed in the larger context of citations that do not include URLs. Findings are not generalizable across all theses and dissertations but are instead intended to inform stewards of these collections, graduate student thesis writers, and administrators, instructors, and librarians. When citing a source, authors do not necessarily need to know the specific likelihood of content drift or link rot. Rather, they need to realize that it can and will happen, and to consider approaches that mitigate against link rot or content drift. Despite the limitations of the size of the sample analysis, overall rates of reference rot suggest that larger samples from different disciplines may have similar patterns.

When citing a source, authors do not necessarily need to know the specific likelihood of content drift or link rot. Rather, they need to realize that it can and will happen, and to consider approaches that mitigate against link rot or content drift.

Categorizing the different ways that links break presented challenges. For example, the researchers looked at URL redirects, code that sends browsers attempting to open a Web page to a different URL. A successful redirect, representing an actively managed site, acts as a forwarding device and will take the reader to a page at a new address that contains the content in the citation. Other redirects were more problematic. One redirected link for the English-language landing page for the Eurasian Economic Commission worked for one researcher but not another. Some redirected pages resolved to general landing pages that no longer contained the specific information cited. The researchers decided to code as having diminished functionality only redirects that significantly impaired the user from reaching the intended material.

Discussion

Examining a small corpus of master's theses in performance studies (2012–2020), this study considers whether Web-based materials linked as evidence and cited by the thesis authors are findable and verifiable and what the barriers are to accessing and verifying the Web-based materials cited. In light of these topics, this study further examines how an analysis of lost evidence in master's theses can guide interventions.

Reference Rot, Ephemerality, and Web at Large Materials

This study revealed much higher levels of reference rot than the researchers had anticipated, with nearly half the weblinks included in citations showing evidence of reference rot. This is an alarming rate of rot, particularly given that the analysis pool was limited to theses less than nine years old at the time of the coding study. Given the clear influence of time on the integrity of the links, we can infer that reference rot problems will be more pronounced in older theses and that they will worsen over time. More problematic than the general reference rot over time, however, is the total loss not just of the link integrity but also of the xscited material itself. Web at large materials are at considerable risk of becoming lost, possibly forever.

This study revealed much higher levels of reference rot than the researchers had anticipated, with nearly half the weblinks included in citations showing evidence of reference rot.

Ephemerality and Reproducibility: The Case of Performance Studies

As a field, performance studies is acutely aware of ephemerality, context, and perspective. As Richard Schechner explains: "In performance studies, texts, architecture, visual arts, or any other item or artifact of art or culture are . . . studied 'as' performances. That is, they are regarded as practices, events, and behaviors, not as 'objects' or 'things.'"[50]

Performance studies' emphasis on experiential research, its tradition of studying ephemeral theater and dance performances, and its tendency to treat artifacts as situated and interpreted and to interrogate them as such lend the discipline some authority in preparing for concerns about lost access to Web resources referenced in citations. But this field, too, is concerned about such issues as reproducibility and replicability, issues that have seized social psychology and medicine, in what psychologist Simine Vazire has termed "the credibility revolution."[51]

As an interdisciplinary field, performance studies is a magpie, borrowing from other disciplinary traditions.[52] Some performance studies scholars rely on the restudy, a tradition drawn from anthropology and ethnography. As Alan Merriam explains in *The Anthropology of Music*, in restudies "an area or a problem is checked a second time either by the same or by a different investigator." The goals include aspects of replicability—or reevaluating the initial work—or of being able to substantively build on the initial work as "a baseline against which to measure and evaluate change."[53] Evidence and documentation are central to this goal. Restudies can be thwarted by scholarship that relies on changing or prone-to-loss Web at large references.

YouTube, which has emerged as a tremendous source for many researchers, including those in performance studies, is worth a special mention in this discussion. We can expect greater reliance on YouTube and other online communities and repositories, particularly as sites for performance became unavailable during the COVID-19 pandemic. As film archivist Rick Prelinger has explained, "YouTube is not itself an archive. Preservation is neither its mission nor its practice."[54] But, Prelinger elaborates: "YouTube might as well be an archive . . . in the public mind it is not simply an archive but an ideal form of archive."[55] The challenges of reliance on ephemeral recordings, encountered as externally hosted digital media, can be expected to grow. Interventions in reference rot must be attentive to audiovisual material, which is especially challenging to capture in Mementos. Such material may also be subject to more complex rights considerations and, in the case of YouTube, may be posted, maintained, and withdrawn by registered users of a centralized site.

