WEB FRAGILITY

INTRODUCTION

The internet is part of our daily lives, but it is surprisingly fragile. The hyperlinks that connect the internet together can easily break if the pages they link to are moved, deleted, or renamed. When links break, this is called **link rot**. Even if the link still works, the content of the site can change significantly if the website is taken over by new people. Simply updating a site to keep it current can mean removing information that other pages are linked to. This is called **content drift**. Estimates place the lifespan of a web page somewhere between a few months and a few years. While this has the potential to be inconvenient on a personal level, it also poses a serious problem for people who rely on stable links to build bodies of knowledge. If links referenced in articles become inaccessible, it is impossible to verify the information that was cited, which undermines that article and its conclusions – whether that article is a piece of local news, a supreme court ruling, or the results of a medical study. Archivists and historians argue that preserving the content of the internet is important because of its historical significance, and that if we fail to do so, we could lose much of today's knowledge, culture, and art as web pages deteriorate or disappear.

WEB ARCHIVING

Web archiving is one response to the threats of link rot and content drift – a process that uses a computer program (usually a web crawler) which navigates the internet and takes snapshots of the websites it encounters in order to preserve those pages and their relationships to other pages on the internet. The snapshots are then stored in an archive in a format that can later be accessed and consulted, even if the original website no longer exists. These snapshots allow you to read and interact with them as though they were live web pages. The most well-known web archive is the Internet Archive's Wayback Machine, which began in 1996 and currently contains over 70 petabytes of data. Many institutions, including Concordia (see Concordia's Archive-It page), have ongoing web archiving projects that preserve online journalism, government webpages, and the pages of organizations related to a variety of topics.

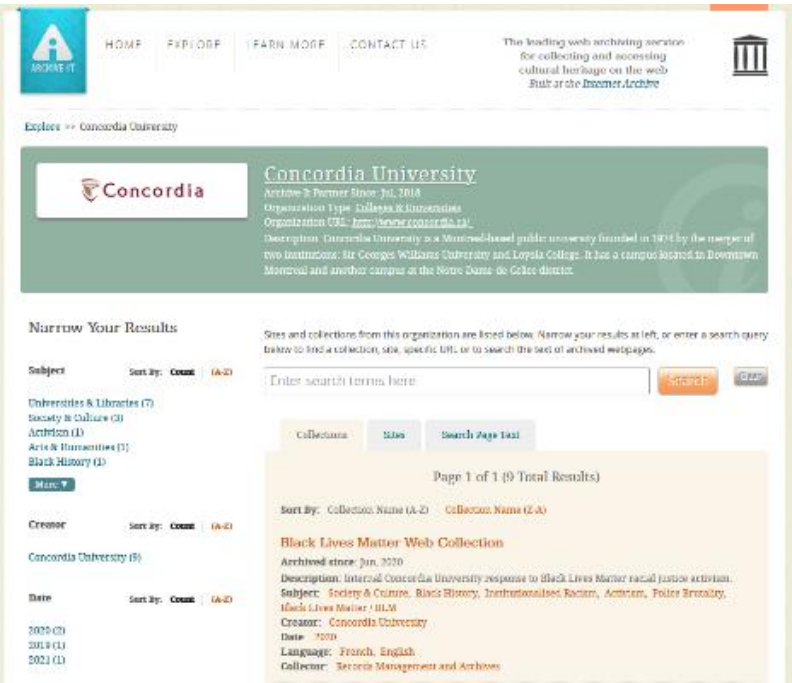


Image description: webpage Concordia University's Archive-It archive

RIGHT TO BE FORGOTTEN

While there are clearly many reasons why web archiving is an urgent and important project, there are some people who object to online content becoming part of the historical record, especially when it comes to personal information. People who have embarrassing or incriminating information about them posted online can face serious consequences in their relationships and jobs. In 2014, the European Union's Court of Justice ruled that individuals have the right to ask that outdated or inaccurate material be delisted from search engines like Google. Though that protection may seem like a good thing, some experts feel that this threatens the integrity of the public record and is a kind of selective editing of history, especially since web archives have been used as evidence in legal cases. They argue that it is important to be able to check facts about events and people's actions, particularly when it comes to holding government and public officials accountable. Canada does not yet have clear legislation about whether search engines must de-list content at the request of individuals, but a recent Federal Court of Appeal ruling in 2023 took a step in that direction, deciding that search engines are subject to existing Canadian privacy laws.

