Journalist's Resource
Research on today's news topics

The growing problem of Internet "link rot" and best practices for media and online publishers

The Internet is an endlessly rich world of sites, pages and posts — until it all ends with a click and a "404 page not found" error message. While the hyperlink was conceived in the 1960s, it came into its own with the HTML protocol in 1991, and there's no doubt that the first broken link soon followed.

On its surface, the problem is simple: A once-working URL is now a goner. The root cause can be any of a half-dozen things, however, and sometimes more: Content could have been renamed, moved or deleted, or an entire site could have evaporated. Across the Web, the content, design and infrastructure of millions of sites are constantly evolving, and while that's generally good for users and the Web ecosystem as a whole, it's bad for existing links.

In its own way, the Web is also a very literal-minded creature, and all it takes is a single-character change in a URL to break a link. For example, many sites have stopped using "www," and even if their content remains the same, the original links may no longer work. The rise of CMS platforms such as WordPress and Druple have led to the fall of static HTML sites, and with each relaunch, untold thousands of links die.

Even if a core URL remains the same, many sites frequently append login information or search terms to URLs, and those are ephemeral. And as the Web has grown, the problem has been complicated by Google and other search engines that crawl the Web and archive — briefly — URLs and pages. Many work, but their long-term stability is open to question.

This phenomenon has its own name, "link rot," and it's far more than just an occasional annoyance to individual users.

Nerdy but important context

A 2013 study in *BMC Bioinformatics* looked at the lifespan of links in the scientific literature — a place where link persistence is crucial to public knowledge. The scholars, Jason Hennessey and Steven Xijin Ge of South Dakota State University, analyzed nearly 15,000 links in abstracts from Thomson Reuters' Web of Science citation index. They found that the median lifespan of Web pages was 9.3 years, and just 62% were archived. Even the websites of major corporations that should know better — including Adobe, IBM, and Intel — can be littered with broken links.

A 2014 Harvard Law School study looks at the legal implications of Internet link decay, and finds reasons for alarm. The authors, Jonathan Zittrain, Kendra Albert and Lawrence Lessig, determined that approximately 50% of the URLs in U.S. Supreme Court opinions no longer link to the original information. They also found that in a selection of legal journals published between 1999 and 2011, more than 70% of the links no longer functioned as intended. The scholars write:

[As] websites evolve, not all third parties will have a sufficient interest in preserving the links that provide backwards compatibility to those who relied upon those links. The author of the cited source may decide the argument in the source was mistaken and take it down. The website owner may decide to abandon one mode of organizing material for another. Or the organization providing the source material may change its views and "update" the original source to reflect its evolving views. In each case, the citing paper is vulnerable to footnotes that no longer support its claims. This vulnerability threatens the integrity of the resulting scholarship.

To address some of these issues, academic journals are adopting use of digital object identifiers (DOIs), which provide both persistence and traceability. But as Zittrain, Albert and Lessig point out, many people who produce content for the Web are likely to be "indifferent to the problems of posterity." The scholars' solution, supported by a broad coalition of university libraries, is perma.cc — the service takes a snapshot of a URL's content and returns a permanent link (known as a permalink) that users employ rather than the original link.

Resources exist to preserve a comprehensive history of the Web, including the Internet Archive's WayBackMachine. This service takes snapshots of entire websites over time, but the pages and data preserved aren't always consistent and comprehensive, in part because many sites are dynamic — they're built on the fly, and thus don't exist in the way that classic HTML pages do — or because they block archiving.

The Hiberlink project, a collaboration between the University of Edinburgh, the Los Alamos National Laboratory and others, is working to measure "reference rot" in online academic articles, and also to what extent Web content has been archived. A related project, Memento, has established a technical standard for accessing online content as it existed in the past.