While the results of this study are not generalizable to other disciplines, they suggest implications for other fields. Performance studies can, in a sense, provide a framework for considering the many types of materials employed for study in other areas of scholarship. The Web artifacts of performance, including texts, videos, blogs, news reports, and webpages, are integral data points for many scholars in the humanities and social sciences. A video of a violin performance cited in a performance studies thesis could just as easily appear in a history, ethnomusicology, literature, or economics thesis. Performance studies is a unique field, but it takes a broad view of what is considered as performance. This study invites librarians to consider how the types of ephemeral materials often cited by scholars of performance studies may also be present in other disciplines.

Marginalization and Vulnerability

Efforts such as Archiving the Black Web (https://www.archivingtheblackweb.org/, archived at https://perma.cc/AB7M-YPPR) and the Internet Archive's Community Webs (https://communitywebs.archive-it.org/, archived at https://perma.cc/DFW6-6QTK) have demonstrated the need for Web archiving dedicated to collecting and preserving Black history and underrepresented cultural heritage online. The researchers theorized that Web at large resources authored by or documenting marginalized populations would be particularly vulnerable to loss or change. This concern was heightened by the theses themselves, most of which (see Appendix B) focused on populations that would be considered marginalized by their national origin, race, ethnicity, age, religious affiliation, incarceration, sexual orientation, or gender identity. The researchers initially considered coding Web-based resources according to the WEBCCCHAM framework for naming normativity. The framework was first suggested by Hope Olson, adapted and extended by Marika Cifor, and invoked by Michelle Caswell to describe the white, ethnically European, bourgeois, cis, Christian, citizen, heterosexual, able-bodied, male mainstream that "masquerade[s] as unnamed universals."[56] But, as the researchers attempted to code for normativity, it was unclear what level of analysis would best address this research question: was it more important, when analyzing a cited resource, to consider the author, topic, venue, or argument? Ultimately, the researchers limited their analysis to simply coding for formal qualities of Web resources that might be related to marginalization, such as identification of works in languages other than English or hosted outside the United States. However, given the small sample size and the constraints of quantification for such a corpus, observations were limited and inconclusive. Future research is needed to develop a clear methodology of assessment to gauge a correlation between marginalization and the vulnerability of Web at large resources in theses.

Sociotechnical Interventions to Prevent Loss

Interventions that address reference rot and evidentiary loss will necessarily involve both human and technical components. A number of tools have become available over the past two decades to address and combat reference rot. The use of persistent identifiers such as DOIs and Handles represents a formal, institutionally supported approach to ensuring that scholarly artifacts, such as books and journal articles, are fixed and findable on the Web. The human and technical infrastructure around persistent identifiers has been instrumental in assuring researchers that digital journals, books, and other scholarly publications are as trustworthy and reliable as their print precursors. As Shawn Jones, Martin Klein, and Herbert Van de Sompel explain, the persistent identifier approach functions well for artifacts that are not designed to change, with invested custodians who are committed to ensuring their ongoing accessibility and integrity.[57] Persistent identifier approaches are thus not considered either feasible or appropriate for more ephemeral, changeable Web at large resources. Such materials are better suited for capture as snapshots, or Mementos, in large-scale Web archives.

A number of tools have become available over the past two decades to address and combat reference rot.

It is beyond the scope of this study to provide an overview of relevant Web archiving tools and practices. The researchers are, however, currently designing user studies with a goal of exploring the feasibility of building URL extractors and checkers, as well as tools to generate permalinks and Mementos, into existing open source thesis submission and management tools. Beyond the tools that can be developed to prevent loss, education about the importance of stopping loss is crucial for changing author behavior. Such education must reach many audiences, including thesis authors, graduate administrators, instructors and advisers, collection managers, and other campus stakeholders. For lasting impact, communities of practice and open source software communities must also be included.

Author Practices

The researchers observed and categorized practices that affected evidentiary loss. Here, the researchers briefly review actions taken by thesis authors associated with preserved or diminished functionality.

Practices to Preserve Functionality

Some student thesis authors demonstrate an awareness of the potential ephemerality of their objects of study, including Web-based resources. Authors used several different methods to address or mitigate against the loss of ephemeral evidence in their theses.

Encapsulated Evidence

The importance of documentation within performance studies is underscored by the practice of encapsulating evidence within the text itself. Graduate authors in this study have included screenshots or transcribed texts—including tweets and interviews—into their theses, rather than pointing to these materials online or providing them as separate files. These practices not only indicate researchers' awareness of ephemerality and efforts to fix and capture relevant evidence but also align with emergent best practices for preservation management of digital publications, such as New York University Libraries' recommended strategies for identifying and preserving "third-party media content that is a core intellectual component of the work."[58]

The Use of Permalinks

Four thesis authors included DOIs when citing digitally hosted journal articles. Another four authors included Internet Archive Wayback Machine permalinks when referencing a variety of ephemeral online content—news stories, personal or special websites, and a Tumblr page. By including Wayback Machine links, these authors ensured that their reader could access the version of the material that they referenced when writing their thesis. Authors were not systematic in using Wayback Machine links—only five are included in the overall corpus, distributed among four authors. An examination of the resources referenced using Wayback Machine permalinks shows that these materials had already disappeared from the live Web by the time thesis authors used them (that is, the authors did not mint Wayback Machine permalinks to ensure that materials they consulted persisted; rather, they used Wayback Machine to access Web resources that had already been taken offline). The inclusion of Wayback Machine permalinks signals an awareness of the instability of the live Web, if not necessarily a familiarity with proactive practices for including Web at large resources in the Internet Archive.