Image description: a woman with sunglasses looks into a car's rear-view mirror. Source: by Jacob Lund Photography on Noun Project. Creative Commons Attribution-Non-Commercial-No Derivatives 2.0 licence.

INTERNET INFRASTRUCTURE

Though it is easy to forget that the internet has physical parts other than the device you are using, the physical infrastructure that the internet relies on is also vulnerable. The cables and satellites that transport data (also called the internet backbone) and the server farms that store data are some of the most essential parts that keep the internet going. These are vulnerable to digital security breaches, as well as power fluctuations and physical threats ranging from shark attacks to natural disasters. When major servers or services go down, it can cause significant chunks of the internet to become inaccessible for hours at a time. If servers are sufficiently damaged or if the internet were downed for long enough, lots of valuable information could be lost. An important part of digital preservation initiatives is keeping backup copies of web archives, so that if a server is damaged or inaccessible, another in a different location will still have a copy.

Image source: Submarine cable map by OpenStreetMap with cable data from Greg Mahlknecht. Creative Commons Attribution-Share Alike 2.0 Generic licence.

CONCLUSION

While web archiving offers a partial solution to preserving the ever-changing internet, it isn't perfect. Many web crawlers struggle to capture certain media types and content created on social media platforms. There are also copyright issues with web crawlers: anyone running a web archiving project should be getting permission from the creators of websites before taking a snapshot, since the creators hold the copyright for that content. In practice, though, it is sometimes very difficult to get in touch with website creators, especially for websites that are no longer being maintained and are at the greatest risk of disappearing. Some people are advocating for changes in research and publishing processes that would require an academic or journalist to submit archived copies of any resource that they link to, ensuring that their citations remain stable and informative. The internet is still a relatively new technology, and best practices around preserving the internet's information are still being worked out. We will see new solutions emerge as the internet continues to change and grow.

QUIZ

Which of these options is an example of link rot?

a) When someone makes a typo in a link

b) When the page someone links to is deleted or moved

c) When the content of a page someone links to is significantly changed

d) All of the above

Correct answer: b

Feedback: Link rot is when the link no longer functions because it has been removed or the website navigation has been changed.

Which of these options is an example of content drift?

a) When a website is updated with an organization's new mission statement

b) When a government agency changes its website after a new government comes to power

c) When a new person takes over a website and rewrites some of the content

d) All of the above

Correct answer: d

Feedback: All of the 3 examples can be content drift or can result in content drift because the website's content has changed, making the original content no longer visible.

How does a web crawler work?

a) It makes an exact copy of the entire internet

b) It takes JPEG images of all websites it visits

c) It makes interactive copies of all websites it visits

Correct answer: c

Feedback: A web crawler archives websites by making copies of the website and some pages at a specific point in time.

Under the 2014 EU Court of Justice ruling, what kinds of information can people ask to be delisted from Google?

a) Anything that they disagree with

b) Information that damages their business

c) Information that is outdated or irrelevant

Correct answer: c

Feedback: Outdated information can usually be removed without having any impact on the historical record.

What are some negative impacts that the right to be forgotten could have?

a) Loss of important historical details

b) Loss of information about public officials

c) Loss of data needed to decide legal cases

d) All of the above

Correct answer: d

Feedback: All of the factors could have a negative impact.

Is there currently a law in Canada that protects the right to be forgotten?

a) Yes

b) No

Correction answer: b

Feedback: There is currently legislation on this topic in Europe, but not in Canada.

What kinds of issues threaten the internet's infrastructure?

a) Natural disasters

b) Cyber attacks

c) Power outages

d) All of the above

Correct answer: d

Feedback: Infrastructure is a physical system; its safety is threatened by all 3 of the factors mentioned in the question.

Why does keeping multiple copies of data keep it safer?

a) It distracts hackers by giving them multiple targets

b) If one server is damaged or unavailable, another will still have a copy of the data

c) It doesn't actually keep it safer

Correct answer: b

Feedback: Keeping multiple copies ensures that you don't lose your data when a server is down or damaged.

ACTIVITY

 Purpose: This activity gives you an opportunity to see web archiving in action.

 Task: Follow the instructions below to capture a website and compare previous versions.

