What is Web Archiving?

Archival Processing
Distinctive Collections | Northwestern University Libraries

Web archiving is the process of copying portions of the World Wide Web periodically and saving them on a separate server. This is completed by robots, called web crawlers, which automatically assess and copy web pages when instructed to do so.

Websites evolve and are updated with different information over time. Because of this ephemeral nature, we are at risk of losing primary resources representing our cultural heritage. When websites are crawled at specific points in time, we’re able to preserve much more fixed information. Furthermore, if the same website is captured periodically throughout its life, a database of the website’s history is available for browsing.

Once a website has been captured by a web crawler, the copy of it is saved in a separate database. Instead of visiting the website’s original URL in a browser, researchers interact with the database to view the historic inventory of web captures for that site, perform keyword searches, and browse the crawls.

When web crawlers capture websites, the archived versions of these pages maintain the majority of the original website’s functionality, allowing researchers to use the website as it existed at the time of its creation. However, there are some known challenges which prevent perfection level capturing for all websites. For instance, the functionality of dynamic content, such as zooming on images or downloading files, may not work when captured by web crawlers. Similarly, any streaming media often cannot be rendered by the database. Any web content that exists behind a protected site, such as an intranet or other password-protected authentication, cannot be crawled.

While the robots conducting web crawls run automatically, archivists take an active role in ensuring that data is captured as accurately as possible to resemble the original website. Understanding it is impossible and unnecessary to save everything on the web, archivists also spend time curating the frequency and amount of web crawls.

Current examples of our web archives can be found online: https://archive-it.org/home/Northwestern