The National Archives

The UK Government Web Archive

Guidance for digital and records management teams

Contents

1. Introduction

1.1 This guidance is provided to enable government digital teams and records managers to understand how to manage and maintain websites to ensure that the presence of government on the web can be archived successfully and made permanently accessible in the UK Government Web Archive (UKGWA). Information is provided here on how the web archiving process works, how the technical processes inform the web archiving service and its timetable, the limitations of what can be captured and made accessible through this preservation medium and the circumstances under which content can be removed.

1.2 Information on using the web archive (including the search function and redirection component) can be viewed on this page.

1.3 This guidance replaces TG105 Archiving websites.

1.4 The National Archives capture the public websites of UK Central Government, as part of their statutory duty to preserve the public record. All central government websites must be capable of being archived and so their suitability for this should be considered when designing, managing and decommissioning a website. The scope of the web archiving programme is defined in The National Archives' Operational Selection Policy OSP27: UK Central Government Web Estate.

1.5 The National Archives needs to be informed of new websites and any closures or substantial changes to an existing website so that their capture can be incorporated into the web archiving programme. Archiving may also occasionally take place on request, for example, in the event of a machinery of government change, at the discretion of The National Archives.

1.6 Web archiving is a complex process, involving many potential challenges. If the principles outlined in this document are not followed, it is likely that problems will occur. It may not always be appropriate or technically possible to archive some content through web archiving. In such cases, the responsible department should make alternative digital transfer arrangements with The National Archives.

The National Archives' web archive team can be contacted at webarchive@nationalarchives.gov.uk

1. Overview of the web archive process

2.1 Step One: Identification and selection of target website

There are currently over 4,000 websites in the UKGWA. We aim to capture all known in scope websites at least once and usually according to a schedule. In most cases each website is captured twice per year. Where possible, a crawl will also be made within the 6 month period prior to its closure.

Website owners should check the A-Z list and get in touch if a website isn't listed. Contact should also be made if a website that is listed is going to close or is about to be overhauled. The National Archives will confirm whether the website is in scope for archiving, determine whether any additional crawls will take place and, if so, define a suitable ongoing crawl schedule.

2.2 Step Two: Crawl

The National Archives uses the remote harvesting method for web archiving as it is the most scalable and efficient way to capture public web content. On average, 100 websites are archived each month.

The web archiving process may take several weeks to complete depending on the size and complexity of the website. This is from the beginning of the crawl to when the archived website becomes publicly-available. The Web Archive team can provide advice about how long it is likely to take to capture a particular site. No content should be removed from the website after the web archiving process has started as this will compromise the quality assurance and patching processes. Any content added to the website during this time will probably not be archived until the next scheduled crawl of the website. It is essential that the website remains accessible at all times during this period.