 Time: This activity should take about 15 minutes to complete.

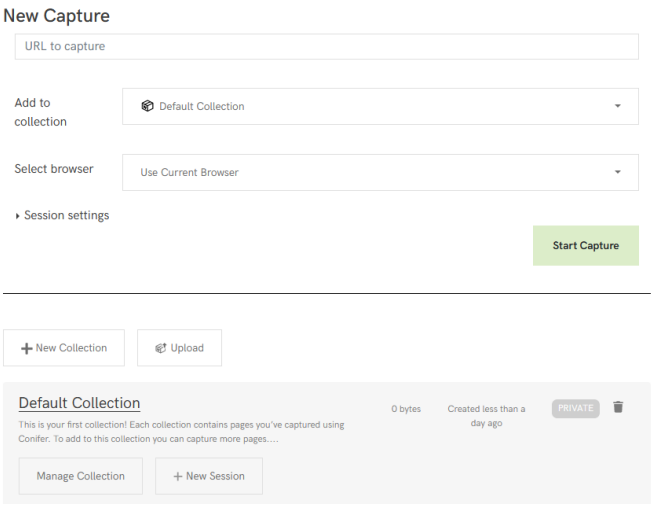
1. PICK A WEBSITE

   For example, the Concordia Library homepage: https://library.concordia.ca/

2. CREATE A CONIFER ACCOUNT

   Navigate to https://conifer.rhizome.org/, click the "Create a Free Account" button, and follow the instructions to create and confirm your account.

3. SAVE THAT SITE USING CONIFER



Login and navigate to the home page, which will show the "New Capture" screen

Paste your chosen URL into the first box

Create a collection (optional) and choose the browser you want Conifer to use to interface with the site

Click "Start Capture"

Wait for a slot and then navigate through the website, showing Conifer which pages and links you want it to preserve (note that you can switch between browsing and recording the site in the top right corner of the interface)

4. EXPLORE THE CAPTURED VERSION OF THE WEBSITE

Load the archived version of your site in Conifer by clicking the title of your collection, selecting the "Browse All" tab, and clicking the site you want to look at

Try the same navigation that you did while you were capturing

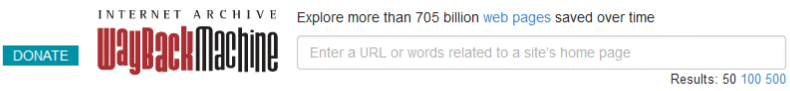
Try to click on a few things that you didn't click during the capture

What works? What doesn't work?

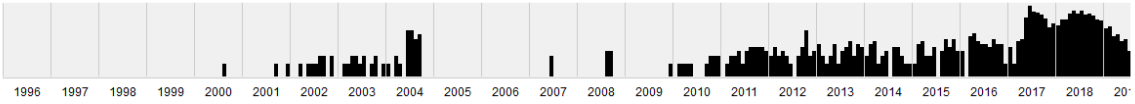
5. CHECK THE WAYBACK MACHINE FOR THE OLDEST SAVED VERSION OF THAT SITE

Navigate to https://web.archive.org

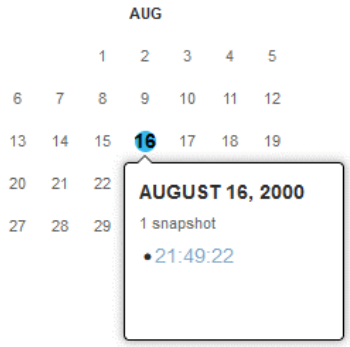
Enter your chosen URL in the search box



Scroll all the way to the left in the timeline and click on the earliest year



Then, select the earliest capture from that year



6. COMPARE IT TO THE LIVE SITE AND THE ONE YOU CAPTURED IN CONIFER

   What is useful about having access to old versions of sites?

   How are the Conifer and Wayback Machine interfaces different?

   In fact, while the Wayback Machine uses a traditional web crawler that explores a website automatically (by essentially clicking and recording all the links on a page), Conifer uses a different kind of program that records a user's session (only recording the pages that they visit and the links that they use to get there). How do you think this changes the results or your archiving?

FURTHER RESOURCES

Video: How to Use the Wayback Machine

This introduction to using the Internet Archive's Wayback Machine includes information about searching the archive and how web pages are captured.