Awareness of Instability

Students' awareness of the instability of their sources—if not a parallel awareness of the potential of Web archiving to mitigate against this instability—also appears in the theses. Alexandra Simpson references "evidence . . . found in the original, though now removed, posts" on Tumblr in her thesis.[59] Danielle Sather reflects: "Performance is ephemeral. Those three words have been ringing painfully in my ears since my first semester in graduate school. 'Performance is ephemeral.' It cannot be archived. It cannot be captured. It cannot be contained. Performance cannot be lived twice in the same exact way. It is even more fluid than water."[60] Sather elaborates that fluidity extends to "the archived text, which always undergoes interpretation through the process of reading. Moreover, as time passes, information expands and technology grows, and our world context changes as do perspectives and processes of interpretation. Text is no less in danger of losing its original meaning than performance."[61] Thesis authors writing about lack of fixity as they study the Web will be especially well-served by outreach that introduces Web archiving tools. In addition to their promise of longer-horizon persistence, Web archiving tools and practices will help ensure that the evidence is not lost.

Practices Associated with Diminished Functionality and Loss

There were consistent author practices that led to links that did not resolve. Additionally, in some cases authors provided links that are at considerable risk of diminished functionality due to core content drift and loss of the resource itself.

Author Errors and Typos

A surprising finding from this study's examination of graduate student citation practices was the frequency of author error in generating URLs. The URLs included in citations contained typographical errors, pointed to overarching websites rather than specific resources, and included personalized code that rendered them unusable to others. As an example of the latter, several authors included URLs that failed to resolve because they included EZproxy. EZproxy is authentication and access software administered by the nonprofit library organization OCLC and widely implemented by academic libraries. It was designed to allow users to access subscription resources via IP (Internet Protocol) authentication or login.[62] URLs that include EZproxy, however, are unusable outside the institutional subscription. A student author citing a journal article accessed through the JSTOR database, for example, included a URL of http://www.jstor.org.lib-ezproxy.tamu.edu:2048/stable/27510726 rather than https://www.jstor.org/stable/27510726. Curiously, these errors had the effect of thwarting immediate access to Web-based resources such as online journals, which are typically considered safe from reference rot. The challenge in this instance is not rot or drift but confusion over identifying and including appropriate URLs, a confusion amplified by formal scholarly publications posted online with multiple associated URLs and URIs.

Reliance on Fragile Resources

Not all breakages happened evenly across theses. In the most extreme case, one thesis had only 4 perfectly functional links out of 45 total. A number of factors likely contributed to this especially large gap in cited link functionality. The thesis was older, dating to 2012, and the author heavily relied on two now-defunct sources—the Television Without Pity Web forum and the Baseline New York Times movie database—as well as a number of YouTube videos that had been removed at the time of coding. But seemingly ephemeral Web content can be challenging to predict—two of the four perfectly functional links cited by the author were to a celebrity's Twitter account, which persisted beyond the death of the tweeting celebrity.

Informing Interventions

From an outreach perspective, the researchers have found it valuable to communicate to student authors that links break increasingly over time, often fairly quickly, and some links can be expected to break faster than others. Initial outreach efforts suggest the importance of including data when communicating. More tailored data, specific to discipline or type of publication (for example, thesis, article, or court case), can be compelling to students as well as faculty advisers. Large-scale studies on link rot draw attention to the threat that rot poses to the integrity of the scholarly record. Closer studies of electronic theses and dissertations (ETDs), coupled with resources, can help empower graduate students to rot-proof their ETD references.

. . . links break increasingly over time, often fairly quickly, and some links can be expected to break faster than others.

This study demonstrates that, when designing tools and workflows to address reference rot, librarians may anticipate that authors will make mistakes when including URLs. These mistakes include typos, references to the wrong item, references to containers rather than items, inaccurate or unusable links (for example, EZproxy rather than DOIs or permalinks). Some libraries, aware of researcher confusion over such terms as Handles, DOIs, and URIs, have adopted more user-friendly language such as "citable link" in their own platforms to assist researchers who are selecting URIs and URLs for inclusion in citations.[63] Finally, while ETD management can and should be done at the institutional level, there is always a place for disciplinary standards and norms. Performance studies may have a different approach to evidentiary capture than does agriculture, civil engineering, or education.

