Please cite as:

Levy, D., Klinner, C., Melzack, G, Stanton, K., Bourke, P., Using a Web-Archiving Service - How to ensure your cited web-references remain available and valid. University of Sydney, 2017, CC BY 2.5 AU.

USING A WEB ARCHIVING SERVICE

HOW TO ENSURE YOUR CITED WEB-REFERENCES REMAIN AVAILABLE AND VALID

Date **2017/07**

In today's electronic information age, academic authors increasingly cite online resources such as blog posts, news articles, online policies and reports in their scholarly publications. Citing such webpages, or their URLs, poses long-term accessibility concern due to the ephemeral nature of the Internet: webpages can (and do!) change or disappear[1] over time. When looking up cited web references, readers of scholarly publications might thus find content that is different from what author/s originally referenced; this is referred to as 'content drift'. Other times, readers are faced with a '404 Page Not Found' message, a phenomenon known as 'link rot'[2]. A recent Canadian study[3] for example found a 23% link rot when examining 11,437 links in 664 doctoral dissertations from 2011-2015. Older publications are likely to face even higher rates of invalid links.

Luckily, there are a few things you can do to make your cited web references more stable. The most common method is to use a web archiving service. Using a web archiving service means your web references and links are more likely to connect the reader to the content accessed at the time of writing/citing. In other words, references are less likely to "rot" or "drift" over time.

As citing authors, we have limited influence on preserving web content that we don't own. We are generally at the mercy of the information custodians who tend to adjust, move or delete their web content to keep their site(s) current and interesting. All we can do to keep web content that we don't own but want to cite intact so that our readers can still access it in years to come is to create a "representative memento" of the online material as it was at the time of citing. This can be achieved by submitting the URL of the webpage we want to cite to a web archiving service which will generate a static ('cached') copy of it and allocate it a new, unique and permanent link, also called 'persistent identifier'. We can then use this new link to the *archived* webpage rather than the ephemeral link to the original webpage for our citation purposes.

There are a range of web archives available. This guide contains a list of trusted web archiving services.

Note: The first step is to check whether a copy of the webpage in question has already been archived in one of the publicly available archives (STEP 1). If this is not the case, you can submit a request to publicly archive a copy of the website (STEP 2) or privately archive a copy of the webpage(s) you want to cite (STEP 3). Before you start, please review the limitations of web archiving, as outlined below.

LIMITATIONS OF WEB ARCHIVING

Web archiving is only suitable for pure web material available in the public domain. For documents such as articles and policies posted on the web, it is better to find an appropriate repository or other permanent publication ID, e.g. DOI, instead (see postscript below).

Archived webpages are effectively 'frozen' versions of the original web content with only limited functionality. For websites that have embedded material such as audio or visual files, or a significant interactive component, or that link to or display database content, web archiving is not a suitable solution. Options for archiving such complex webpages include:

- **Audio/visual files:** The only real option on this front is to contact the creator of the audio or video file and ask them to submit the file to a public data repository such as figshare or Dryad, so that a doi can be created for permanent citation. Otherwise, you'll just have to take the risk and cite the audio/video as is.
- **Dynamic websites:** Some interactive components of websites can be captured by web archives such as the Wayback Machine. However, these archives generally aren't able to capture dynamic pages that include forms, JavaScript, or other elements that require interaction with the original host. A snapshot of a user interacting with a site can be captured using Webrecorder, but again, there may still be elements

that can't be captured in full. If in doubt, review the archival copy carefully to ensure the information you need is captured.

- **Database content:** There's really no other option than to cite the URL of the database as is. If the result of a specific database query is being cited then it may be permissible to extract the relevant subset of data. Always check the database terms and conditions before extracting database content, or undertaking text or content mining activities.

Not all websites can be archived by web archiving services that use crawlers and web harvesting tools to ingest websites. This can be because the site is password protected, blocks web crawlers and harvesting robots, or has explicitly requested that it not be archived**.**

Web archive links can be cumbersome (very long URLs).