Podcast: Right to be Forgotten

An episode of the Radiolab podcast about how journalists at one publication decide to unpublish some of the content they previously published.

Webpage: Harvard Law Review study about link rot

Article that discusses how to make legal scholarship more permanent.

REFERENCES

Zittrain, J., Albert, K., and Lessig, L. Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations. *Harvard Law Review* 2014, 127(176).

Ott, D. E. Reference Hygiene and Death on the Internet – Decay, Rot, Half-Life, Deterioration, and Corruption. *JSLS : Journal of the Society of Laparoscopic & Robotic Surgeons*. 2022; 26(1): e2021.00082.

Nguyen, H. and Weber, M. S. Internet Archives as A Tool for Research: Decay in Large Scale Archival Records. *2015 IEEE International Congress on Big Data*, New York City, NY, USA: IEEE, 724-727.

Zittrain, J., Bowers, J. and Stanton, C. The Paper of Record Meets an Ephemeral Web: An Examination of Linkrot and Content Drift within The New York Times. *SSRN* 2021, May 4.

Brügger, N. Web Archiving - Communication. *Oxford Bibliographies* 2018, August 28.

Quartz. The Early Internet Is Breaking - Here's How the World Wide Web from the 90s on Will Be Saved [Youtube Video]. 2019 July 18.

Internet Archive. How to Use the Wayback Machine [YouTube Video]. 2021 January 13.

Zittrain, J. The Internet Is Rotting. *The Atlantic* 2021 June 30.

Anat, B.-D. and Amram, A. The Internet Archive and the Socio-Technical Construction of Historical Facts. *Internet Histories* 2022; 6(1-2): 179–201.

McDonald, S. The Right to Be Forgotten: The Potential Effects on Canadian. *Dalhousie Journal of Interdisciplinary Management* 2019, 15.

Webster, M. Right to Be Forgotten [Podcast]. *RadioLab* 2019 August 23.

Vavra, A. N. The Right to Be Forgotten: An Archival Perspective. *The American Archivist* 2018, 81(1): 100–111.

Grandhi, S. A., Plotnick, L. and Hiltz, S. R. An Internet-Less World?: Expected Impacts of a Complete Internet Outage with Implications for Preparedness and Design. *Proceedings of the ACM on Human-Computer Interaction* 2020, 4(no. GROUP): 1–24.

Greenstein, S. The Basic Economics of Internet Infrastructure. *Journal of Economic Perspectives* 2020, 34(2): 192–214.

Vox. How Does the Internet Work? [YouTube Video]. *Glad You Asked S1 E2* 2020 January 8.

VIDEO TRANSCRIPT

Aeron MacHattie, Teaching and Research Librarian

Intro slide: Concordia Library logo

Interviewer: I've heard of the concepts of link rot and content drift. Can you please explain them? [question displayed on the screen]

Aeron MacHattie: Sure. So, the internet's made up of hyperlinks, so there are different pages that are linked together with hyperlinks. And those links are what allow you to navigate the Internet, find your way around bookmarked pages and so on. So, when those links break, the Internet breaks and that can happen for various reasons.

Visual: search on Concordia Library website and enter wrong address, the web shows file not found.

Aeron MacHattie: If the pages are deleted, if they're moved, if they're named even. And when those links break, that's called link rot. Content drift is a related issue, and it's what happens when the content of a web page changes over time. That sometimes happens if somebody new takes over a web page.

Visual slide text: the web is fragile. Link rot: hyperlinks can easily break if webpages are deleted or renamed. Content drift: webpages change over time

Aeron MacHattie: So, for example, if a new government comes to power, government ministries and agencies might have their web pages changed dramatically. But it can also just happen in the course of day-to-day updates that happen to websites and say your local public library could have information on it about a knitting program that you linked to. That then is later taken down because the program no longer exists, or it's been moved to a different part of the website. So, the link remains the same. The link is still functional, but the information that was linked to is no longer

there. So, while the link works, it’s kind of meaningless and you're losing the contextual information that makes that link makes sense.

Visual: video clip shows a web page is being scrolled down

Interviewer: But how are these related to me? [question displayed on the screen]

Aeron MacHattie: So, on an individual level, it can definitely be inconvenient or frustrating to not be able to either find the information you want or to follow the link that you're used to following.

