The National Archives

Web Archiving Guidance

Any enquiries regarding the content of this document should be sent to Archives Sector Development asd@nationalarchives.gsi.gov.uk

This document/publication is also available at nationalarchives.gov.uk/archives-sector

CONTENTS

1 Introduction

1.1 What is the purpose of this guidance?

1.1.1 This guidance explains what web archiving is and how it can be used to capture information which is published online. It is aimed at people who are new to the concept of web archiving, and those who may have heard about it, but are unsure of potential options and methods available.

1.1.2 After reading this guidance, users will know:

- What web archiving involves
- Why web archiving is important
- Potential approaches to archiving websites.
- Who to approach to discuss archiving their website externally.
- Hints and tips for good practice.

They will also know:

- The importance of websites as records of activities and functions of their organisation or group.
- Why web archiving is different to making a back-up copy of a website.
- Some risks of saving web content to removable media such as CD, which is vulnerable to loss or degradation.
- Some technical considerations for discussion with web managers.
- Sources of further information.

1.1.3 This document compliments the detailed guidance for web managers in central government produced by The National Archives and Central Office of Information: http://coi.gov.uk/guidance.php?page=239 (The detailed guidance is aimed at government web managers, though may be generally useful for web managers in developing websites which are easier to archive.)

1.2 Who is this guidance for?

1.2.1 This guidance is suitable for archivists, records managers and those with responsibility for archives, records and information in:

- central government
- local government and the wider public sector;
- religious and private institutions;
- businesses;
- charities and voluntary organisations;
- community groups;
- and for people with their own websites.

It can also be used to *influence* discussions with web managers and IT staff in those groups and includes a section with some basic technical detail to support this.
It is of potential use by any organisation, group or individual with an interest in preserving their website as a record of functions and activities and to support their ongoing activities or business processes.

2 Web Archiving

2.1 What is web archiving?

2.1.1 Web archiving is the process of collecting websites and the information that they contain from the World Wide Web, and preserving these in an archive. Web archiving is a similar process to traditional archiving of paper or parchment documents; the information is selected, stored, preserved and made available to people. Access is usually provided to the archived websites, for use by government, businesses, organisations, researchers, historians and the public.[1] As in traditional archives, web archives are collected and cared for by archivists, in this case 'web archivists'.

2.1.2 As the Web contains a massive amount of websites and information, web archivists typically use automated processes to collect websites. The process involves 'harvesting' websites from their locations on the live Web using specially designed software. This type of software is known as 'a crawler'. Crawlers travel across the Web and within websites, copying and saving the information as they go. The archived websites and the information they contain are made available online as part of web archive collections. These can be viewed, read and navigated as they were on the live web, but are preserved as 'snapshots' of the information at particular points in time.

2.1.3 Some organisations use simple tools and processes to archive their own web content. National libraries, national archives and various groups and organisations are also involved in archiving culturally important Web content in detail. Commercial web archiving software and services are also available to organisations that need to archive their own web content for their own business, heritage, regulatory, or legal purposes.[2] The largest web archiving organisation crawling the Web is the *Internet Archive* which aims to maintain an archive of the entire World Wide Web.

2.2 Types of web archiving

2.2.1 There are 3 main technical methods for archiving web content: client-side web archiving, transaction-based web archiving, and server-side web archiving.[3] Client-side archiving is the most popular method and can be carried out remotely and on a large scale. Transaction-based and server-side approaches require active collaboration with the server owners and need to be implemented on a case-by-case basis.

---

[1] Wikipedia, *Web archiving* http://en.wikipedia.org/wiki/Web_archiving

[2] Wikipedia *Web archiving* definition http://en.wikipedia.org/wiki/Web_archiving

[3] These definitions have been borrowed from *Web Archiving*, Julien Masanès (ed.), (Springer, 1998). See Chapter 1, in particular.

