

# Introduction to Web Archives for Research Use

Web Archive Engagement Manager, British  
Library

# The Legal Deposit Libraries

**UKWA**  
UK WEB ARCHIVE



**Bodleian Libraries**  
UNIVERSITY OF OXFORD



**UNIVERSITY OF  
CAMBRIDGE**



**National Library of Scotland**  
Leabharlann Nàiseanta na h-Alba



**LLYFRGELL GENEDLAETHOL CYMRU  
THE NATIONAL LIBRARY OF WALES**

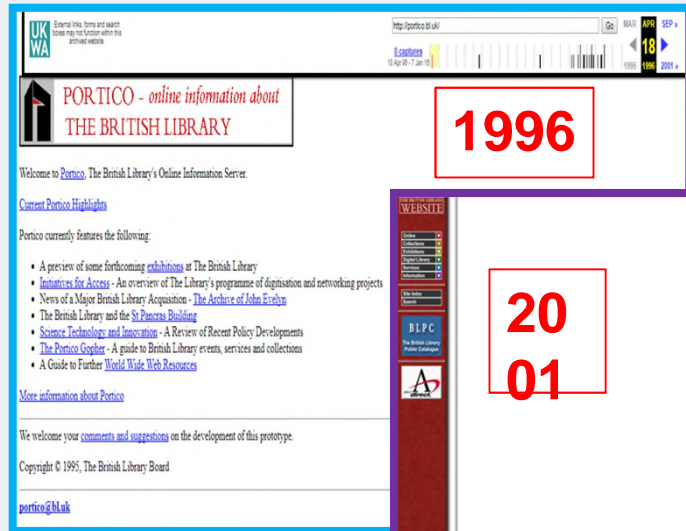


**Trinity College Dublin**  
Coláiste na Tríonóide, Baile Átha Cliath  
The University of Dublin

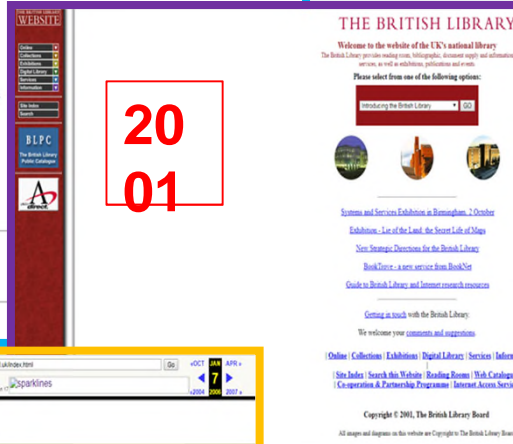
**LIBRARY  
HSILIRB**

**LIBRARY  
HSILIRB**

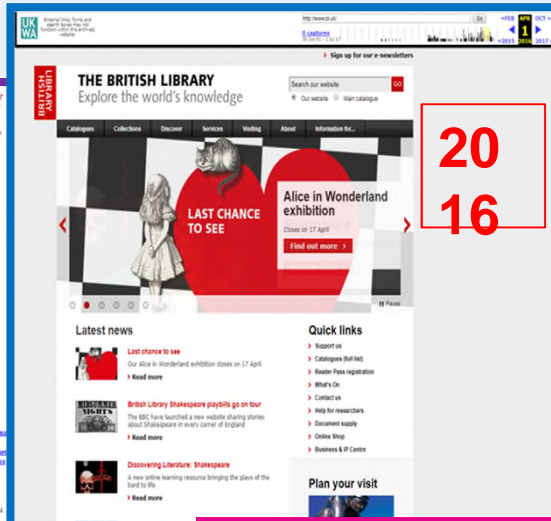
# What is a web archive?