2.3 Step Three: Quality assurance (QA)

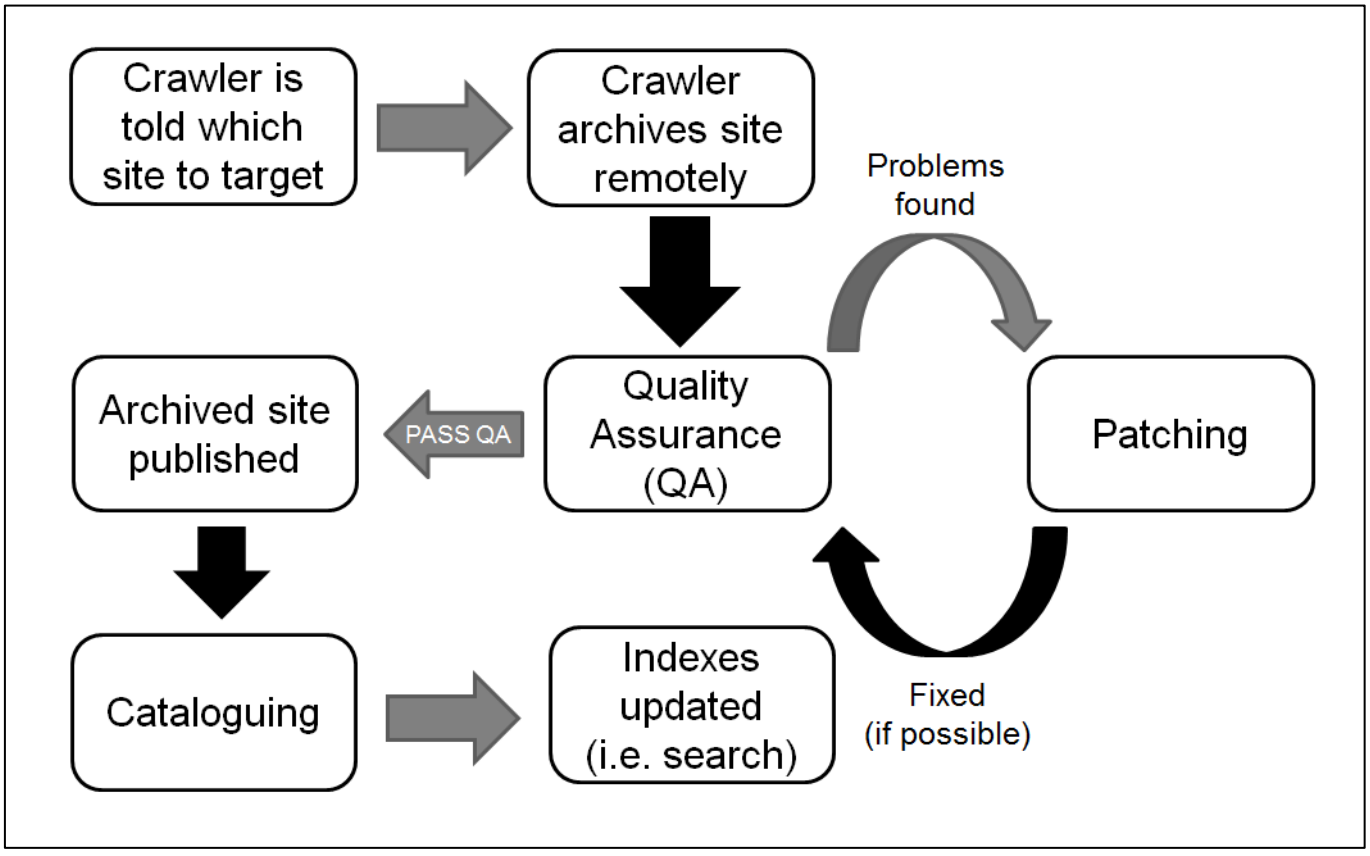
The QA process commences at a temporary URL, involving The National Archives' Web Archiving Team, our service providers and, ideally, the website owners. The aim is that the archived website should resemble the live website as closely as possible. Any problems found are fixed, if possible, through a process known as patching.

2.4 Step Four: Publication

The archived website moves to a permanent URL, where it will remain accessible online. If a resource has been archived, you can add the *http://webarchive.nationalarchives.gov.uk/*/* prefix to the original, or "live", URL to see all the captures it. For example: http://webarchive.nationalarchives.gov.uk/*/http://data.gov.uk/dataset/gb-road-traffic-counts

2.5 Step Five: Access

A description of the website is added to Discovery, The National Archives catalogue, to enable it to be viewed alongside the wider record of government. The index that underpins the UKGWA full text search function provided is updated monthly. The web archive is also open to all web search engines and is Memento compliant.



Please note: steps two to four may take **several weeks** to complete.

Appendix B of this guidance contains a checklist for website owners that are managing the closure of a website.

3. Archiving Websites: Technical Guidance

3.1 Overview

3.1.1 Website managers should be aware of the limitations of web archiving technology. Designing a web archive friendly website from the outset can save on a lot of work later and making content “machine-reachable” will go a long way to making it suitable for web archiving. The following guidance should be followed when setting up a new website, maintaining it and finalising a website content prior to its closure.

3.1.2 Websites must allow access to MirrorWeb’s crawlers. MirrorWeb provide web archiving services to The National Archives. We aim to crawl in the most polite manner possible but crawl rates can be made slower if necessary. Blocking or slowing down our crawlers with anti-robot technology will mean that the website cannot be successfully archived. Our crawler can be identified through a specific user-agent string. The Web Archiving team can be contacted for these details.

