Preserving the Web

Digital Preservation Topical Note 10

DigitalPreservationCoalition

What is web archiving?

Preserving the web, or 'web archiving', refers to the practice of taking a copy of a website or of particular content published on the web to act as a record. A web record might consist of an entire website or only the text from a few pages. Web records require urgent attention because the web by nature is ephemeral. Statistics from the UK Web Archive – which takes a full copy of the entire UK web domain once a year – show that after one year, only 10% of the web remains live and unchanged [1]. There are a number of reasons web content is archived. It might be used as evidence to support compliance with legal or regulatory requirements. It might be archived to support on-going work or as corporate memory. Organisational policy may also require that all web records be preserved.

When should I archive web content?

It is especially important to capture web content when it is the only version of a record. Many records published on the web may be archived discretely. For instance, an organisation may archive reports or other documents through local records management. In other cases, an organisation may share content through their website that is not captured anywhere else. An archiving policy governing how different records are managed may be available to provide guidance on whether web content needs to be archived. It is also important to know the laws and regulations that apply to the web content you create. Many organisational websites constitute official records and are subject to Freedom of Information requests. Organisations may be held accountable for commitments they have made to the public via the web as an official channel of communication.

Key Term: **Web Crawl (or Harvest)**

Refers to the act of browsing the web automatically and methodically to index or download content and other data from the web. The software to do this is often called a web crawler or harvester. A web crawler allows users to take a copy of a website on a set frequency, such as once a day or once a year; a large domain-level web harvest can often take months. Heritrix, an open-source tool created by the Internet Archive, is a common type of web crawler.

Key Term: **WARC file**

WARC (Web ARChive format) is a file format (.warc) used to hold web harvests; a type of file 'container', or 'wrapper', that holds multiple types of digital information and metadata in one computer file

This Digital Preservation Topical Note was produced with the kind support of the National Archives of Ireland

[1] Webster, Peter (20 March 2015) 'How fast does the web change and decay? Some evidence', *Web Archives for Historians*, https://webarchivehistorians.org/2015/03/

How is the web archived?

There are several ways to archive the web. For very large web archives, such as the UK Web Archive and the Internet Archive, the most common method is to use a *web crawler*. A web crawler navigates through websites (like a spider) taking a copy as it goes. The crawler then converts this information into a WARC file that can be rendered (or played back) by a tool such as the Wayback Machine. For more targeted content, a smaller tool might be used. HTTrack, for instance, is a free, open source tool that can be downloaded onto a desktop computer. HTTrack will take a copy of a website and store it in structured folders. Printing to PDF might also work effectively for capturing a very narrow scope of web content, such as one or two pages of text. However, tools like HTTrack and print to PDF 'flatten' the web content. If there are any hyperlinks or other dynamic functionality necessary to understanding the meaning of content on the web, these tools may not adequately capture the record. In some cases, it may be more economical to have a specialist third party provide web archiving as a service.

How can I create web content to make archiving easier?

There are a few simple steps for creating web content that will simplify the process of web archiving. For example, dynamic content such as JavaScript, embedded calendars, or social media feeds pose difficulties for web archiving tools. Limiting the use of these types of content makes web archiving more straightforward. URLS and pathnames can change, leading to '404s' or 'Page Not Found' results. Keep this in mind when linking to external websites. In some cases, it may be useful to provide a textual description of the relevant content on the linked page in case that page disappears.

How does copyright affect archiving the web?

It is also important to keep copyright restrictions in mind when creating web content. Web pages are made of many different components. Features such as comments, social media feeds, or other functions that allow members of the public to share information could introduce risks of copyright infringement. Content posted by an external party, or shared on behalf of an external party, may also introduce risk of copyright infringement. Gaining permission both to publish and to copy for the purpose of preservation can reduce restrictions to archiving valuable material in the future.

For more information on Digital Preservation visit the DPC Website: https://www.dpconline.org