Conclusion

Academic librarians have an interest in addressing the preservation issues raised by pervasive link rot within electronic theses and dissertations. Theses are held in and published through libraries' collections as part of institutional research output, serving as evidence of students' scholarly preparedness. Librarians both create and maintain ETD collections and instruct emerging and established scholars about how best to communicate and disseminate their texts. The issue of reference rot in ETDs can—and should—galvanize us as stewards and as educators. Massicotte and Botter make a compelling case for libraries to intervene in ETD link rot. Describing the "new universe of responsibility" for preservation that librarians incurred through the development of

ETD programs, they argue: "Since an ETD comprises a unique form of scholarly output produced by universities, and simultaneously satisfies the parent institution's degree-granting apparatus, as well as reflecting its academic stature on the international stage, the presence of reference rot in this body of literature is of particular concern and worthy of immediate attention."[64]

This study is unique in its attention to a small corpus of electronic theses, which allows a closer look at the types of loss incurred, as well as in its consideration of the disciplinary implications of performance studies. Working with a smaller sample set, bounded by a single department, this study re-creates a close reader's experience attempting to validate Web-based evidence referenced in the theses. It describes the challenges a reader will face in locating and verifying Web-based content, particularly less stable Web at large materials. It also reports methods that some authors have taken to mitigate against these challenges, including Web archiving their references, using permalinks, and encapsulating snapshots of Web-based evidence directly in their texts. This close look informs the article's recommendation to develop scaled approaches to intervene in and mitigate against evidentiary loss. Several potential modes of intervention in addressing reference rot are considered in the discussion, focusing on the need to raise awareness of this concern with audiences such as graduate students authoring ETDs, graduate advisers overseeing this work, developers of potential technical solutions, and curators responsible for stewardship of ETD collections.

This study's findings point to the need for and importance of sociotechnical interventions to serve urgent preservation needs related to Web-based references. It is exceptionally important that students citing Web-at-large resources know to pay attention to Web archiving and reference rot. While disciplines that do not rely on Web-at-large resources for their evidence may pose a less immediate need for intervention, the ephemerality of Web-based material is not confined to performance studies alone. By developing an awareness of discipline-specific practices and more closely examining both the incidence of and implications for reference rot, librarians will be better positioned to guide emerging scholars and graduate students through the process of incorporating techniques for preserving evidence and to build tools, systems, and workflows to support preservation of fragile Web-based content referenced in electronic theses.

It is exceptionally important that students citing Web-at-large resources know to pay attention to Web archiving and reference rot.

Acknowledgments

The authors gratefully acknowledge *portal* Editor Ellysa Stern Cahoy and our anonymous peer reviewers. Thank you to Leonardo Cardoso, our Performance Studies colleague at Texas A&M University, for encouraging and responding to early findings. Thank you to Sam Abrams, for reading and responding to an article draft, and to John Anders, for assisting with statistical analysis. Thank you to Martin Klein, for foundational research on reference rot and for cultivating scholarship that promotes the preservation of research artifacts.

Appendix A

Coding Manual

Category	Coded answer or free text
Coder	Name
Thesis author	Copied from thesis
Thesis title	Copied from thesis
Year	Copied from thesis
Full citation	Copied from thesis
Link	Copied from thesis
Resource type, to be filled using controlled term list	SIG, ARTICLE, BLOG, ORG, BOOK, NEWS, CONF, GOV, WORKING PAPER, DATA, VIDEO, THESIS, SOCIAL, WIKI, OTHER
Domain	.com, .gov, .edu, etc.
Link in the spreadsheet matches original document/citation	Y/N
If typo, paste corrected link used for testing	Corrected link
Does the link resolve? (i.e., when you click on it or paste the text into a browser, a resource comes up)	Y / N Specialized error pages, including YouTube "Video Unavailable," are considered to be not resolved, though the link remains active.
If link resolves: Flag if everything looks perfectly functional—student has linked the appropriate resource, link resolves to correct page, no evidence of content drift in main content.	X or can't assess or leave blank
If link resolves: Flag if there is evidence of significant, central content drift; describe in notes.	CORE
If link resolves: Flag if there is evidence of minor content drift; describe in column.	free text
If link resolves: Flag if you observe a redirect (e.g., the landing page URL does not match the original URL).	X Only erroneous redirects are flagged as associated with diminished functionality.
If link resolves: Flag for author error in creating the original link (beyond typos).	X Ex: link to wrong YouTube video; link to jstor.org or other nonspecific landing page.

Appendix A, continued.