3.1.3 The copyright status of any government website and its content should always be clearly discernible to the user of the live and archived version of the website. The Government Digital Service has produced this guidance and The National Archives directs users to the statement regarding re-use of the information contained within the web archive (please see the section under re-use of content). Website owners must inform us of any content on their website does not comply with this guidance.

3.1.4 Any content that cannot be captured through web crawling should be raised by the department with their Departmental Records Officer (DRO) and The National Archives in order to assess whether it should be preserved through other methods as part of the public record.

3.1.5 When a website is closing, website managers should consult the below Checklist for website managers and contact the Web Archiving Team at **least two weeks before** any crawl would need to begin (in addition to the amount of time it takes to archive the website) so that the necessary preparations can be made.

3.1.6 When a website closes, ownership of the domain should be retained by the publishing government organisation even after the final crawl of the website has been made. This is usually handled by a DNS team. It is essential as it prevents 'cybersquatting'. It is also an important part of web continuity, as it gives the option to set up redirects to the UKGWA.

3.2 Limitations of web archiving technology

3.2.1 From the outset, it is essential to understand that all web archives are a snapshot, or representation, of what was online and accessible to the crawler at the time of the crawl; an archived website is not a copy of a website. This is because the underlying systems (or “back-end”) of the website cannot be archived using the remote harvesting method. Because of this and other technical limitations, the web archive should not be considered a “backup” of a website

from which the original website can be restored at a later date. In addition, please note that The National Archives cannot accept “dumps” of websites from content management systems, databases, on hard drives, CDs or DVDs or any other external media.

3.2.2 It is not possible to capture content behind logins, or on intranets, even if usernames and passwords are provided. The UKGWA is publicly-accessible and all the material captured in it must be in the public domain. If content is hosted behind login because it is not appropriate for it to be publicly-accessible, it should be managed there until its sensitivity falls away and then published to the open website.

3.2.3 Other limitations of web archiving technology are outlined on our guidance page.

3.3 How to make your website structure and content archive compliant

3.3.1 Present everything on your website using either the HTTP or HTTPS protocol.

3.3.2 It is generally not possible to capture streaming video or audio. Any content of this sort should also be accessible via progressive download, over HTTP or HTTPS, using absolute URLs, where the source URL is not obfuscated.

3.3.3 Where possible, keep all content under one root URL. The web crawler operates on a URL or domain scope basis, meaning that any content hosted under root URLs other than the target domain, sub-domain or microsite (see Glossary) is unlikely to be captured. If this is not possible, tell the Web Archiving Team about all external URLs you use and these URLs can be included on an XML sitemap or supplementary URL list to give to the crawler as additional URLs, or seeds, to capture. Typical examples include documents hosted in the cloud, and when any services that link through external domains are used.

3.3.4 Audio-visual (AV) material should be linked to using absolute URLs instead of relative URLs. For example:

> http://www.mydomain.gov.uk/video/video1.mp4
> rather than
> …video/video1.mp4 or video/video1.mp4

3.3.5 The web crawler relies on being able to discover links in the source code of web pages, therefore “orphaned” content (i.e. content that is not linked to from within your website, or provided to the Web Archiving Team in an XML sitemap or supplementary URL list before a crawl) will not be captured.

3.3.6 The web crawler is unable to discover links in binary files attached to websites (i.e. links included in .pdf, .doc, .docx, .xls, .xlsx, .csv documents). You should ensure that all resources linked to in these files are also linked to on simple web pages or that the links are provided to the Web Archiving Team in an XML sitemap or supplementary URL list before the crawl is launched. If an alternative link is not provided, the content will not be captured.

3.3.7 Links in binary files (i.e. links included in .pdf, .doc, .docx, .xls, .xlsx, .csv documents) do not currently work in the web archive. If a user clicks on a link in a document they will be taken to that location on the live web. Additionally, if I resource is only linked to in a binary file it will not be captured by our crawler. Web teams should ensure all resources are linked to on a standard web page or in an XML sitemap.

