Web archiving

1. Introduction

Websites created by public offices and local authorities (public sector organisations), and the information they contain, are public or local authority records. As such, they must be managed in accordance with the Public Records Act 2005 (the Act) from creation to authorised disposal.

Managing active websites and archiving websites when they are no longer in current use are two different activities. For information on managing website content and website activity, see our guide on *Managing websites as records* (23/G23).

2. When to archive websites

Archiving websites should happen in times of change, for example, when a public sector organisation is intending to:

- develop a new website
- make major updates or redevelopments to an existing website, or
- decommission a whole website or part of a website.

Some archiving processes can capture web records that might not be captured by other systems or processes. For example, if web records are managed by a Content Management System (CMS) which does not provide rollback functionality, it may be necessary to take a snapshot of the website at risk determined intervals. Organisations need to be aware that it may be necessary to keep a copy of the entire website to ensure the records associated with it can be referred to for recordkeeping and business purposes.

Websites may also be archived to preserve them long term for cultural/historical purposes. Websites captured for this reason may not always be managed by the creating organisation.

3. Harvesting and snapshots

Harvesting is the process of capturing a whole website or specified parts of a website, usually with commercial tools such as external site crawlers. Crawlers are software packages or pieces of code that index or copy websites in a methodical manner, then save the selected elements as static pages to a disk. The resultant data is a snapshot of the site at a known point in time.

There are benefits and risks to using this approach such as:

- although the context of the web information is preserved, creation, rollback and metadata may not be available
- the 'look and feel' of the website may be preserved but it produces a static version of information that may have been originally presented in a dynamic or personalised manner.
- may only capture public facing pages but not intranet pages or secured content
- some content such as multimedia formats may not be captured if located on a different server to the HTML pages.

3.1. National Library web harvesting

The National Library primarily collects New Zealand websites and individual documents published on websites under legal deposit legislation (National Library of New Zealand Act 2003, Part 4). This includes regular annual harvests of public sector organisation websites.

Selective harvesting priorities are outlined in the New Zealand and Pacific Published Collections collecting plan. An annual harvest of the New Zealand internet is also undertaken. This includes websites that are not necessarily included in the selective web harvesting programme, such as schools and tertiary education institutions. For more information about the National Library's web harvesting programme, see their website.

Be aware however, that for public sector organisations, the National Library's web harvesting programme does not constitute authorised transfer or disposal under the Act. Also, as there are difficulties in harvesting large, complex websites, there is no guarantee that this will meet other legislative recordkeeping requirements, such as the Official Information Act 1982, the Local Government Official Information and Meetings Act 1987, and the Privacy Act 2020. Organisations should assess the risk if unable to access vital web records and consider the use of other systems such as a CMS for maintaining a usable copy of a discontinued website until it can legally be disposed of.

4. Transaction logs

Transactional logging is the recording of actions that occur to a web page, information or artefact. Almost all CMS products enable the recording of actions known as a transaction. Actions may be the creation of a new page, the publishing of a new content item, or the submission of a form. Collated lists of transactions are called transaction logs. They are often saved to a database table or text file within the application that generated the transaction.

There are benefits and risks to using this approach such as:

- can be easily set up with most database driven applications, but has limited accessibility
- captures raw information, but context is often lost.

5. Digital preservation approaches

Normalisation and format migration are two digital preservation strategies that can be used for web archiving. Normalisation usually involves converting or 'normalising' web records to a different standard file format (either a static document format such as PDF or as an HTML document) or using an emulation tool to access or read obsolete or uncommon formats. Format migration involves migrating web records in old file formats to newer versions when they are at risk of becoming obsolete.

Both strategies can be automated but quality control is incredibly important:

- Normalisation creates homogenous, easier to manage collections and means that users only need to know how to use a few file types.
- Format migration means that files can be accessed in current IT environments, but careful consideration must be given to migration pathways to avoid loss of data and functionality.

For further advice about digital preservation approaches to web archiving, contact us.