Category	Coded answer or free text
If link resolves: Code if link resolves to a resource with blocked or restricted access.	Ex: paywall, account required, private group, other. Only significantly restricted access, e.g., requiring private group membership to access, is flagged as associated with diminished functionality.
If link does *not* resolve: Code the failure to resolve.	Ex: 404, time-out, error in link, not designed to resolve, page not found, expired domain, EZproxy, other
If the resource is missing, can it be located via search?	Y / N
Flag permalinks or persistent identifiers	X Ex: Internet Archive, DOI, ARKs [archival resource keys], Handles, PURLs [persistent URLs], JStor permalink, perma.cc, etc.
Describe sites/resources in languages other than English	Free text
Interesting about the link/Notes	Free text
Interesting about links in the overall thesis	Free text
Date of coding	Date
Browser	Browser used
Coder	Name
Date of confirming coordinated response	Date

Appendix B

Theses Included in Corpus

Date	Author	Title	Handle
2012	Micu	Humor Alert: Muslim and Arab Stand-Up Comedy in Post-9/11 United States	http://hdl.handle.net/1969.1/ETD TAMU-2012-05-10988
2012	Nasir	"Dusty Muffins": Senior Women's Performance of Sexuality	http://hdl.handle.net/1969.1/ETD-TAMU-2012-08-11493
2012	Piepenbrink	Rules of Engagement: Performance and Identity in the War on Terror	http://hdl.handle.net/1969.1/ETD-TAMU-2012-05-10899
2012	Sayre	Queer Utopian Performance at Texas A&M University	http://hdl.handle.net/1969.1/ETD-TAMU-2012-05-10892
2013	Powell	Heavy Metal Humor: Reconsidering Carnival in Heavy Metal Culture	http://hdl.handle.net/1969.1/151037
2013	Roby	Crust Punk: Apocalyptic Rhetoric and Dystopian Performatives	http://hdl.handle.net/1969.1/151027
2014	Kalashnikova	The Carmen-Suite: Maya Plisetskaya Challenging Soviet Culture and Policy	http://hdl.handle.net/1969.1/152783
2015	Adamy	Diva Performativity: Female Body and Voice through Euro-Classical Vocal Pedagogy	http://hdl.handle.net/1969.1/155241
2015	Hardi	Old Army Fight: The Intersection of War Memorials and Veterans at Texas A&M University	http://hdl.handle.net/1969.1/155253
2015	Johnson	Dance Floor Reverberations: Affect and Experience in Contemporary Electronic Dance Music	http://hdl.handle.net/1969.1/155447
2015	Liddell	The Public Body: Individual Tactics and Activist Interventions on the Street in Delhi, India	http://hdl.handle.net/1969.1/155478
2015	Simpson	Shipping the Margin to the Centre: Excavating Tumblr;	

		Filling In The Self	http://hdl.handle.net/1969.1/155239
2016	Elder	From Genderfuck to Nonbinary: Negotiating Gender in Performance	http://hdl.handle.net/1969.1/157097
2016	Lee	Beyond the Comfort Zone: Female Gugak Musicians Responding to 21st Century Korea	http://hdl.handle.net/1969.1/157722
2016	Sather	Liminal Showers: A Ritual Performance in Prisoner Advocacy	http://hdl.handle.net/1969.1/157096
2016	Valle	"El Baile del Pueblo [the people's dance]:" A 60-Year Legacy of Performing a History of Cubans of African Descent through Casino Salsa	http://hdl.handle.net/1969.1/157091
2018	Adeyemo	What's in a Construct? Perceptions of African Americans in Houston's Nigerian Central	http://hdl.handle.net/1969.1/174304
2019	Aguilar	An Alternative View on the Machismo Culture of Brownsville, TX	http://hdl.handle.net/1969.1/186173
2019	Bendana Rivas	Creating an Interactive/Immersive Classical Music Concert	http://hdl.handle.net/1969.1/186162
2019	Garcia	"Molotov Cocktail Party": Protest and Humor in the Music of Molotov	http://hdl.handle.net/1969.1/186263
2019	Tanska	The Voice of Ukraine: Mediating Nationalism and Cosmopolitanism	http://hdl.handle.net/1969.1/187584
2020	Adinku	The Afro-Diasporic Imagining of African Power, Identity, and Subjection Through Dress	https://hdl.handle.net/1969.1/189540

Sarah Potvin is an associate professor at Texas A&M University in College Station. She may be reached by email at: spotvin@tamu.edu, and her ORCID ID is 0000-0001-6025-5054.

Tina Budzise-Weaver is an associate professor at Texas A&M University in College Station. She may be reached by email at: tmweaver@tamu.edu.