3.3.8 Internal links, other than those for AV content, should be relative and should not use absolute paths. This means that the links are based on the structure of the individual website, and do not depend on accessing an external location where files are held. For example:

> *or*
>
> *rather than*
>
>

3.3.9 Web archiving technology cannot always access URLs that are dynamically generated using scripting languages like JavaScript. The crawler may also have problems reading page content generated dynamically using client-side scripting. This may affect the archiving of websites constructed in this way. If the code is maintained in readily-accessible script files, as recommended, this will facilitate the diagnosis and fixing of problems. Moreover, if possible, the page design should make use of the <noscript> element to ensure that content is still readable, and links can still be followed.

3.3.10 Website managers should ensure that any client-side scripting is publicly viewable over the internet. This is normally the case unless specific measures such as encryption are taken to hide the script and, as far as practicable, is maintained in separate script files (e.g. with .js extension) rather than coded directly into content pages.

3.3.11 Avoid using dynamically-generated date functions on a website. This provides a poor user experience of the website in the web archive as any date shown will always display today's date. Instead, a website should use the server-generated date, rather than the client-side date.

3.3.12 The use of Flash technology (and similar) should be avoided where possible as we often have difficulty archiving these resources and retaining their functionality. If it cannot be avoided, alternative methods for accessing and displaying content should be provided. Bear in mind that where visualisations are used, the underlying data itself should always be accessible in as simple a way as possible.

3.3.13 Where possible, use meaningful URLs such as *http://mywebsite.com/news/new-report-launch* rather *than http://mywebsite.com/5lt35hwl*. As well as being good practice, this can help when you need to redirect users to the web archive.

3.4 How to make the navigation of your website archive compliant

3.4.1 Keep navigation as basic as possible, by providing static links, link lists and basic page anchors, rather than using JavaScript and dynamically generated URLs. If using scripting (such

as JavaScript) on your website, provide plain HTML alternatives. This supports accessibility for users, search engines and for web archiving.

3.4.2 Interactive functionality is unlikely to archive well. This includes search functionality, and anything else that requires a "submit" operation such as drop-down menus, forms, radio buttons, checkboxes, and so on. Alternative routes to accessing this content should always be provided as this will assist with crawling and will provide a better user experience to those accessing your website in the web archive. Make sure that any URLs that are problematic to reach or relate to the most significant content are provided in an XML sitemap or supplementary URL list, and tell the Web Archiving Team about them.

3.4.3 Web archiving functionality is limited when databases are used to support web pages. The web archive can only capture snapshots of database-driven pages if these can be retrieved via a query string, but cannot capture the database used to power the pages. For example, we should be able to capture the content generated at *http://www.mydepartment.gov.uk/mypage.aspx?id=12345&d=true* since the page will be dynamically generated when the web crawler requests it, just as it would be for a standard user request. This works where the data is retrieved using a HTTP GET request as in the above example.

3.4.4 It is not possible to archive content that relies on HTTP POST requests, since no query string is generated. Using POST parameters is fine for certain situations such as search queries, but website managers should ensure that the content is also accessible via a query string URL that is visible to the crawler, otherwise it will not be captured.

The following is an example of a complex combination of JavaScript, which will cause problems for archive crawlers, search engines, and some users:

```
<a
href="javascript:__doPostBack('ctl00$ContentPlaceHolder1$gvSectionItems','Page$1')">1
</a>
```

A well-designed URL scheme with simple links is a preferred alternative, for example:

```
<a href="content/page1.htm"
onclick="javascript:__doPostBack('ctl00$ContentPlaceHolder1$gvSectionItems','Page$1')
">1</a>
```

Website Managers should provide XML sitemaps, supplementary URL lists or, ideally, html pages with lists of hyperlinks to such content. This will help with the capture and QA process and will provide a route for users to access the information in the web archive.

3.4.4 XML sitemaps or supplementary URL lists should also contain the URLs on websites where pagination (../page1 , ../page2 and so on) is used, as the crawler can sometimes

misinterpret recurrences of a similar pattern as a crawler trap and therefore may only crawl to a limited depth.