In some cases or jurisdictions (including in Australia where we do not have provisions for text, data or content mining, or a flexible "fair use" exception to infringement), web archiving may be interpreted as infringing copyright, or a breach of terms and conditions of use. To minimise risk, always check the website's terms and conditions of use before you archive.

STEP 1: PUBLICLY ARCHIVED COPIES

The Wayback Machine/Internet Archive, https://archive.org/web/

The Wayback Machine is a collection of digitally archived ('cached') web content created by the Internet Archive, a US-based non-profit organisation. The Internet Archive currently contains some 286 billion cached web pages. It enables users to search for and retrieve archived versions of webpages from the past 20+ years (since 1996), many of which are no longer 'live' at the original site.

The Internet Archive also provides an option to submit new webpages for archiving that can then be used as trusted, that is, stable and permanent, citations.

→ Further information on the functionality of the Wayback Machine/Internet Archive

PANDORA, http://pandora.nla.gov.au/

PANDORA, Australia's Web Archive, is a growing collection of Australian online publications, established initially by the National Library of Australia in 1996, and now built in collaboration with nine other Australian libraries and cultural collecting organisations.

PANDORA is the preferred archive to use when citing Australian websites.

Australian Government Web Archive, http://webarchive.nla.gov.au/gov/

The Australian Government Web Archive (AGWA) is a web archiving initiative of the National Library of Australia which complements the Library's long established PANDORA Archive. The AGWA is a collection of Commonwealth Government websites. Initially content collection commenced in June 2011 but earlier content sourced from other web archiving activities is being added progressively.

The AGWA is the preferred archive to use when citing Australian Government websites.

Other web archives

Many other national libraries and archives have their own web archives. These often specialise in archiving local national and government websites. For example:

- Government of Canada Web Archive
- New Zealand Web Archive
- UK Web Archive
- UK Government Web Archive
- US Library of Congress Archived Web Sites

Many state and university libraries also have web archive collections, so check the local libraries and archives that may have archived the material you need to cite.

Archive-It, https://archive-it.org/

Established in 2006, Archive-It is a member-based a web archiving service for collecting and accessing cultural heritage of digital content. Archive-It partners can collect, catalogue, and manage their collections of archived content. Content is hosted and stored at the Internet Archive data centres. A full list of Archive-It Partners and their collections can be found here. Archived content is accessible to anyone with an Internet connection (no subscription/membership required); start your search here.

STEP 2: SUBMIT A REQUEST TO PUBLICLY ARCHIVE A COPY

If the webpage you wish to cite has not already been archived in one of the public archives listed under STEP 1, you may be able to submit a request to have a public web repository archive it. Always check the website terms and conditions of use before you submit your request. If the website terms and conditions of use prohibit the webpage you wish to cite from being archived by a public web archiving service, privately archive a copy of the website instead (see STEP 3 below).
You can submit requests to archive websites to the following public web archiving services:

The Wayback Machine https://web.archive.org/
Note that the Wayback Machine is both, a collection of digitally archived web content as well as a public archiving service that individuals can use to request the preservation of a specific web resource. To submit a request to archive a website, go to https://web.archive.org/, enter the URL of the web page you want to archive in the field under 'Save Page Now' and click on the button SAVE PAGE.

WebCite, http://www.webcitation.org
WebCite was founded in 1997 and is hosted at the Centre for Global eHealth Innovation, University of Toronto, Canada.
Users (citing authors) can:

- Either manually initiate the archiving of a single cited webpage (by using either the WebCite bookmarklet or the archive page) and then manually insert the link to the archived webpage in their manuscript, or
- Upload an entire manuscript to the WebCite server via the comb page, which initiates WebCite to comb through the manuscript and archive cited URLs.

→ Background information and FAQs on WebCite

STEP 3: PRIVATELY ARCHIVING COPIES

If the webpage you wish to cite has not already been archived in one of the public archives listed under STEP 1 or STEP 2, you may want to privately archive it. In general, you can archive a copy of a legitimately accessed resource for your personal research and study under the fair dealing exceptions in the *Copyright Act 1968 (Cth)*. Always check the website terms and conditions of use before you archive. Note: It is unlikely that you can republish your archived copy to the web, or share that archived copy with other researchers.