Kathy Christie Anders is an associate professor at Texas A&M University in College Station. She may be reached by email at: kanders@tamu.edu.

Notes

1. See, for example, Williams College LibGuide, "Citing Your Sources," 2023, https://libguides.williams.edu/citing, archived at https://perma.cc/SM9R-JHN2.
2. Jill Lepore, "Can the Internet be Archived?" *New Yorker* (January 19, 2015), https://www.newyorker.com/magazine/2015/01/26/cobweb, archived at https://perma.cc/64DL-NZQK.
3. Lepore, "Can the Internet be Archived?" Lepore observes a distinction between obviously lost Web-based work—marked with 404 errors or "page not found"—and changeable, drifting resources. As will be discussed in the literature review, scholarship in Web archiving and Internet preservation has formalized the distinction, particularly in the context of citations or references that include links (URLs/URIs) to Web-based content.
4. Jonathan L. Zittrain, Kendra Albert, and Lawrence Lessig, "Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations," Social Science Research Network (2013), https://doi.org/10.2139/ssrn.2329161.
5. Barney G. Glaser and Anselm L. Strauss, *The Discovery of Grounded Theory: Strategies for Qualitative Research* (Chicago: Aldine, 1999).
6. 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," *PLoS ONE* 9, 12 (2014), https://doi.org/10.1371/journal.pone.0115253. Klein and his coauthors introduce the concept of "web at large resources" in their article on reference rot in scholarly journal articles, defining the category to include "a wide range of web content, distinct from journal articles." This study adopts their concept of "web at large."
7. Klein, Van de Sompel, Sanderson, Shankar, Balakireva, Zhou, and Tobin, "Scholarly Context Not Found."
8. In addition to the studies described in more detail below, see Ali Sadat-Moosavi, Alireza Isfandyari-Moghaddam, and Oranus Tajeddini, "Accessibility of Online Resources Cited in Scholarly LIS Journals: A Study of Emerald ISI-Ranked Journals," *Aslib Proceedings* 64, 2 (2012): 178–92, https://doi.org/10.1108/00012531211215196.
9. See Diomidis Spinellis, "The Decay and Failures of Web References: Attempting to Determine How Quickly Archival Information Becomes Outdated," *Communications of the ACM* [Association for Computing Machinery] 46, 1 (2003): 71–77, https://www.spinellis.gr/pubs/jrnl/2003-CACM-URLcite/html/urlcite.pdf, archived at https://perma.cc/XN6F-ZN4C; Mary F. Casserly and James E. Bird, "Web Citation Availability," *Library Resources & Technical Services* 52, 1 (2011): 42–53, https://doi.org/10.5860/lrts.52n1.42; Robert Sanderson, Mark Phillips, and Herbert Van de Sompel, "Analyzing the Persistence of Referenced Web Resources with Memento," *ArXiv* (2011), http://arxiv.org/abs/1105.3459.
10. Wallace Koehler, "Web Page Change and Persistence—A Four-Year Longitudinal Study," *Journal of the American Society for Information Science and Technology* 53, 2 (2002): 162–71, https://doi.org/10.1002/asi.10018.
11. Steve Lawrence, D. M. Pennock, G. W. Flake, R. Krovetz, F. M. Coetzee, E. Glover, F. A. Nielsen, A. Kruger, and C. L. Giles, "Persistence of Web References in Scientific Research," *Computer* 34, 2 (February 2001): 26–31, https://doi.org/10.1109/2.901164.
12. Lawrence, Pennock, Flake, Krovetz, Coetzee, Glover, Nielsen, Kruger, and Giles, "Persistence of Web References in Scientific Research."
13. Lawrence, Pennock, Flake, Krovetz, Coetzee, Glover, Nielsen, Kruger, and Giles, "Persistence of Web References in Scientific Research."
14. Daniela V. Dimitrova and Michael Bugeja, "Consider the Source: Predictors of Online Citation Permanence in Communication Journals," *portal: Libraries and the Academy* 6, 3 (2006): 269–83, https://doi.org/10.1353/pla.2006.0032.
15. Edmund Russell and Jennifer Kane, "The Missing Link: Assessing the Reliability of Internet Citations in History Journals," *Technology and Culture* 49, 2 (2008): 420–29, https://doi.org/10.1353/tech.0.0028.