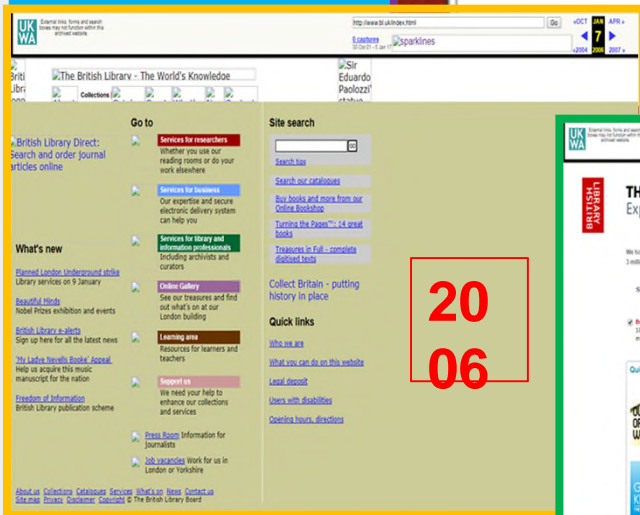
1996



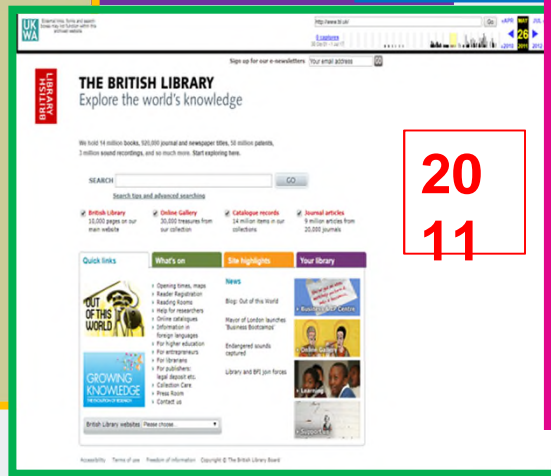
2001



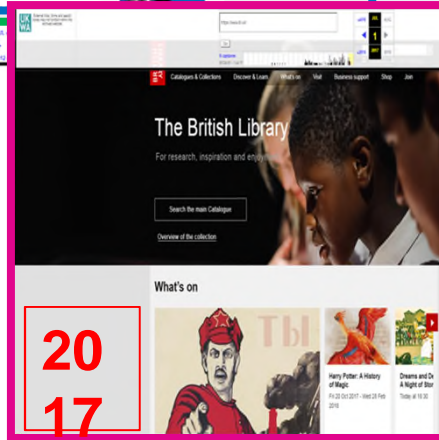
2016



2006



2011



2017

# Portico

The screenshot shows the Portico website interface. At the top, there is a navigation bar with a search box containing 'http://portico.bl.uk/' and a 'Go' button. To the right of the search box is a calendar for April 1995, with the 18th highlighted. Below the search box, there is a link for '17 captures' and a date range '18 Apr 95 - 7 Jan 16'. On the far right of the navigation bar are links for 'Close', 'Cymraeg', and 'Help'. Below the navigation bar is a header section with the Portico logo (a stylized 'A' shape) and the text 'PORTICO - online information about THE BRITISH LIBRARY'. The main content area starts with a welcome message: 'Welcome to [Portico](#). The British Library's Online Information Server.' This is followed by a link for 'Current Portico Highlights'. Below this, it states 'Portico currently features the following:' and lists six bullet points: 'A preview of some forthcoming [exhibitions](#) at The British Library', '[Initiatives for Access](#) - An overview of The Library's programme of digitisation and networking projects', 'News of a Major British Library Acquisition - [The Archive of John Evelyn](#)', 'The British Library and the [St Pancras Building](#)', '[Science Technology and Innovation](#) - A Review of Recent Policy Developments', and '[The Portico Gopher](#) - A guide to British Library events, services and collections'. Below the list is a link for 'More information about Portico'. At the bottom of the main content area, there is a message: 'We welcome your [comments and suggestions](#) on the development of this prototype.' and a copyright notice: 'Copyright © 1995, The British Library Board'. At the very bottom, there is a link for 'portico@bl.uk'.

