January 26, 2015 Issue

By Jill Lepore

Can The Internet Be Archived?

Andrea Chang
Core 2 Interaction Studio

By Jill Lepore

The Beginning of the Internet

- In 1963, researcher Marvin Minsky at M.I.T points out how books are good tools for displaying inofrmation but bad at storing, organizing, & retrieving.
- He was one of the first to reject “printed page as a long-term storage device”

Common Misconceptions of the Internet

- "What 's on the web will stay on the web forever"
- Average life of a webpage is 100 days
- Sometimes webpages such as posts are deliberately deleted or sites hosted by corporations die with its hosts. (ex: Myspace)

...this has led to many problems

So many webpages disappear and/or are moved, many people have a hard time finding it or sourcing it. This is known as the **"content drift"** or **"reference rot"**.

Many legal scholars, lawyers, judges, etc rely on these webpages to cite in their footnotes as their proof of evidence.

A 2013 study found that by the end of six years, nearly 50% of URLS cited no longer work.

By Jill Lepore

Internet Archive and the Wayback Machine

- Later on, the Internet Archive (archive.org) & the Wayback Machine was founded & invented by Brewster Kahle.
- Motto: “Universal Access to All Knowledge”
- Inspired by the Library of Alexandria, except it’s open for everyone to see & use.

By Jill Lepore

What is the WayBack Machine and How does it work?

The WayBack Machine is a web archive & collection of old webpages. It is essentially a robot that scours through the internet & makes a copy of every webpage it can find.

People can also manually add onto the archive collection by copying an URL into a service called Archive It (archive-it.org).

The public can then access the information copied from the machine in the Internet Archive, which also has a physical library located in San Francisco, California. However, not everything the machine copies can be accessed due to copyright laws.

By Jill Lepore

How do archives deal with copyright?

Compared with the Internet Archive, the Library of Congress usually asks for permission before collecting the information. This is known as the **opt-in policy**.

On the other hand, the Wayback Machine will collect anything it finds, unless, that page is blocked with a text file, “robots.txt” at the root of the web site. The robot will honor that file & at the same time also remove all past versions of the site. This is known as the **opt-out policy**.

By Jill Lepore

2002

- Later in 2002, Kahle proposes an idea where the Internet Archive would collaborate with other national libraries & become the head of a worldwide organization of web archives.
- Goal: make information “freely available to the world in the face of increasinginly restrictive digital options”
- However, the plan was cancelled because of issues with national laws such as copyright and privacy.

By Jill Lepore

The Future

- In 2014, a tool called Perma.cc was launched by the Harvard Library Innovation Lab & allows users to create an archived version of the page link. By clicking on the link in the footnote, the user will be transported to the permantly archived version of the webpage.

- Memento is another tool that will be launched & it works as a web extention that allows users to navigate through every major public web archive worldwide & find the closest match to the webpage based on the time period you want.

- The same group as Memento will also be launching Time Travel, which is essentially a web portal.

Further Questions...

Are there any webpages the WayBack Machine cannot find & store? Such as the Dark Web?

Is information found in the Internet Archive safe & accurate?