16. Sanderson, Phillips, and Van de Sompel, "Analyzing the Persistence of Referenced Web Resources with Memento."
17. Sanderson, Phillips, and Van de Sompel, "Analyzing the Persistence of Referenced Web Resources with Memento."
18. Klein, Van de Sompel, Sanderson, Shankar, Balakireva, Zhou, and Tobin, "Scholarly Context Not Found."
19. Michael Bugeja and Daniela V. Dimitrova, "Exploring the Half-Life of Internet Footnotes," *Iowa Journal of Communication* 37, 1 (2005): 77–86.
20. Bugeja and Dimitrova, "Exploring the Half-Life of Internet Footnotes."
21. Bugeja and Dimitrova, "Exploring the Half-Life of Internet Footnotes."
22. Shawn M. Jones, Martin Klein, and Herbert Van de Sompel, "Robustifying Links to Combat Reference Rot," *Code4Lib Journal* 50 (2021), https://journal.code4lib.org/articles/15509.
23. Russell and Kane, "The Missing Link."
24. Lawrence, Pennock, Flake, Krovetz, Coetzee, Glover, Nielsen, Kruger, and Giles, "Persistence of Web References in Scientific Research."
25. Klein, Van de Sompel, Sanderson, Shankar, Balakireva, Zhou, and Tobin, "Scholarly Context Not Found."
26. Klein, Van de Sompel, Sanderson, Shankar, Balakireva, Zhou, and Tobin, "Scholarly Context Not Found."
27. Zittrain, Albert, and Lessig, "Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations."
28. John Bowers, Clare Stanton, and Jonathan Zittrain, "What the Ephemerality of the Web Means for Your Hyperlinks," *Columbia Journalism Review* (May 21, 2021), https://www.cjr.org/analysis/linkrot-content-drift-new-york-times.php, archived at https://perma.cc/TP4L-MBRB.
29. Bowers, Stanton, and Zittrain, "What the Ephemerality of the Web Means for Your Hyperlinks."
30. Analyses of ETD (electronic theses and dissertations) reference rot include Mark Edward Phillips, Daniel Gelaw Alemneh, and Brenda Reyes Ayala, "Analysis of URL References in ETDs: A Case Study at the University of North Texas," *Library Management* 35, 4–5 (2014): 293–307, https://doi.org/10.1108/LM-08-2013-0073; and Jeffrey D. Kushkowski, "Web Citation by Graduate Students: A Comparison of Print and Electronic Theses," *portal: Libraries and the Academy* 5, 2 (2005): 259–76, https://doi.org/10.1353/pla.2005.0028.
31. Mia Massicotte and Kathleen Botter, "Reference Rot in the Repository: A Case Study of Electronic Theses and Dissertations (ETDs) in an Academic Library," *Information Technology and Libraries* 36, 1 (2017): 11–28, https://doi.org/10.6017/ital.v36i1.9598.
32. Massicotte and Botter, "Reference Rot in the Repository."
33. Russell and Kane, "The Missing Link."
34. Massicotte and Botter, "Reference Rot in the Repository."
35. Massicotte and Botter, "Reference Rot in the Repository," 24.
36. Our dataset excluded theses that were under embargo.
37. The dataset included seven URLs with errors that coders identified as typos. For example, one thesis included the reference: "Veteranartists.org. Veteran Artits. 10 Feb. 2012. <http://veteranartitsts.org/>." The URL extracted from the thesis was corrected to http://veteranartists.org.
38. Margrit Schreier, "Content Analysis, Qualitative," in *SAGE Research Methods Foundations*, Paul Atkinson, Sara Delamont, Alexandru Cernat, Joseph W. Sakshaug, and Richard A. Williams, eds. (Thousand Oaks, CA: SAGE, 2019), https://dx.doi.org/10.4135/9781526421036753373.
39. Kimberly A. Neuendorf, "Defining Content Analysis," in *The Content Analysis Guidebook*, 2nd ed., ed. Kimberly A. Neuendorf (Los Angeles: SAGE, 2017): 1–35.
40. "List of HTTP Status Codes," Wikipedia (January 20, 2023), https://en.wikipedia.org/w/index.php?title=List_of_HTTP_status_codes&oldid=1133159407.