# 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](http://www.webarchive.org.uk) (1)



[Home](#) [Topics and Themes](#) [Save a UK website](#) [About Us](#) [Contact Us](#)

Language ▾

## Search the UK Web Archive

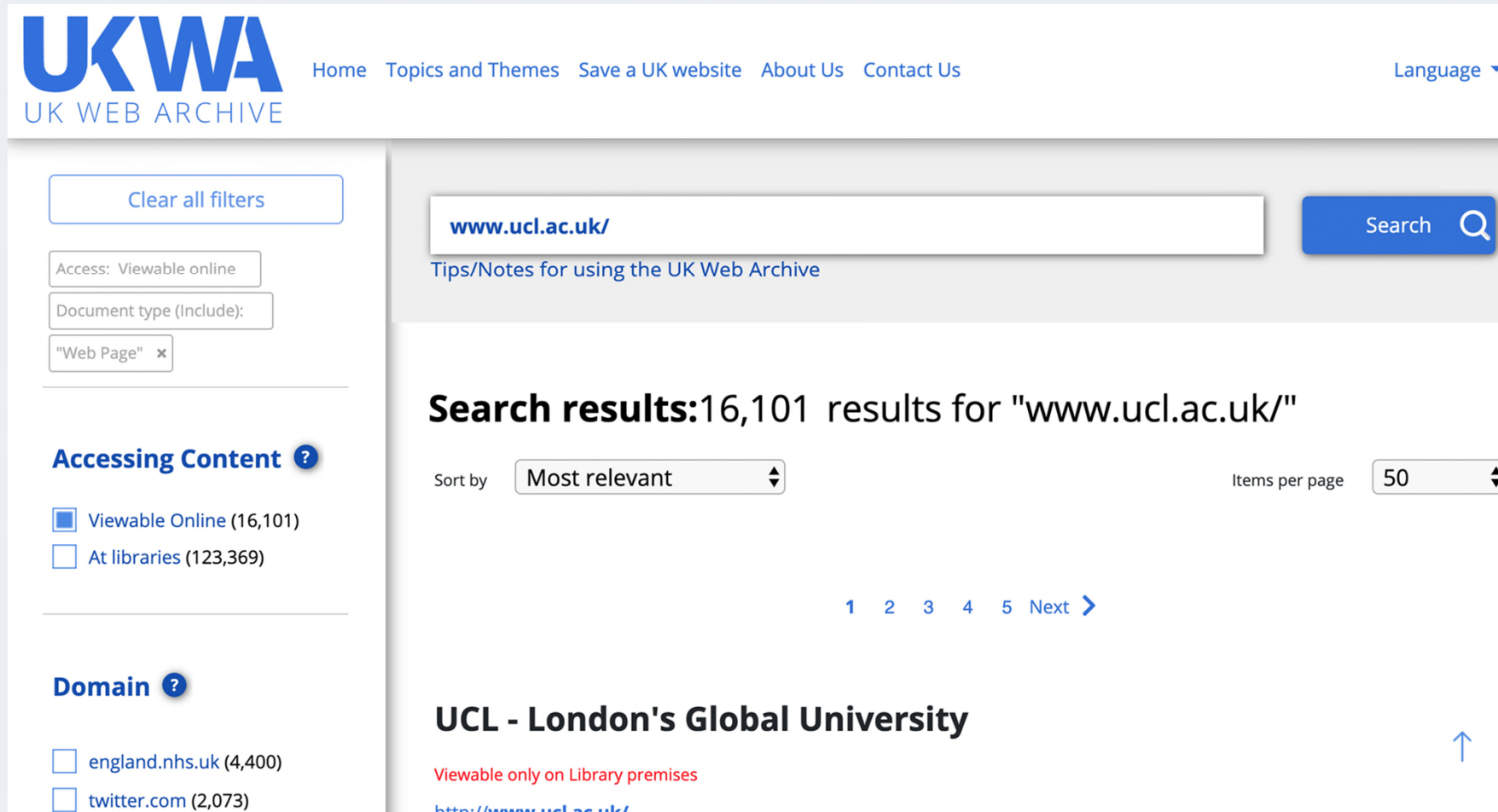
Search 🔍

Enter a specific website URL (e.g. [www.bl.uk](http://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.

