The National Archives

Finding Content in the UK Government Web Archive

This publication is available for download at nationalarchives.gov.uk.

Contents

Introduction

This document's purpose is to give advice on how to find resources in the UK Government Web Archive (UKGWA).

We welcome suggestions on how we could improve in this. You can contact us at webarchive@nationalarchives.gsi.gov.uk

Further information can be found on our webpages and on Wikipedia.

Please note that, in this document, 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 will not be available in the web archive. Search, "tick boxes" and certain drop-down menus will also not function.

Despite these limitations, we hope that the following suggestions will help you find what you need. There are several ways you can access content in the web archive and the method, or combination of methods, you choose will depend on your needs and your knowledge: how much you already know about what you are looking for, whether you know where it was located and what technical skills you have at your disposal.

Full-text search

The UKGWA has a full-text search tool which searches the text of webpages and most text-based file formats (e.g. PDF).

You can simply search using keywords, or you can perform narrower searches. For example, you can exclude certain terms from search results, restrict the search so that only results from archived versions of a particular website are displayed, or only those from archived websites in a particular area of government activity, for example “Environment”. Detailed guidance can be found at http://webarchive.nationalarchives.gov.uk/search_help/

Full text search in web archives are known to be limited in certain respects. The web archive contains much near-duplicate and duplicate material, over multiple dates, which produces some ‘noise’, or undesired results.

The A-Z list and categories

The A-Z list contains the more than 3,000 websites we have captured over the years. We’ve designed the list as a finding aid so, if you know the name of the organisation you are looking for, try a “find on this page”, “Ctrl+f” or “Command-F” search.

Check the * index

Our browse index (also known as the * index) provides access to all the versions of a particular resource 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/*/**. This method will work for all resources in the UKGWA.

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

If the page has even been archived you will see an index page with dated links to all the versions of the page contained in the web archive. These dates are the dates on which the resource was captured. If there are no captures of the resource, an error page will be given.

Another typical example follows. 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, 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. This displays the * index page for a particular resource.

The * index Bookmarklet

There is a bookmarklet available on our information on web archiving page. This will help you quickly check the * index page to see if and when a resource has been archived. Simply navigate to the resource on the live website in your browser and then activate the bookmarklet. To install the bookmarklet, drag it from the above page and into your bookmarks or favorites bar. This should work in most browsers.

Use the + 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 UKGWA, if the resource is no longer available on the live website. It uses the web archive to reduce broken links on central government websites.

It can also be used manually by prefixing any URL with **http://webarchive.nationalarchives.gov.uk/+/** . This will access the latest captured version of the resource in the 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.nationalarchives.gov.uk/

Memento

Memento is a tool that allows the user to browse through different representations of a website resource by using web archives. For further information see http://www.mementoweb.org/

See http://www.nationalarchives.gov.uk/webarchive/information.htm#memento for UKGWA implementation. You can configure Memento for use only with UKGWA, but you can add other “Timegates” to search other web archives too.

A note on “complete crawls” and “partial crawls”

Complete crawls are those that are targeted and go through our quality assurance processes. As most websites are crawled 2-3 times per year, if the frequency of captures for a resource is greater than this, it is likely that some of these are 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 an extra date can appear on the * index page, or an extra resource in the + index, but many of the links may be broken or will resolve to a date either before or after the date of the archived website being viewed. This can be confusing to our users as complete crawls and partial crawls cannot currently be differentiated. However, we are planning to address this to improve the user experience by identifying partial crawls. If you need to know when a complete crawl of a website was made, please contact us.

Application Programming Interfaces (APIs)

Full-text Search API

Access to an RSS-based full text search API is available on request. It is based on the technical specification described at http://web.archive.org/web/20120514083520/http://opensearch.a9.com/spec/opensearchquerysyntax/1.0/.

The results are in OpenSearch 1.0 RSS 2.0 format and they are compatible with http://www.opensearch.org/Specifications/OpenSearch/1.1#OpenSearch_response_elements

Please note that no more than 10 results at a time can be requested, but you can use the "start" parameter to get more results by issuing other queries.

If you would like access to this API, please contact us.

Wayback-Style API

This is an XML-based, publicly-accessible API.

http://webarchive.nationalarchives.gov.uk/xmlquery?type=urlquery&startdate=20090101000000&enddate=20130131000000&url=http://number10.gov.uk/

It is an implementation of the replay and capture Internet Archive Wayback API. It follows the specifications at http://wwwoh-access.archive.org/wwwoh/waybackapi.htm with the below alterations.

To avoid any ambiguity as to which URL the query parameters belong to, the "url" query parameter must be last. Everything after it will be considered part of the target URL.

Capture request

The digest value is a hash of the whole resource, HTTP headers included.

Pagination is not implemented. The capture query always returns the results starting from the first one; wayback.request.resultsrequested is -1.

The following tags have not been implemented:

* wayback.results.result.file

* wayback.results.result.compressedoffset

Replay

No rewriting is performed (it is just like the publicly-accessible web archive). The resource is served as it was captured.

For example, http://webarchive.nationalarchives.gov.uk/replay?date=20130109092234&url=http://number10.gov.uk is identical to http://webarchive.nationalarchives.gov.uk/20130109092234/http://number10.gov.uk

Preservation of Web Archive Data

UKGWA "replays" the data crawled on the web over the web, via the http://webarchive.nationalarchives.gov.uk/ prefix. Therefore, all web data types are captured (.html, .css, .js, .xml, .gif, .pdf and so on) and most can be replayed.

UKGWA data is held in ARC files (see http://www.digitalpreservation.gov/formats/fdd/fdd000235.shtml) for both access and preservation.

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.gov.uk