Introduction to Web Archives for Research Use

Web Archive Engagement Manager, British Library
The Legal Deposit Libraries

Bodleian Libraries
UNIVERSITY OF OXFORD

National Library of Scotland
Leabharlann Nàiseanta na h-Alba

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

UK WA
UK WEB ARCHIVE

UNIVERSITY OF CAMBRIDGE

LLYFRGELL GENEDLAETHOL CYMRU
THE NATIONAL LIBRARY OF WALES

BL.uk
What is a web archive?
What is the UK Web Archive?

The UK Web Archive (UKWA) collects millions of UK websites each year, preserving them for future generations.
UKWA Mission

Collect the entire UK web space
How long have we been archiving the UK web?

2005 - Collecting websites on a selective basis only, with permission of the owners. Thousands of websites in 15 years.

2013 - Collecting websites under Non-Print Legal Deposit Act regulations (without owners permission). Millions of websites every year.
How frequently do we collect the web?

‘Everything’ - Once per year as part of our ‘Annual Domain Crawl’

Selected ‘targets’ (including News) - Daily, Weekly, Monthly, Quarterly, Six-monthly
Access to the collections

**Thousands of websites** - Available from anywhere

**Millions of websites** - Only available in the reading rooms of a UK Legal Deposit Library (nine locations).
www.webarchive.org.uk (1)

Search the UK Web Archive

Enter a specific website URL (e.g. www.bl.uk) or any word or phrase...

What we do

The UK Web Archive (UKWA) collects millions of websites each year, preserving them for future generations. Use this site to discover old or obsolete versions of UK websites, search the text of the websites and browse websites curated on different topics and themes.
100s of curated collections

**FTSE 100**
This collection, curated by staff at the British Library...

**Family History**
A collection of websites curated by staff at the UK Legal...

**Fashion**
Websites selected around the theme of fashion including t...

**Festivals**
A collection of websites related to the various festivals...

**First World War Centenary, 2014-18**

**Food Archive**
A food related collection

**Forth Bridge 125th Anniversary**

**G8 Summit 2005**
Selection of web sites
This collection of websites was curated between 2012 and 2014 by Dr. Saskia Huc-Hepher, a Senior Lecturer at the University of Westminster, on the subject of the French community in London. The Collection was a fundamental component of Dr. Huc-Hepher’s thesis on the French community in London. It is hoped that the collection will serve both as an innovati...
JISC UK Web Domain Dataset (1996-2013)

In partnership with the Internet Archive and JISC, we have obtained access to the subset of the Internet Archive’s web collection that relates to the UK. The JISC UK Web Domain Dataset (1996-2013) contains all of the resources from the Internet Archive that were hosted on domains ending in ‘.uk’, or that are required in order to render those UK pages.

The collection was deposited with us in two separate tranches. The 1996-2010 tranche is composed of 470,466 files (mostly .arc.gz) and the total size is 32TB. The 2011-2013 tranche runs up to April of 2013 (i.e. until the enactment of the UK’s Non-Print Legal Deposit legislation), is composed of 203,502 files with a total size of 30TB. The first version of this dataset corresponded to the first tranche, whereas the current version of the dataset includes both tranches.

This dataset cannot be made generally available here, but can be used to generate secondary datasets, and these can be made available under open license terms.

Before interpreting results from this dataset, or any secondary-datasets based upon it, please refer to the known issues with this dataset.

Secondary datasets

- JISC UK Web Domain Dataset (1996-2010) Format Profile
- JISC UK Web Domain Dataset (1996-2010) Geoindex
- JISC UK Web Domain Dataset (1996-2010) Host Link Graph
- JISC UK Web Domain Dataset (1996-2013) Crawled URL Index
SHINE - www.webarchive.org.uk/shine (1)
GLAM Workbench - Jupyter Notebooks

Compare two versions of an archived web page

Works with AWA, NZWA, IA, & UKWA

This notebook demonstrates a number of different ways of comparing versions of archived web pages. Just choose a repository, enter a url, and select two dates to see comparisons based on: page metadata, basic statistics such as file size and number of words, numbers of internal and external links, cosine similarity of text, line by line differences in text or code, and screenshots.

- Download from GitHub
- View using NBViewer
- Run live in Appmode on Binder

Observing change in a web page over time

Find all the archived versions of a web page
Harvesting collections of text from archived web pages
Harvesting data about a domain using the IA CDX API
Find and explore Powerpoint presentations from a specific domain
Exploring subdomains in the whole of gov.au
Exploring change over time
Compare two versions of an archived web page
Observing change in a web page over time
Create and compare full page screenshots from archived web pages
Using screenshots to visualise change in a page over time
Display changes in the text of web pages
glam-workbench.net/web-archives/
Other web archives

- UK Government Web Archive: https://www.nationalarchives.gov.uk/webarchive/
- UK Parliamentary Web Archive: https://archives.parliament.uk/online-resources/web-archive/
- Common Crawl: http://commoncrawl.org/
- Internet Archive: https://archive.org/
Useful UKWA links

Email: jason.webber@bl.uk
Twitter: @UKWebArchive

www.webarchive.org.uk
www.webarchive.org.uk/shine
data.webarchive.org.uk
www.webarchive.org.uk/blog
glam-workbench.github.io/
Thank you