Linking best practices

As of September 2014, the Journalist's Resource website had more than 10,000 internal and external links — and we're a tiny site compared to many. We use a WordPress extension to regularly check our links, and 10 or more can break every week — our own little universe of link rot. Many of these are caused by sites that update their design or infrastructure, PDFs that move, press releases that expire, and so forth. While there's nothing we can do about many of these changes, by carefully choosing when and how to link, we try to minimize the odds that we'll be affected. Every media organization should do this — the cost in time and resources is minimal, and the long-term benefits for both organizations and users can be substantial.

Below are some suggested linking "best practices," with an emphasis on stability and transparency rather than search-engine optimization and page ranking. The goal is to reduce the probability that outbound links will go bad, minimize your work going forward and maximize

your site's long-term utility to users. Of course, in many journalistic situations — breaking news, Twitter and live blogs, for example — the calculus necessarily changes. Speed counts, and the resources to which you link may intrinsically be ephemeral.

As time allows, however, keep the basic philosophy in mind. To paraphrase the author Michael Pollan and his famous rules for a good diet, it can be summed up in a simple mantra: “Useful links. Stable sources. Be transparent.”

1. Put in only essential links.

- Every link has the potential to go bad over time, and the more you put in, the higher the chance that one will break. If something is not central to the subject at hand and the information can be easily found with a simple Web search — institutional websites, well-known individuals, and so forth — there’s no point in linking. Doing so only increases your risk.
- For your users’ sake, don’t link too much. If you have a forest of links in your writing, it can become difficult to know what to click on — what may be behind a link, or why it’s even there. Choose your links carefully and strategically.

2. Ensure that links are clearly visible, yet don’t obscure your text.

- Single words (“told,” “study,” “reasons”) are too easy to overlook, yet linking entire phrases can be distracting and come off as overly emphatic. Link text of two to five words works well.
- The link color and style should be distinct from unlinked text, but not overshadow it completely. Keep Web accessibility for all in mind.

3. Choose linking text carefully.

- The link text should let users know what they’ll find if they click. Options include nouns with descriptive information (“2014 Yale study”), a person and an active verb (“Micah Sifry wrote”) or an interesting statistic (“97% of social scientists”). This also helps demonstrate accuracy and openness, as Oxford’s Reuters Institute put it in a 2014 report.
- Avoid structures such as "A new University of Pittsburgh study (link here) reveals the incidence of concussions among younger football players." The insertion just slows down readers, and at this point in the Internet's evolution, people know what a hyperlink looks like. That said, if this is your style, be consistent.
- Avoid stacking links tightly in a sentence — for example, "Three new studies provide a research perspective on concussions in sports." It may work for insider coverage of issues that have received extensive online attention, and you need to pack in a lot of links, but the chance for reader confusion is significant.
- To better indicate content, you can use hover text that appears when users mouse over a link. However, you should be thoughtful and consistent about this — go all in, or avoid hover text.
- A side-benefit of informative link text is that if the URL later goes bad, you have information that will simplify the search for the content — you know what to look for.

4. URL and content stability is essential — except when its ephemerality is part of the story.

- Unless you're covering breaking news, try to avoid linking to anything that might go away — personal or short-term project websites that may disappear, draft versions of documents or press releases. Fast-moving stories may require linking to content that could be taken down or modified, however, and the solution is to use website tools that monitor link validity in real time.
- Link to primary sources whenever possible, unless the secondary source is central to your coverage. For example, if you're writing about a new U.N. report, link directly to it. However, if you're dissecting how the report has been misinterpreted, you'll want to link to both the primary document and what you see as faulty coverage.
- Because of concerns about Wikipedia's accuracy, reliability and potential for bias, link to the site only when it's the subject at hand. If you do choose to link to a page, click on "cite this page" and use the "permanent link" displayed. This will lead to a snapshot of that particular version of the Wikipedia page, unaffected by subsequent edits.
- When you have a choice of sites to which to link, chose stability. For example, at Journalist's Resource we tend to favor PubMed, even if its user interface (UI) isn't the snazziest. Beyond their having 24 million citations and counting, they're part of the National Institutes of Health and are going to be around for a good long time.
- If you're linking to scholarly content, beware drafts on authors' websites. They can be open, unlike the versions on many academic journals, but they aren't the final content, and you owe it to your readers to point them to the real thing.
- If you're linking to an academic paper with a DOI number, consider using that (the domain to use is "http://doi.org/", followed by the DOI number). Persistent URLs (PURLs) also offer greater longevity, but there's some debate over the wisdom of using them for archival purposes.
- For major reports that are regularly updated — say, the State Department's work on human trafficking — link to the report landing page rather than specific documents (more on this below). This way your link will continue to work even as documents and sub-pages change. On the other hand, if you're referencing a particular statistic or fact, don't link to a generic page with content that might change. Instead, find a source that is both specific and stable.