3.4.5 Website managers should use simple, standard web techniques. There are few limits to a website builder's creativity when using the standard World Wide Web Consortium (WC3) recommendations. Using overly complex and non-standard website design increases the likelihood that there will be problems for users, for web archiving, and for search engine indexing.

3.4.6 In most cases, a website that has been designed to be W3C Web Accessible should also be easy to archive. Client-side scripting, along with other formats not covered by Web Content Accessibility Guidelines (WCAG), should only be used if it is determined that they are most appropriate for their intended purpose. It is good practice for website managers and developers to make any scripting transparent and to provide alternative methods for accessing information.

3.4.7 The UKGWA can archive and replay all versions of HTML to date.

3.4.8 Websites should support browsers that don't support JavaScript, or have it disabled, and provide alternative methods of access to content. It is also good practice to provide alternatives to animation, as well as transcripts of audio and video content.

3.4.9 The best way to design websites that use JavaScript is to follow a "progressive enhancement" approach. This works by building your website in layers:

i. Code semantic, standards-compliant (X)HTML or HTML5

ii. Add a presentation layer using CSS

iii. Add rich user interactions with JavaScript

3.4.10 Our technology is usually unable to capture content which is protected by a cross domain file. This usually affects content which is embedded in web pages but is hosted on another domain, such as multimedia content. If this is the case for any of your content please ensure that it is made available to the crawler in an alternative way.

3.4.11 Including a human-readable HTML sitemap in your website is good practice. It makes content more accessible, especially when users are accessing the archived version, as it provides an alternative to interactive functionality.

3.5 How to ensure that the datasets on your website are archive compliant

3.5.1 Since June 2010, the UKGWA has been comprehensively archiving datasets published by departments under the Government's Transparency initiative. Datasets, including those published and indexed on data.gov.uk, are captured approximately every 6 months, at the same time the data.gov.uk website itself is captured. This is designed to give the best user experience.

3.5.2 Underlying Data Publication Guidance (TG135) sets out the specific requirements for publication of datasets. In general, if the guidance set out in TG135 and in this document is followed, the web archiving of datasets should present no special problem.

3.5.3 The web archive can capture all file types but consideration should be given to future access and use of the data, especially where complex visualisation layers and browse search functionality is used, which are likely to cause problems in the web archive. Alternative, crawler-accessible navigation should always be provided to allow capture and access.

3.6 How content published via social media services can be added to the archive

3.6.1 Content posted on social media websites such as Twitter, Facebook, Flickr and YouTube represent a challenge to web archiving as the technology developed to deliver instantaneous and visually-appealing content to users is often incompatible with web archive technology for a variety of the reasons covered above. These constraints still apply to most embedded social media content on website.

3.6.2 However, The National Archives has been archiving blog content for a number of years and, from 2013, this was extended to include Twitter and YouTube accounts. In both cases, a specific account is targeted, so only the content hosted on it is archived. Additionally, customised access interfaces have been developed to suit the requirements of the projects.

3.6.3 If you need your social media archived, please contact the Web Archiving Team. Currently, this is limited according to our Operational Selection Policy. Your organisation's social media may also need to be managed and preserved either via your own website, or by other electronic systems, and in line with your information management policies. Your Departmental Records Officer (DRO) can provide further guidance.

3.7 What to do if you need to request content to be removed from the web archive

3.7.1 Content can only be removed from archived websites in exceptional circumstances, and only when it adheres to one or more of the criteria set out in The National Archives' Takedown Policy. Website owners should familiarise themselves with this policy and make sure that all content put on the web is fit for being in the public domain and to be made perpetually available in the web archive.

3.7.2 The underlying code of an archived website cannot be altered in the web archive. Captured website content is stored in an ARC or WARC file to preserve its integrity.

3.7.3 If website owners wish to remove content from their website that should be archived they should first check that it has been captured. The simplest way to do this on a URL by URL basis is by prepending an original URL with the *http://webarchive.nationalarchives.gov.uk/*/* prefix. If the content has been captured, it will display a list of capture dates.