2.2.2 Note: All 3 approaches are different from a website 'back-up' which merely allows for a site to be put back together from saved files in the event of problem. The methods described above concern *archiving* of websites and this means that sites can be collected, preserved, accessed and navigated by users in ways similar to the original live site.

Further detail on these types of technical approaches to web archiving is available at Appendix A.

2.3 Why archive websites?

2.3.1 Many organisations create websites as part of their communication with the public and other organisations as they are powerful tools for sharing information. Websites document the public character of organisations and their interaction with their audiences and customers. In addition, information published on the web is increasingly becoming the only place where it is available. Because of this, the website is a crucial part of the records and identity of an organisation or individual.

2.3.2 Because the web provides access to up-to-date information, often websites are regularly updated and are constantly evolving. This is one of the web's great strengths, but also means that information supplied this way can sometimes be viewed as ephemeral in nature and as having little or no ongoing value. This means that it can be lost before being captured as evidence, for business or historical purposes.

2.3.3 Much of the early web and the information it once held has now disappeared forever; from early online content in the early 1990s to around 1997, very little web information survived. This was before the recognition of the ongoing value of legacy information published online, and before the first web archiving activities which began in 1996.[4]

Since the 1990s, as well as becoming culturally significant, the web has become even more significant as a hub of information. As a result, the web has become integrated into other activities, such as research, referencing and quotation. These activities that used to rely on physical records now increasingly use and link to pages and documents held on websites. Web archiving is a vital process to ensure that people and organisations can access and re-use knowledge in the long-term, and comply with the needs of retrieving their information.

2.3.4 Web archiving can be a relatively low-cost and efficient process, depending on the approaches used. Ideally, web archives should be harvested in their original form and be capable of being delivered as they were on the live web, providing a record of web content as it was available at a specific date and time. When a website is archived, the context of the information it provides is maintained, meaning that users can view the information in the context in which it was originally presented.

2.3.5 Back-up copies of websites do not always result in viable web archives, especially where websites use active scripts. Back-up copies where websites use active scripts would just contain the programming

[4] *Internet Archive, About* webpage www.archive.org/about/about.php

code and are not harvested from the web and time-stamped. (Time-stamping is a computer-readable date and time that the crawler applies to each file it harvests. This ensures that the archived website is a viable representation of the website at the time the website was archived).
For websites which use only flat HTML and for personal archiving of your own website, back-up copies are acceptable where they include dates of creation and changes within the back-up files.

2.3.6 Archiving websites gives organisations the chance to provide access to legacy information that they may not necessarily want to keep on their 'live' website. Evidence from the Web Continuity initiative at The National Archives shows a significant and ongoing user demand for access to older content that an organisation may consider out of date or unimportant. Archiving and providing access to this content becomes part of wider information and records management activities. As such, web archiving can contribute to a positive image of an organisation's ability to manage its information effectively.

3 Records and information management

3.1 Websites as records

3.1.1 The management of information on a website should be part of a wider approach to information and records management. It should be managed, reviewed and selected by following the same practices used for any other records created by your organisation.

3.1.2 Records and information need to be managed and retained: for ongoing business use; for legal purposes; as evidence; and for historical and cultural purposes. Just like paper and digital files, websites support the current and future activities of you and your organisation. If websites are valued as records and for the information they contain, capturing, managing and retrieving that information for as long as it is needed is a powerful and positive contribution to management of all of your essential records and information.

3.1.3 For people new to records management, basic guidance is available from The National Archives: www.nationalarchives.gov.uk/information-management/projects-and-work/records-management-code.htm . This is part of a wider selection of guidance for records managers. It is aimed at government and public sector organisations, though has general principles which are useful for records and information management in general.

3.1.4 Remember that back-up copies of your website are not intended to be used as web archives; they are useful for ongoing business purposes. Do bear in mind that it may not be possible to recreate your website from the saved files without detailed work by a web manager or designer.

3.1.5 Always check and double-check content before it is published online; even if it is only on the live web for a day, it may have been archived somewhere, or cached by search engines.