The following services can be used to create your own privately archived copy of a website:

Webrecorder, https://webrecorder.io/
Webrecorder provides an integrated platform for creating high-fidelity, interactive web archives while browsing, sharing, and disseminating archived content. Users may try the service anonymously or login and create a permanent online archive.

Webrecorder Player, https://github.com/webrecorder/webrecorderplayer-electron
Webrecorder Player can be used to view websites archived by Webrecorder if they're downloaded and saved offline. It is available for OS X, Windows and Linux.

Web Archiving Integration Layer (WAIL), http://machawk1.github.io/wail/
Web Archiving Integration Layer (WAIL) is a graphical user interface (GUI) atop multiple web archiving tools intended to be used as an easy way for anyone to preserve and replay web pages. It is currently available for MacOS X and Windows (7+).

OTHER RECOMMENDATIONS

Apart from archiving cited webpages, other recommendations[4] for citing authors include to:

- …provide formal citations along with web citations whenever possible
- …provide enough contextual information to enable readers to search the Internet to track down invalid links.

A FEW FINAL TIPS

- Never assume webpages will remain stable and keep reflecting the exact same content you once accessed and cited.
- Before submitting a manuscript for publication, check it for cited web references. Where possible, convert ephemeral web citations into stable and permanent 'representative memos' by archiving them with one of the above archiving services, paying attention to any copyright issues.
- If in doubt, discuss with your Academic Liaison Librarian, who can provide tailored advice on information resources, and can refer you to copyright support.

POSTSCRIPT

This guide addresses the problem of citing online resources that were created and are maintained *by others* and that do not already have a stable and permanent URL.

The possibilities are different if you are the creator or owner of a citable resource, e.g. the author of a paper, book, report or scholarly blogpost. In this case, and if you haven't transferred your copyright to someone else, you have the option, and are strongly encouraged, to deposit your resource into a repository, e.g. the Sydney eScholarship Repository, where it is allocated a persistent identifier. A persistent identifier is a web address for a resource that will remain the same regardless of where the resource is located. Thus, links to the resource will continue to work even if the resource is moved. Most academic publishers today work with the persistent identifier system DOI (Digital Object Identifier). Institutional repositories, including the University of Sydney's, often use the Handle system. For a discussion on the two systems, see here.

If you are the owner of an entire website (rather than an individual document), there are other options of safeguarding your online content; see for example the National Library of Australia, Managing Web Resources for Persistent Access and Safeguarding Australia's web resources: guidelines for creators and publishers, as well as Kille L. W., The growing problem of Internet "link rot" and best practices for media and online publishers.

ACRONYMS

URL: Uniform Resource Locator
URI: Uniform Resource Identifier

FURTHER RECOMMENDED READING

Dellavalle, R. P., Hester, E. J., Heilig, L. F., Drake, A. L., Kuntzman, J. W., Graber, M., & Schilling, L. M. (2003). Going, going, gone: Lost Internet references. *Science*, *302*(5646), 787-788.

Klein M. & Van de Sompel H., Content referenced in scholarly articles is drifting, with negative effects on the integrity of the scholarly record, *LSE*, The Impact Blog, 23 Feb 2017

Jones, S. et al. Scholarly Context Adrift: Three out of Four URL references lead to changed content. *PLOS ONE*|DOI:10.1371/journal.pone.0167475 December2, 2016

Lawrence, Steve, et al., Persistence of web references in scientific research. *Computer* 34.2 (2001): 26-31.

REFERENCES

1 Dellavalle, R. P. *et al.* Going, going, gone: Lost Internet references. *Science* **302**, 787-788 (2003).

2 Jones, S. M. *et al.* Scholarly Context Adrift: Three out of Four URI References Lead to Changed Content. *PloS one* **11**, e0167475 (2016).

3 Massicotte, M. & Botter, K. Reference Rot in the Repository: A Case Study of Electronic Theses and Dissertations (ETDs) in an Academic Library. *Information Technology and Libraries (Online)* **36**, 11 (2017).

4 Lawrence, S. *et al.* Persistence of web references in scientific research. *Computer* **34**, 26-31 (2001).