For more information, see our web pages or contact the web archiving team at webarchive@nationalarchives.gov.uk

Appendix A: Glossary of Web Archiving Terms

Access	In an archived website, being able to navigate to, or otherwise access, a captured resource.
ARC / WARC file	The storage file formats used in the web archive.
Capture (or crawl, harvest, snapshot)	The process of copying digital information from the web to the web archive using a crawler.
Crawler (or spider, robot, scraper)	Software that explores the web by following links and extracting data. In web archiving, it is configured to follow seeds on target domains and capture web-based resources.
Crawler trap	A feature of a website that causes a crawler to make an infinite number of requests and in so doing can disrupt the crawling process.
Domain (or target domain, target website)	The website that the crawler is instructed to capture. This is an important factor in what is captured during a crawl.
Exceptional crawl	A crawl in addition to the regular crawl schedule, only used in exceptional circumstances at the discretion of The National Archives.
Index page	The access page with the *http://webarchive.nationalarchives.gov.uk/*/* prefix. By adding the original resource URL to the end, it shows whether a resource has been captured and, if so, lists all of the dates of capture.
Internet Memory Foundation	A not for profit organisation that carries out the technical web archiving and hosting process on behalf of The National Archives.
Microsite	A directory named after the root web address e.g. *http://domain.gov.uk/mysite/*.
Partial crawl	When, in the process of archiving another in scope website, the crawler leaves the target domain to a limited depth, resulting in a very shallow

	crawl of another website. Care should be taken not to confuse this with a full crawl of a targeted website as it is likely to be less complete.
Quality Assurance (QA)	The process of checking the completeness of capture by using a combination of automatic and manual tools and techniques.
Remote Harvesting	The act of harvesting content across the web. The National Archives' chosen method for scalability reasons.
Root URL	Typically the shortest form of URL on any website, off which everything else is contained (e.g. http://www.gov.uk/)
Scheduled Crawl	A crawl of a website performed at regular intervals. This can be any interval between 1 and 12 months.
Seed	A URL the crawler is instructed to capture.
Sub-domain	A directory named before the root web address, which typically looks like http://mysite.domain.gov.uk/ (where mysite is the sub-domain).
Supplementary URL list	A text file listing URLs on a website, especially those that are difficult for crawler technology to reach or those that are hosted externally. An alternative to an XML sitemap.
Web Archive	A collection of web resources stored in an archive. The UKGWA is among the few that are publicly-accessible.
Web Continuity	An initiative led by The National Archives in collaboration with other central government departments to eradicate broken links on government websites through redirecting users to the UKGWA.
XML sitemap	A list of URLs encoded in XML that assists crawlers, especially search engines, in indexing content.

Adapted and expanded from the Archive-It glossary (accessed March 2014)

Appendix B: Website closure checklist for website managers

We need at least eight weeks' notice to complete the web archiving process. This is the time from the launch of a crawl to when it is publicly-accessible. Here is a countdown checklist that may help you plan the archiving of your website.

At least 10 weeks before closure (please contact us as early as possible, particularly if your site is large and/or complex)	✓ Get in touch with us (the Web Archiving Team) ✓ Have a look at your website and its content to see if it poses any potential problems in terms of its structure and technical design. ✓ Use our A-Z or just put *http://webarchive.nationalarchives.gov.uk/*/* in front of any URL on your website to find recent crawls. Look out for recurring problems and raise them with us. There may be a lot of dates there, but most are probably not full archives of your website. Contact us for confirmation of these dates. ✓ Work with us to decide on when to start the crawl. Keep in mind that the crawl, quality assurance and publication process may take several weeks to complete.
1 week before the crawl starts	✓ Make sure we know about all sub-domains and orphaned content that needs to be archived.
From agreed crawl start date	✓ Avoid adding or deleting content from your website. Any content added will probably have to be put elsewhere after the crawl (gov.uk, for example) and deleting content may make the archive less complete
Post-crawl	✓ You know your website best, so please have someone on hand to help with the quality assurance process. ✓ Keep your domain registered – that prevents “cybersquatting” gives you options to redirect to the web archive or another appropriate location and prevents “Page not found” errors.