

## Tips on Finding Content in the UK Government Web Archive

The UK Government Web Archive is freely accessible online and is maintained by the Web continuity department at The National Archives. We take snapshots of all central government websites, including central National Health Service websites, public inquiries and some regional bodies. The UK Government Web Archive also captures websites closing under the government's Website Review programme. Websites are typically crawled every six months, but this can vary. To date, the web archive contains approximately one billion documents.

Web archiving is a relatively new technology and is being constantly developed. We welcome suggestions on how we could improve this.

Further information:

<http://www.nationalarchives.gov.uk/webarchive/>

<http://www.nationalarchives.gov.uk/information-management/policies/web-continuity.htm>

You can find some useful and interesting information about web archiving, which explains some of the background and terminology, here:

[http://en.wikipedia.org/wiki/Web\\_archiving](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. The search functionality on websites does not work when they are archived, as along with some other navigational functions, including 'tick boxes' and certain drop-down menus.

### 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://collections.europarchive.org/tna/quick\\_search/](http://collections.europarchive.org/tna/quick_search/)

The search function can be a useful tool for locating resources, especially if what you are looking for is very specific, and it appeared on the web earlier than autumn 2008. As an example, try searching for "*Extractive Industries Transparency Initiative*".

An advanced search function is also available. 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 available at: [http://collections.europarchive.org/tna/adv\\_search/](http://collections.europarchive.org/tna/adv_search/)

Both full text search tools are, however, known to be limited in certain respects:

- 1) The Web Archive contains a lot of near-duplicate and duplicate material, which produces a lot of 'noise', or undesired results.
- 2) The full-text search tool is currently out-of-date. It only runs up to autumn 2008, missing a large amount of content since then. The index will be updated shortly.
- 3) Enhancements to searching, such as over a date range, are not available in the current version.

A help guide for search can be found at:  
[http://collections.europarchive.org/tna/search\\_help/](http://collections.europarchive.org/tna/search_help/)

The National Archives is currently working to provide users with easier ways to find content.

### **Check the \* index**

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/\\*/](http://webarchive.nationalarchives.gov.uk/*/)

For example if you wanted to see all instances of the Number 10 website you would add the url: <http://www.number10.gov.uk/> to the prefix to make this url:

[http://webarchive.nationalarchives.gov.uk/\\*/http://www.number10.gov.uk/](http://webarchive.nationalarchives.gov.uk/*/http://www.number10.gov.uk/)

You will then 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 (due to, 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. This method works for all resources contained in the Web Archive.

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 select 'Copy shortcut' (in Internet Explorer) or 'Copy link location' (in Firefox) 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:

[http://webarchive.nationalarchives.gov.uk/20100330193812/http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH\\_066309](http://webarchive.nationalarchives.gov.uk/20100330193812/http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_066309) 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](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. As a 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 plan to address this issue to improve the user experience by identifying partial crawls.

### **Use + index 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/+/> 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://www.number10.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/>.

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 engine indexing, 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](mailto:webarchive@nationalarchives.gsi.gov.uk)

---

**The National Archives**

**June 2011**