5. Whenever possible, link to pages rather than PDFs.

- Many online resources are present in both Web page and PDF form — for example, the Reuters Institute paper, "Accuracy, Independence and Impartiality: How Legacy Media and Digital Natives Approach Standards in the Digital Age," has a landing page and is also available in a full-text PDF. Given this choice, go for the landing page. This allows users to quickly assess the content without having to download it, and also offers the option of an executive summary.
- Landing pages are generally more stable than PDFs. Because the latter are documents, they tend to be renamed or move around on websites. They can also be updated, potentially invalidating the reason for your original link, yet this won't necessarily be indicated to you or your users.

- PDF filenames are more likely to contain characters considered "unsafe" in URLs — commas, spaces, accented characters and so on. While these are automatically translated to Web-safe codes (more information below), they can impact link reliability.
- If a PDF is large, the required download can cause browsers to time out. They also depend on specific software being installed on users' computers. Yes, most people have Adobe Reader and compatibility is built into many browsers, but you can't count on that.
- PDFs can contain copyrighted material, and linking directly to them might raise legal issues (more on this below). They may also be behind paywalls.
- If you do choose to link directly to a PDF, it can be helpful to signal this to users: "A new Scholars Strategy Network post on the immigration crisis (PDF) sheds light on some persistent myths," for example. This is a matter of local style, however, and whatever approach you choose, be consistent.

6. Always look for the most compact and direct URL available, and ensure that it's clean, with no unnecessary information after the core of the URL. (This process is often referred to as "URL normalization" or "URL canonicalization.")

- If a URL contains an ".html" or other Web page extension, in most cases anything thereafter can and should be removed — it's just dead weight and could, down the line, break a link that's actually good. Verify that the slimmed URL works and if so, use that for your hyperlink.
- When there's a "?" character in a URL, check whether it and everything thereafter is mandatory for the link's functioning. For example, in http://journalistsresource.org/ballot-framing?utm_source=JR-email, everything from "?" on can and should be deleted. Note that the "?" sometimes precedes post or category information; that's fine, and you at least verified that this was required rather than, say, useless search terms or tracking codes.
- With multiple "?" characters, you can often "peel back" the URL, progressively removing unnecessary elements from the end until you get down to the smallest and most stable link possible.
- Exercise caution with links that have "%" in them — the symbol precedes codes that replace characters considered "unsafe" for URLs. For example, PDF document with the name "skating basics.pdf" would be "skating%20basics.pdf" because space characters are not valid for URLs. While URLs with codes may function, they can be unstable in the long run.
- Watch out for URLs that contain references to resources that may not be universally accessible — Google Drive, for example, or login or session information. These could work perfectly well for you, but could fail for others. If you do see such information encoded in a URL, use Google or another search engine to find a direct path to the desired content.
- Do some research before linking deeply into websites (this is called deep linking, but is distinct from the similarly named but completely different practice in mobile applications). Long URLs are intrinsically more vulnerable, and you could be inadvertently violating copyright or jumping over paywalls.