3.1.6 As mentioned in section 2, websites change and evolve over time as they are updated. Evidence of how your website appeared, and the information it contained at certain points in time is a valuable record and would be essential if needed for evidential or legal purposes.

3.2 Selecting and collecting

3.2.1 Websites are a record in themselves, of how an organisation wanted to present itself to the public and what information it communicated to them. Websites can also contain documents such as board minutes, reports, policies and plans. All of the information and documents on websites are records of the activities that created them. The have value as assets to the people and organisations that created them, as time and money has been invested in their creation and management.

3.2.2 Consider the value of the website and its content. Does it contain content that is of business or historical value? Is this content kept or preserved elsewhere, for example, in a shared network drive or an electronic records management system?
You can use the principles of business, evidence and historical value to evaluate the information and documents on your website and how these relate to other records that you need to keep. From this, you can decide which information to keep and how long you need to keep it for. For example, financial records are usually kept for at least 7 years.

3.2.3 Overall, approaches should be format-neutral. This means that records and information are managed according to why they were created and what they were used for. They are not managed differently because they are in a different format, such as spreadsheets, PDF documents, images, websites and so on.

3.2.4 Consider how often you need the website to be archived, for example, once a year, twice a year, every three months or even just once. This will depend on how frequently the website and its content changes, and the relative importance of the content. A website may need archiving more often at certain times, for example, if there is particularly important event that means the website is changing regularly then more frequent archiving might need to be arranged.

3.2.5 The other (non-technical) consideration for web archiving relates to the scope and scale of collecting. This can also help decide which technical approach to use, depending on what needs to be kept and why. Websites of central government are selected according to Operational Selection Policy 27, which is available online here: nationalarchives.gov.uk/documents/information-management/osp27.pdf

3.2.6 If the website and content are considered to be of value then arrange for it to be archived. **See section 5, Archiving your website: what you can do.** Bear in mind that not all content on a web page can be captured through web archiving.

3.3 Hints and tips on archiving and websites

3.3.1 Saving and archiving

- Don't save all your web content to removable media such as CD or DVD – these are vulnerable to loss and physical degradation over time, especially if you only have one copy.
- When you create back-up copies of your website for business purposes, it is better to save these to a network drive, which is in turn backed up to another data source.
- If your only option is to save all your web content to CD or DVD make multiple copies, store at least one copy at another location, check all copies to see that they work every 6 months, and copy content to new CDs or DVDs every 3 years. This helps to prevent loss if a disc fails, if the discs are damaged or stolen, or if the discs deteriorate over time.

3.3.2 Websites and archiving

There are some things that can be done to help make websites easier to archive. Details of these items have been included below. It will be useful to discuss the items with your website manager or designer.

- Where possible, keep all content under one root URL. URL stands for 'Uniform Resource Locator', the global address of resources and documents on the web. This means that everything in a website can be easily crawled and therefore archived. It makes it easy to identify that the entire site has been captured.
- Use 'friendly' URLs. Human-readable or friendly URLs are good practice for a number of reasons, including usability, security, and search engine optimisation.
- If using scripting (such as JavaScript) on your website, provide plain HTML alternatives – this supports accessibility for users and supports archiving. Provide static links or 'basic page anchors' where possible, rather than dynamically generated URLs. If your website includes rich media and streaming content, provide alternatives such as progressive downloads alongside streamed content
- Review your website for accessibility according to the standard produced by WC3. A website that has been designed to be W3C Web Accessible should also be easy to archive. [5]

[5] Information on accessibility is available at: www.w3.org/ ; for an overview and tutorials see www.w3schools.com/ .

- Provide XML site maps, which list and link to all of the content of your website. This is useful for users, makes your website more findable by search engines and supports archiving. [6]

- Ensure that content is not being added and removed between acquisition sessions as this content will not be captured and preserved.