41. Donna Lanclos, "Making Space for the 'Irrational' Practice of Anthropology in Libraries," *Canadian Journal of Academic Librarianship* 6 (December 2020): 1–22, https://doi.org/10.33137/cjal-rcbu.v6.34621.
42. Glaser and Strauss, *The Discovery of Grounded Theory*.
43. Klein, Van de Sompel, Sanderson, Shankar, Balakireva, Zhou, and Tobin, "Scholarly Context Not Found."
44. Klein, Van de Sompel, Sanderson, Shankar, Balakireva, Zhou, and Tobin, "Scholarly Context Not Found."
45. Shawn M. Jones, Herbert Van de Sompel, Harihar Shankar, Martin Klein, Richard Tobin, and Claire Grover, "Scholarly Context Adrift: Three out of Four URI References Lead to Changed Content," *PLOS ONE* 11, 12 (2016), https://doi.org/10.1371/journal.pone.0167475.
46. Massicotte and Botter, "Reference Rot in the Repository," 22. Even as reference rot researchers consider the need to weigh the relative importance of content drift in cited resources, they may point to the many barriers to doing so. Jones, Van de Sompel, Shankar, Klein, Tobin, and Grover, whose assessment of content drift is purely quantitative, argue that establishing that "the impact of content drift may be less important for some URI references than for others" would "require surveying authors about the intentionality of their references and readers about their perception of the appropriateness of referenced content." They further assert that such an assessment would have "to involve automatically characterizing URIs as referring to, for example, a project, organization, or a tool versus specific content." See Jones, Van de Sompel, Shankar, Klein, Tobin, and Grover, "Scholarly Context Adrift."
47. For a discussion of the correlation between age and reference rot, as well as a review of an older literature examining the lifespan of websites, see Klein, Van de Sompel, Sanderson, Shankar, Balakireva, Zhou, and Tobin, "Scholarly Context Not Found."
48. Adapted from Klein, Van de Sompel, Sanderson, Shankar, Balakireva, Zhou, and Tobin, "Scholarly Context Not Found."
49. "Television Without Pity," Wikipedia, last edited November 3, 2022, https://en.wikipedia.org/w/index.php?title=Television_Without_Pity&oldid=1119715752.
50. Richard Schechner, "Foreword: Fundamentals of Performance Studies," in *Teaching Performance Studies*, ed. Nathan Stucky and Cynthia Wimmer (Carbondale: Southern Illinois University Press, 2002).
51. Simine Vazire, "Implications of the Credibility Revolution for Productivity, Creativity, and Progress," *Perspectives on Psychological Science* 13, 4 (2018): 411–17, https://doi.org/10.1177/1745691617751884.
52. Schechner, "Foreword: Fundamentals of Performance Studies."
53. Alan P. Merriam, *The Anthropology of Music* (Evanston, IL: Northwestern University Press, 1964).
54. Rick Prelinger, "The Appearance of Archives," in *The YouTube Reader*, ed. Pelle Snickars and Patrick Vonderau (Stockholm: National Library of Sweden, 2009), 268.
55. Prelinger, "The Appearance of Archives."
56. Michelle Caswell, "Dusting for Fingerprints: Introducing Feminist Standpoint Appraisal," *Journal of Critical Library and Information Studies* 3, 2, special issue on Radical Empathy in Archival Practice (2021), 7. Caswell cites Hope Olson, "Patriarchal Structures of Subject Access and Subversive Techniques for Change," *Canadian Journal of Information and Library Science* 26, 2–3 (2001): 4. Brian M. Watson has further developed this framework into WEB3CH2A2MS, incorporating allosexual, monogamous, and settler. See Brian M. Watson and Beck Schaefer, "Handicapped Has Been Cancelled: The Terminology and Logics of Disability in Cultural Heritage Institutions," *First Monday* (January 2023), https://doi.org/10.5210/fm.v28i1.12898.
57. Jones, Klein, and Van de Sompel, "Robustifying Links to Combat Reference Rot."

58. Jonathan Greenberg, Karen Hanson, and Deb Verhoff, "Guidelines for Preserving New Forms of Scholarship," NYU Libraries, 2021, https://doi.org/10.33682/221c-b2xj.
59. Alexandra Louise Elizabeth Simpson, "Shipping the Margin to the Centre: Excavating Tumblr; Filling in the Self," Master of Arts thesis, Performance Studies, Texas A&M University (May 2015), 40.
60. Danielle Nicole Sather, "Liminal Showers: A Ritual Performance in Prisoner Advocacy," Master of Arts thesis, Performance Studies, Texas A&M University (May 2016), 23.
61. Sather, "Liminal Showers," 24.
62. "EZproxy," OCLC, 2023, https://www.oclc.org/en/ezproxy.html, archived at https://perma.cc/KE5R-WET5**.**
63. Shelley Barba, "Question about TDL URLs," email to the Texas Digital Library DSpace Users Group Listserv (January 19, 2023). Barba reports that Texas Tech University has "changed the 'name' of the URI as it was confusing for some researchers, and instead call it a 'Citable Link'" in their DSpace institutional repository. An additional challenge, outside of the scope of this study, is what Jonathan Blaney and Judith Siefring describe as "the culture of non-citation" of digital sources, wherein researchers do not acknowledge that they have relied on Web-based materials. See Jonathan Blaney and Judith Siefring, "A Culture of Non-Citation: Assessing the Digital Impact of British History Online and the Early English Books Online Text Creation Partnership," *Digital Humanities Quarterly* 11, 1 (2017), http://www.digitalhumanities.org/dhq/vol/11/1/000282/000282.html.
64. Massicotte and Botter, "Reference Rot in the Repository," 12.