Tips on Finding Content in the UK Government Web Archive

Introduction

This document aims to provide our users with advice on how to access the UK Government Web Archive. The limitations discussed below are due to the fact that web archiving is a relatively new technology and it is being constantly developed. We welcome suggestions on how we could improve in this, and you can contact us using the email address at the end of this document.

Further information can be found here:
http://www.nationalarchives.gov.uk/webarchive/
and
http://www.nationalarchives.gov.uk/information-management/policies/web-continuity.htm

You can find some useful introductory information about web archiving at
http://en.wikipedia.org/wiki/Web_archiving

Please note that a ‘resource’ is a page or file (e.g. .html, .pdf, .xls, .jpg, .doc etc). A ‘crawler’ is software which collects the resources to form an archived website. Only publically-available and machine-accessible content can be captured by our web archiving technology. Therefore content hosted behind log-in screens cannot be captured and is not available in the web archive. Search in archived websites will not function and other navigational functions, such as ‘tick boxes’ and certain drop-down menus cannot be archived.

We hope that the following suggestions will assist you in finding what you need.

Full-text search

The UK Government Web Archive has a full-text function, which is hosted by the Internet Memory Foundation. It can be accessed here:
http://webarchive.nationalarchives.gov.uk/quick_search/

An Advanced Search function is also available at
http://webarchive.nationalarchives.gov.uk/adv_search/

This allows you to be more specific about what you are searching for. For example you can exclude certain terms from search results. It is also possible to restrict the search so only results from archived versions of a particular website or from archived websites in a particular category are displayed.

Both full text search tools are, however, known to be limited in certain respects. The web archive contains much near-duplicate and duplicate material, over multiple dates, which produces a lot of ‘noise’, or undesired results.

Enhancements to Search features such as a date range restriction are not available in the current version.
The National Archives is currently working on projects to provide users with easier ways to find content.

An additional, in depth, help guide for search can be found at http://webarchive.nationalarchives.gov.uk/search_help/

**Check with the “*”**

Our browse index (known as the * index) provides access to all the versions of a particular page contained in the Web Archive. The easiest way to find a page in the archive is to add the URL of that page to this prefix:

http://webarchive.nationalarchives.gov.uk/*/

For example if you wanted to see all instances of the Number 10 homepage you would add the URL: http://www.number10.gov.uk/ to the prefix to make

http://webarchive.nationalarchives.gov.uk/*/http://www.number10.gov.uk/

You will see an index page containing links to all the versions of the page contained in the Web Archive. The links are listed as dates on which the sites were crawled.

Due to partial crawls or occasional incomplete crawls of websites (because, for example, a website being restructured or closing before it can be fully crawled), sometimes checking the index page for a resource means that it can be accessed from a different date.

Simply insert the URL of the resource after the /*/ and, if it is present in the web archive, it will list all the dates at which the resource was captured. Please note that this method works for all resources contained in the Web Archive.

Please note that it is often possible to copy and paste the URL of a resource from the Web Archive even if the link does not work. If you hover your cursor over the link and right click and then select ‘Copy shortcut’ (in Internet Explorer) or ‘Copy link…’ (in Firefox or Chrome) the URL will be stored on your clipboard. You can then paste the URL into your browser and replace the date with the * sign. If any versions of the resource are available in the Web Archive the index page will be shown.

For example, if you copy and paste this URL from a page:

you would replace the date portion with a * sign to create this URL: http://webarchive.nationalarchives.gov.uk/*/http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_066309 to view the index page for a particular resource.

**A note on ‘partial crawls’**

Partial crawls occur when the crawler ventures outside its defined ‘scope’ and captures content on another website, but only to a very limited ‘depth’. This means that a page can appear on the * index page, but many of the links appear broken. This can be confusing to our users as partial crawls and ‘complete’ crawls are not differentiated. The
result is that it may appear that a document has never been archived, when in fact it probably was, albeit at a different date.

We are planning to address this issue to improve the user experience by identifying partial crawls.

**Use the “+” for the latest archived version**

The Web Continuity redirection component works by sending users to the latest captured version of a resource in the UK Government Web Archive, if the resource is no longer available on the live website. It uses the web archive to eliminate broken links on central government departments.

It can also be used ‘manually’ by using the [http://webarchive.nationalarchives.gov.uk/+](http://webarchive.nationalarchives.gov.uk/+)/ prefix. This takes the user to the latest captured version of the resource in the UK Government Web Archive, but any links on that resource will still point to the ‘live’ website.

For example, please see [http://webarchive.nationalarchives.gov.uk/+](http://webarchive.nationalarchives.gov.uk/+)[http://www.nationalarchives.gov.uk/](http://www.nationalarchives.gov.uk/)

**Browse by categories**

The National Archives UK Government Web Archive pages provide categorised lists and can be accessed from [http://www.nationalarchives.gov.uk/webarchive/](http://www.nationalarchives.gov.uk/webarchive/).

These contain links to the index page for the ‘homepage’ of all archived websites.

**Use web search engines**

The UK Government Web Archive is fully open to Google and other search engines and many of our users access its content in this way.

We hope that this document helps you to find the content you need. If you have any suggestions, comments or questions, please contact us at webarchive@nationalarchives.gsi.gov.uk

---

**The National Archives**

**October 2013**