- To be captured by web archiving, information needs to be 'machine reachable'; which means that it can be reached by a web crawler. Information that needs a log in, tick box, pick list or search box to access it is not machine reachable and so cannot be captured by a web crawler. Where possible try to provide alternatives that can be directly downloaded, such as an A-Z list or site map.

- Remember that content posted on externally owned websites such as Flickr or You Tube may also need to be managed and preserved either via your own website or other electronic systems. This is because these websites are outside the control of your own organisation and cannot always be harvested effectively and completely. See for example the archived versions of the Prime Minister's website, Number 10 which uses Flickr, YouTube and Twitter : http://webarchive.nationalarchives.gov.uk/20110131164346/http://www.number10.gov.uk/

- It is not always clear where the boundaries of any one site are, external links may captured for one 'hop' from the website, or the user of the archived website may link to the live external site.

3.3.3 More detail on how different aspects of websites and formats of online information may affect web archiving are available on the Central Office of Information website: http://coi.gov.uk/guidance.php?page=244 (These are aimed at central government, though can be useful when considering options for all websites.)

4 Access and Preservation

4.1 How long and who for

4.1.1 The reasons for archiving your website will inform how you manage and provide access to the information you collect. This will also determine why it is kept, how often it is collected, how long it needs to be kept for, and who can access it. If the material has been archived for legal or compliance reasons, it will be kept and used according for as long as required. When heritage organisations such as archives, libraries and museums collect web archives on a larger scale, they aim to preserve them permanently.

4.1.2 Archived websites are usually made available in some form to someone – either within the organisation only, or to the wider public for research purposes. Access depends on the reasons that the websites have been archived and whether this allows open, restricted or closed access. Websites collected

[6] XML stands for e**X**tensible **M**arkup **L**anguage. Further information is available on the WC3 schools website: www.w3schools.com/xml/default.asp Site maps allow web crawlers to navigate around all content on a website.

for legal or compliance purposes will usually be accessed and used only within the organisation and the organisations they are accountable to.

4.1.3 Heritage organisations will usually also provide access to researchers onsite and online, unless access to the content is restricted. Copyright or licensing issues may prevent wholesale access to this material on a public website. Archiving of third party copyright material may be permitted under legislation, though the archived material is made available only in certain locations (e.g. a library reading room). In many countries Legal Deposit legislation has been extended to permit extensive web archiving across the whole country domain. Usually access to such archived material is restricted to the onsite reading rooms of the national collecting institutions involved.

4.1.4 Wider online access is made available where permissions or collecting remits allow. The National Archives' *UK Government Web Archive* is one such collection which is freely available online: www.nationalarchives.gov.uk/webarchive/ There are advantages to this approach, as The National Archives encourages government website managers to redirect users to this web archive when pages are no longer available on live government websites: www.nationalarchives.gov.uk/information-management/policies/web-continuity.htm The approach supports good practice in government information management and allows anyone on the web to access the archived content.

4.1.5 The National Archives collects websites of UK government, documenting how our interactions with government are changing, and how government communicates with the public. There are a range of other organisations and groups of organisations archiving websites in the UK and across the world. Details of these collections are available online here: www.nationalarchives.gov.uk/webarchive/other-collections.htm

5 Archiving your website: What you can do

5.1 Heritage institutions

5.1.1 Many heritage institutions have adopted client-side web crawling as an approach because they are involved in large-scale web archiving programmes, and this is a very efficient way of capturing large amounts of material. Large scale web archiving usually involves one of two approaches (or a combination): full domain harvesting and selective harvesting. Full domain harvesting refers to attempts to collect a comprehensive snapshot of a national domain, for example all websites ending '.uk'). Selective harvesting collects websites under certain themes or types of sector.

5.1.2 Archivists in local record offices may consider curating a themed web archive collection. Please contact asd@nationalarchives.gsi.gov.uk for more information.

5.2 Local government

5 2.1 From Summer 2011, The National Archives is to pilot a web archiving model for 7 local authority archives to ensure important online information is preserved for future generations. The pilot will used as the basis for creating a template for procuring web archiving services and guidance on best practice to help archive services across the country develop their own web archives. The pilot will archive local authority websites as well as community or private websites which the archive services think may be of interest to future local historians.