# [www.webarchive.org.uk](http://www.webarchive.org.uk) (2)



The screenshot shows the UKWA (UK Web Archive) search interface. At the top left is the UKWA logo and navigation links: Home, Topics and Themes, Save a UK website, About Us, and Contact Us. A language dropdown menu is located at the top right. On the left sidebar, there are filter options: 'Clear all filters', 'Access: Viewable online', 'Document type (Include):' with a selected 'Web Page' tag, 'Accessing Content' with a help icon, and 'Domain' with a help icon. Under 'Accessing Content', 'Viewable Online (16,101)' is selected. Under 'Domain', 'england.nhs.uk (4,400)' and 'twitter.com (2,073)' are listed. The main search area contains a search bar with 'www.ucl.ac.uk/' and a 'Search' button. Below the search bar is a link to 'Tips/Notes for using the UK Web Archive'. The search results section displays 'Search results: 16,101 results for "www.ucl.ac.uk/"'. It includes sorting options ('Sort by: Most relevant') and 'Items per page' (set to 50). A pagination bar shows '1 2 3 4 5 Next >'. The first result is 'UCL - London's Global University', marked as 'Viewable only on Library premises' and with a blue upward arrow icon. The URL 'http://www.ucl.ac.uk/' is partially visible below the result.

# 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

# London French Special Collection


**UKWA**  
UK WEB ARCHIVE

[Home](#) [Topics and Themes](#) [Save a UK website](#) [About Us](#) [Contact Us](#)

Language ▾

## London French Special Collection

You are here: [Topics and Themes](#) / London French Special Collection

This collection of websites was curated inbetween 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... 

# Open data - data.webarchive.org.uk

home » ukwa.ds.2

## JISC UK WEB DOMAIN DATASET (1996-2013)

[Introduction](#)

[Format Profile](#)

[Geoindex](#)

[Host-level Links](#)

[Crawled URL Index](#)

## UK SELECTIVE WEB ARCHIVE

[Introduction](#)

[Website Classification Dataset](#)

## PROJECTS

[Big UK Domain Data for the Arts and Humanities](#)

[Analytical Access to the Domain Dark Archive](#)

## 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](#)

# [data.webarchive.org.uk](http://data.webarchive.org.uk)

H	I	J	K	L	M
		20071102151717/http://www.wwt.org.uk:80/article/217/502/toad_hall_to_open_at_slimbrick	SE23 3EP		
		20071102152122/http://www.thebathweb.co.uk:80/mtree/Carpenters/	SE23 3EP		
		20071102152122/http://www.thebathweb.co.uk:80/mtree/Carpenters/	SE23 3EP		
		20071102152122/http://www.thebathweb.co.uk:80/mtree/Carpenters/	SE23 3EW		
		20071102152122/http://www.thebathweb.co.uk:80/mtree/Carpenters/	SE23 3EW		
		20071102152122/http://www.thebathweb.co.uk:80/mtree/Carpenters/	SE23 3HA		
		20071102152122/http://www.thebathweb.co.uk:80/mtree/Carpenters/	SE23 3HA		
		20071102152122/http://www.thebathweb.co.uk:80/mtree/Carpenters/	SE23 3HA		
		20071102152122/http://www.thebathweb.co.uk:80/mtree/Carpenters/	SE23 3HA		
		20071102152122/http://www.thebathweb.co.uk:80/mtree/Carpenters/	SE23 3HA		
		20071102152122/http://www.thebathweb.co.uk:80/mtree/Carpenters/	SE23 3HF		
		20071102152122/http://www.thebathweb.co.uk:80/mtree/Carpenters/	SE23 3HF		
		20071102152122/http://www.thebathweb.co.uk:80/mtree/Carpenters/	SE23 3HF		
		20071102152122/http://www.thebathweb.co.uk:80/mtree/Carpenters/	SE23 3HF		
		20071102152122/http://www.thebathweb.co.uk:80/mtree/Carpenters/	SE23 3HF		
		20071102152122/http://www.thebathweb.co.uk:80/mtree/Carpenters/	SE23 3HF		
		20071102152122/http://www.thebathweb.co.uk:80/mtree/Carpenters/	SE23 3HG		
		20071102152122/http://www.thebathweb.co.uk:80/mtree/Carpenters/	SE23 3HN		
		20071102152122/http://www.thebathweb.co.uk:80/mtree/Carpenters/	SE23 3HN		
		20071102152122/http://www.thebathweb.co.uk:80/mtree/Carpenters/	SE23 3HN		
		20071102152324/http://www.thebathweb.co.uk:80/mtree/Home/Carpet_and_Rugs/	SE23 3HN		
		20071102152324/http://www.thebathweb.co.uk:80/mtree/Home/Carpet_and_Rugs/	SE23 3HN		
		20071102152324/http://www.thebathweb.co.uk:80/mtree/Home/Carpet_and_Rugs/	SE23 3HN		
		20071102152324/http://www.thebathweb.co.uk:80/mtree/Home/Carpet_and_Rugs/	SE23 3HN		
		20071102152324/http://www.thebathweb.co.uk:80/mtree/Home/Carpet_and_Rugs/	SE23 3HN		
		20071102152324/http://www.thebathweb.co.uk:80/mtree/Home/Carpet_and_Rugs/	SE23 3HN		
		20071102152324/http://www.thebathweb.co.uk:80/mtree/Home/Carpet_and_Rugs/	SE23 3HN		
		20071102152324/http://www.thebathweb.co.uk:80/mtree/Home/Carpet_and_Rugs/	SE23 3HN		