7. Avoid link-shorteners, with two exceptions.

- Bitly, TinyURL and other such services are essential for Twitter and other contexts where URL length is tightly constrained. However, for text hyperlinks they should be avoided. While they produce a compact link, it's no more stable than the underlying URL it contains — garbage in, garbage out, as the coders say. You're also dependent on a third party's maintaining your links, and that adds a layer of risk.
- Perma.cc, as described in the introduction, both produces a permalink and archives the target content for at least two years; vested organizations such as law journals and courts have the authority to make links truly permanent.
- WebCite, a project of the University of Toronto and other organizations, provides a similar service to perma.cc, but is open to all.

8. Don't link in a way that violates copyright or breaks through paywalls. While there are a lot of gray areas, do your absolute best to respect all laws and regulations.

- For academic papers, link to abstract page rather than the full-text or PDF version. For paywalled sites, you're indicating to the user where content is, but respecting copyright.
- Link to abstracts even with open journals, as they load quickly and allow users to judge whether to go for the full-text version. This also protects you down the line if a study that's initially free and accessible moves behind a paywall.
- For media sites, respect paywalls, even if you can find the direct link to full content by using a search engine.
- Exercise caution with links to YouTube and other media-sharing sites. Because videos are uploaded by users who may or may not have copyright, they can be taken down for infringement — don't assume such links are permanent.
- Avoid linking to documents on sites such as Academia.edu where the users' right to upload content isn't always clear.

9. Verify after publication and check your links at regular intervals.

- Check all your links after you publish. Some content-management systems can manipulate URLs during the production process, and the end results may not work.
- If possible, use an application or service that regularly checks the validity of your site's links. When you do find broken links, fix them promptly. Also be aware that valid links can, in a sense, be "broken" when the content you were originally pointing to changes without notice.

10. As you're building and maintaining your own blog or website, remember that other sites link to your content, and you want to keep those links alive.

- Create landing pages for all individual PDF documents, rather than just a page of links to a series of PDFs.
- If you do post PDFs, ensure that their names do not contain any unsafe characters — in particular, no commas, periods or spaces. The same goes for all your URLs, naturally.

- General-purpose pages can have generic URLs (http://journalistsresource.org/about, for example), but specific content — articles, blog posts, dated reports and so on — should have distinct, long-lived URLs.
- When content is superseded, consider keeping the original material with a note at the top pointing users to the new content.
- If you must change a page's URL, set up a quick redirect to send users from the old URL to the new one.
- When a redesign or infrastructure upgrade requires wholesale changes to your URL structure, build in ways that allow inbound links to the old URLs to connect to the right content.
- Ensure that your server can handle incorrect URLs with upper-case letters — for example, mysite.com/BigBlog should be automatically redirect to mysite.com/bigblog. (Moz.com has a great post on URL best practices.)

At this point, you've done everything you can. Your outbound links will still have the digital rug pulled out from under them from time to time, but the risks will be minimized, and you'll have a fighting chance at fixing them if they do break. For practitioners, educators and others interested in learning more about decoding URLs, we've put some sample exercises beneath the "Media/Analysis Tips" tab above.

This is a big subject, and there are of course many different approaches to effective hyperlinking. We welcome all suggestions, and they can be sent to leighton_walter_kille@hks.harvard.edu or john_wihbey@hks.harvard.edu.

———

Written by Leighton Walter Kille, July 15, 2015.

Many thanks to David Weinberger and Jonathan Zittrain of Harvard's Berkman Center for Internet & Society; Micah L. Sifry, the executive editor of the Personal Democracy Forum; Keely Wilczek and Valerie Weis of the Harvard Kennedy School Library; and Evan Horowitz of the Boston Globe *for their invaluable suggestions and input for this article.*

Keywords: technology, link rot, linkrot, resource rot, hyperlinking, uniform resource identifier (URI), uniform resource locator (URL), uniform resource name (URN), canonical URLs, SEO