5.2.2 In parallel with this work, The National Archives is conducting an automated web crawl of local authority and NHS sites in the next two years to capture a wide variety of locally-held information, including datasets which are not currently preserved by data.gov.uk . The resulting captured material will complement the in-depth work of the pilot.

5.2.3 Records managers in local government can consider several options including archiving their websites using web archiving software and services; archiving with your local record office; or a with a larger heritage institution. Contact asd@nationalarchives.gsi.gov.uk to discuss potential options.

5.3 Businesses and organisations

5.3.1 Some organisations may look at web archiving on a small scale, perhaps only to preserve their organisation's website or to gather a small collection of websites. This type of archiving can be achieved in a number of different ways, including cloud-based archiving, where an author stores a snapshot of a web resource using a third-party online service provider; creating a citation repository where materials cited in publications are collected and URLs for each resource are provided, and desk-top archiving, where individuals save local copies of web resources that are important to them.[7]

Please see 5.4 below for potential options.

5.4 Communities, projects and individuals

The options below are potentially useful for anyone beyond central government and NHS bodies.

5.4.1 It is possible to archive your own website using web archiving software and services. You can also consider archiving with your local record office or a larger heritage institution.

5.4.2 Some web archiving software is freely available online through open source licences. These include Heritrix, Web Curator Tool and Netarchivesuite. These are available here along with related software tools: http://netpreserve.org/software/downloads.php

Some technical expertise is necessary to set up and run these tools.

[7] For a full discussion of these approaches and the tools available see Digital Curation Centre and UKOLN *Web archiving summary and briefing paper*, p13-23 www.dcc.ac.uk/resources/briefing-papers/technology-watch-papers/web-archiving

5.4.3 Website copier tools, such as HTTrack may be useful in creating offline copies of websites which are suitable for archive purposes. It is available here: www.httrack.com/
Basic technical skills are necessary to set up and run this tool.

5.4.4 Some organisations provide on demand web archiving services. These include the Internet Archive providing Archive-IT: www.archive-it.org/ ; Internet Memory Foundation, providing ArchivetheNet: http://internetmemory.org/en/. Further examples are available by searching online.

5.4.5 Other options are available. You can approach the British Library web archiving team, who may be able to select your website to become part of their growing collection. Selections are made based on nominations, though not all will become part of the collection.
You can nominate your website for archiving at: www.webarchive.org.uk/ukwa/info/nominate
For further information, please contact the British Library's web archiving team:
www.webarchive.org.uk/ukwa/info/contact
The British Library's web archive collection is available at: www.webarchive.org.uk/ukwa/

5.5 Central government

5.5.1 The National Archives comprehensively archives central government websites, to create a historical archive of central government's presence on the web and to prevent 'broken links'. It regularly archives central government websites, including all departments, agencies and non-departmental public bodies (NDPBs).The UK Government Web Archive collection is available online here:
nationalarchives.gov.uk/webarchive/

5.5.2 Government departments and agencies should manage their online contents according to the guidelines set out in 'What to keep'. This initiative helps government ensure that it knows what information to keep and for how long, and to put this into practice. Information is available online at:
www.nationalarchives.gov.uk/information-management/projects-and-work/what-to-keep.htm

5.5.3 To check if your government website is being archived, please check the UK Government Web Archive list at: www.nationalarchives.gov.uk/webarchive/atoz.htm . You should also consider using the redirection tool, which redirects users to archived content when a broken link is found. Bear in mind that if using the redirection tool, users will be redirected to the latest version of the page in the UK Government Web Archive, so if information is continually being added to and removed from the same web page, the redirection will only take you to the latest version. To archive central government websites or discuss the redirection tool, please contact webarchive@nationalarchives.gsi.gov.uk
Further information is available online at: www.nationalarchives.gov.uk/webarchive/webmasters.htm