Visual image description: people with confused expressions

Aeron MacHattie: But it's a much bigger systemic issue when it comes to fields that need big, stable bodies of knowledge and that rely on functioning links to continue to exist. So increasingly, links are used in the references of variety of different kinds of articles. So, it can be any kind of academic journal, could be medical studies, could be legal decisions, could be journal articles or newspaper articles. And when those links stop working, or when the content that's linked to goes away, it undermines the credibility of the article and takes away a huge chunk of information from the scholarly conversation happening in that field.

Visual: screenshot of SpokenWeb with an article named “The Bad Batch”: OpenRefine as a Batch Editing Method in SWALLOW

Visual: screen capture shows a mouse click on “cite this article”

Interviewer: Interesting. I never have thought about this. So, what should we do to save the content? [question displayed on the screen]

Aeron MacHattie: So, web archiving is one response to these issues. The Internet Archive was one of the earliest web archives and it started archiving web pages in about 2005. And usually, the way that web archiving works is that there's a computer program that either navigates the internet automatically taking snapshots of the pages it encounters or a similar program that will record the clicks of a user and then record snapshots of the pages that the user visits.

Visual slide text: web archiving software navigates the internet, takes snapshots, stores the snapshot

Aeron MacHattie: And that way, what happens is it builds up and snapshots of the pages and then also the links between those pages to preserve the relationships and allow you to kind of navigate the pages as if they were still live, even if the pages disappear.

Visual: screen capture of Wayback Machine, and search library.concordia.ca. Clicked 2010, result page was displayed.

Aeron MacHattie: There are some problems with that because some media like videos or content posted to social media sites are a little bit more difficult to capture. So, it's not perfect, but it definitely can record the majority of the information that's there. Archivists and historians argue that it's really important to do this work because it allows us to capture and record the art, culture, and knowledge of our time that is otherwise at risk of disappearing as links break and as content is changed.

Interviewer: Sounds like the right thing to do. But should we preserve everything on the internet regardless of the content? [question displayed on the screen]

Aeron MacHattie: So, while there's definitely a lot of really good reasons to do web archiving, there is some pushback. Some people are concerned that private information, especially information that's either embarrassing or may be incriminating, that information persisting on the Internet in Internet archives, either for formal Internet archives or in the archives of newspapers, that could have really serious repercussions on their lives and their relationships and their jobs, and that they should have a right to have that information removed.

Visual slide text: Two sides of a coin. Preserving the content of the internet is important because of its historical significance! What about content with personal information? Do we have the right to be forgotten?

Aeron MacHattie: So, in 2014, the European Supreme Court ruled that individuals do have the right to request that inaccurate or outdated information be removed from search engines like Google.

Visual: Screenshot of Google search shows some results may have been removed under data protection law in Europe.

Aeron MacHattie: But there's not a lot of legislation in other parts of the world. Another example is there's a blog post posted around the time of the 2016 Rio Olympics that outed a number of queer athletes. And there was a lot of concern then because some of those athletes lived in countries where homosexuality is really stigmatized or criminalized. And so, both the blog post and the snapshot in the Internet Archive were taken down out of concern for the safety of the athletes. And I believe that's the only instance in which the Internet Archive has actually removed content for right to be forgotten reasons. And there isn't any right to be forgotten legislation in Canada so far.

Visual: screen capture of Internet Archive shows page on how to request to remove something from the archive

Interviewer: And it seems contradictory to web archiving, doesn't it? Is there a middle ground? [question displayed on the screen]

Aeron MacHattie: Yeah, that's a really good question. And I think it's a question of balancing the rights of individuals who have privacy concerns with the need to have a public record. So, experts are concerned that if too much is done in the area of the right to be forgotten, that that could represent a kind of selective editing of history, and especially when it comes to governments, big institutions, public figures, elected officials. We want to be able to have a record of the things that people have said and what those institutions have done. Not having access to that information later could have really, really bad ramifications. And in some cases, web archives have actually been used in legal cases to make real legal decisions. So, as I said before, it is a balance. We want to make sure that people feel protected. But it is really, really important to have that public record and to have internet history being recorded for posterity, for legal decisions, for people's personal reference, and to make sure that everybody is held accountable for their actions.

Visual: thank you and credit. CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon and infographics & images by Freepik. Music: Lofi Summer Background by Vladislav Kurnikov on Pixabay.