5.5.4 Information on archiving datasets hosted on government websites is available online here:
www.nationalarchives.gov.uk/webarchive/archiving-datasets.htm

5.6 National Health Service (NHS) bodies

5.6.1 The two-part Records management: NHS code of practice is a guide to the required standards of practice in the management of records for those who work within or under contract to NHS organisations in England. It is based on current legal requirements and professional best practice. It is available online here: www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4131747

5.6.2 To discuss options for archiving central NHS and Department of Health bodies, please contact: webarchive@nationalarchives.gsi.gov.uk

6 Books and online resources

6.1 Books

Brown, Adrian *Archiving websites,* Facet Publishing, UK, 2006

Masanes, Julien (ed), *Web Archiving,* Springer, 2006

6.2 Online resources

Central Office of Information, *Archiving websites* http://coi.gov.uk/guidance.php?page=239

Digital Curation Centre and UKOLN *Web archiving summary and briefing paper* www.dcc.ac.uk/resources/briefing-papers/technology-watch-papers/web-archiving

Digital Preservation Coalition, *Web Archiving and Preservation Task Force* www.dpconline.org/about/working-groups-and-task-forces/524-web-archiving-and-preservation-task-force

Joint Information Systems Committee, *Beginners Guide to Digital Preservation blog* http://blogs.ukoln.ac.uk/jisc-bgdp/

Joint Information Systems Committee, *Preservation of Web Resources (PoWR) Handbook* http://jiscpowr.jiscinvolve.org/wp/handbook/

Museums, Libraries and Archives Council and The National Archives, *Supporting long term access to digital material* www.mla.gov.uk/what/programmes/digital/Supporting_Long_Term_Access_to_Digital_Material

The National Archives, *Digital Preservation FAQs* www.nationalarchives.gov.uk/information-management/projects-and-work/digital-preservation-faqs.htm

Wikipedia, *Web archiving* http://en.wikipedia.org/wiki/Web_archiving

---

Appendices

Appendix A

A.1 Client-side web archiving

This is the most popular method employed, because of its relative simplicity and its scalability. This method allows the archiving of any site that is freely available on the open web, making it attractive to heritage institutions with an interest in preserving websites owned and managed by other organisations or individuals. Crawlers imitate the form of interaction that users have with websites, usually starting from a seed page (often the URL of the top level domain for example: *http://www.department.gov.uk/default.htm*), following and extracting links from pages, and fetching documents and information until they reach the boundary of the domain they are operating in. Typically such crawlers can capture a wide variety of web material – not only documents or text pages, but audio files, images and video, and data files. The success of capture very much depends on how accessible the material is to the crawler – streamed media files and content hidden behind forms or pick-lists are usually not easily gathered. The National Archives takes this approach to web archiving with its UK Government Web Archive: www.nationalarchives.gov.uk/webarchive/

A.2 Transaction-based web archiving

This type of approach is operated on the server-side and so requires access to the web server hosting the web content. It is much less frequently employed as a methodology and records the transactions between the users of a site and the server. Interestingly in this approach, content that is never viewed will never be archived. The main constraint is the need for access to the server, which will require agreement and collaboration with the server's owner. The primary advantage here is that it is possible to record exactly what was seen and when, so this approach is particularly attractive as a method for internal corporate and institutional archiving, where legal accountability or compliance is important.

A.3 Server-side web archiving

This method involves directly copying files from the server. As with the previous method, this can only be employed with the consent and collaboration of the server owner. The main challenge with this approach is to make the copied content usable as a navigable archived website. Many of the problems arise where dynamically-generated content is used, where content is aggregated on-the-fly from various sources in response to a user request. Copying database files, templates and scripts does not mean that it will be easy to regenerate content from the archive. What is required is to run the same environment with the same parameters in the archive, which can be challenging. The main benefits of this approach rest with the possibility of being able to archive parts of the site which are inaccessible to client-side crawlers.