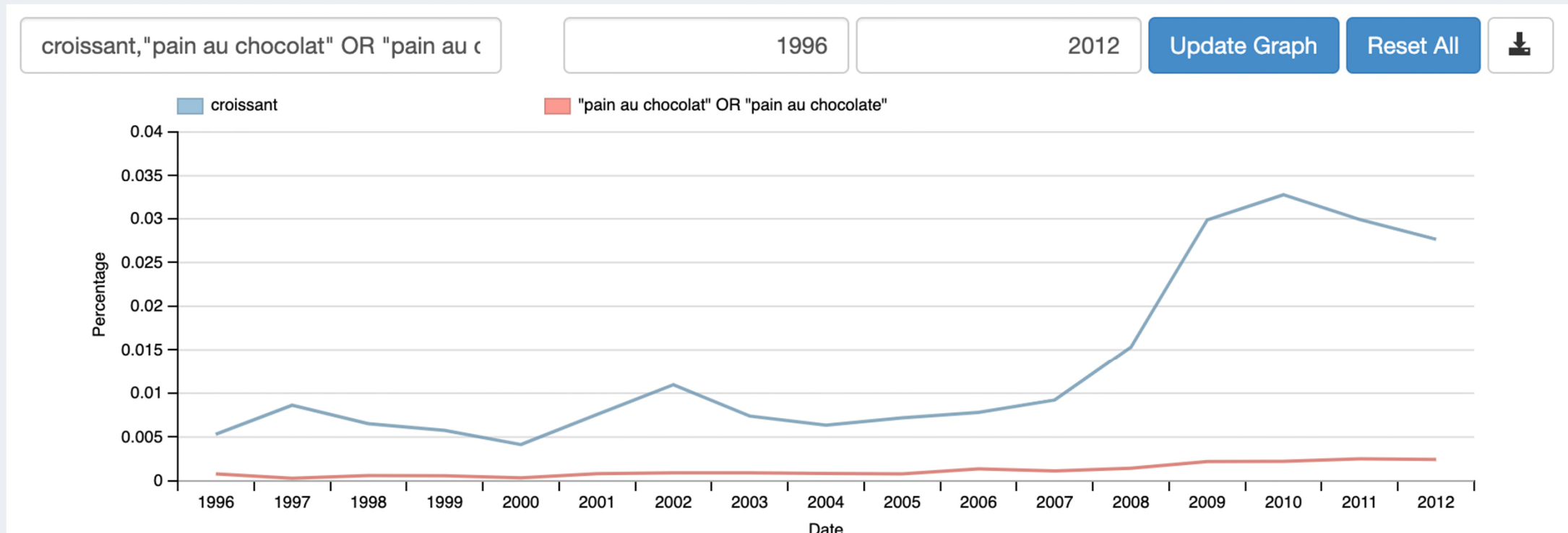
# SHINE - [www.webarchive.org.uk/shine](http://www.webarchive.org.uk/shine) (1)

The screenshot displays the SHINE search interface. At the top, there is a navigation bar with 'UK Web Archive', 'Search', and 'Trends'. Below this, there are tabs for 'Search', 'Advanced Search', and 'Search Tips'. The main area is divided into several sections:

- Filters:**
  - Postcode District:** 10 items. Includes SS0 (23,664), OX1 (4,976), SW1E (4,940), CB2 (4,685), NR3 (3,965), and a 'Show more...' link.
  - Crawl Year:** 10 items. Includes 2004 (91,462), 2012 (91,390), 2006 (65,289), 2008 (64,070), 2007 (63,614), and a 'Show more...' link.
  - Specific Content Type:** 10 items. Includes text/html (424,365).
- Search Bar:** A search input field containing 'aarhus' with 'Sample Mode' dropdown, 'Search', and 'Reset' buttons.
- Search Term(s):** A light blue box displaying '• Search Term(s): aarhus'.
- Results:** A section with 'Results' and 'Concordance' tabs. Below the tabs, it shows 'Results 1 to 10 of 704,028' and options for 'CSV', 'Crawl Date', and 'Asc'. A pagination bar shows '1' selected, with '2', '3', '4', '5', '6' and '«', '»' buttons. An 'Action' dropdown button is also present.



# SHINE - [www.webarchive.org.uk/shine](http://www.webarchive.org.uk/shine) (2)



[buddah.projects.history.ac.uk/bursaries](http://buddah.projects.history.ac.uk/bursaries)

## Big UK Domain Data for the Arts and Humanities



[About](#) [Blog](#) [Contact](#) [Documentation](#) [News and events](#) **[Bursaries](#)**

# 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

# [glam-workbench.net/web-archives/](http://glam-workbench.net/web-archives/)

Jupyter

Edit App

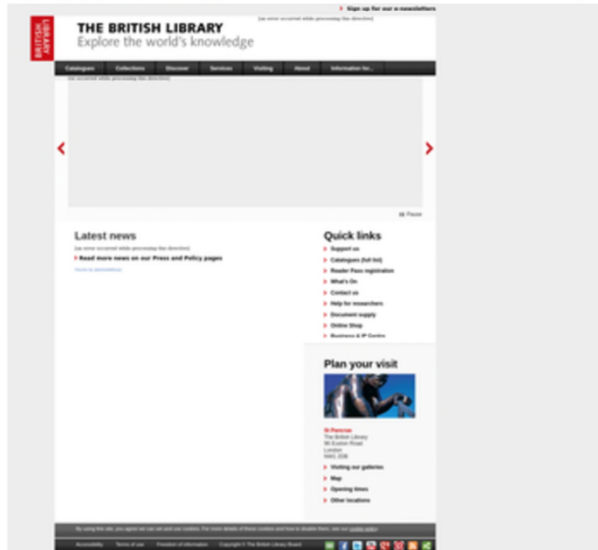
Visit repo

Copy Binder link

## Screenshots

29 April 2014

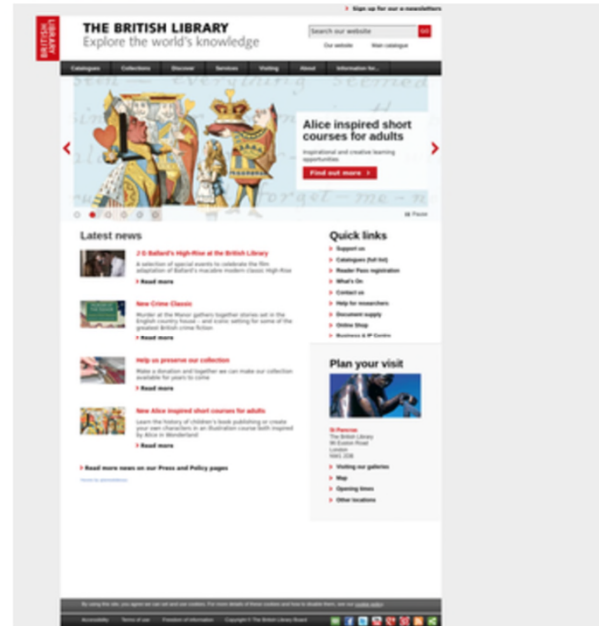
<https://www.bl.uk/>



[Download]

15 February 2016

<http://www.bl.uk/>



# 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/>
- National Library of Ireland: [https://www.nli.ie/en/web\\_archive.aspx](https://www.nli.ie/en/web_archive.aspx)
- Common Crawl: <http://commoncrawl.org/>
- Internet Archive: <https://archive.org/>

# Useful UKWA links

Email: [jason.webber@bl.uk](mailto:jason.webber@bl.uk)

Twitter: @UKWebArchive

[www.webarchive.org.uk](http://www.webarchive.org.uk)

[www.webarchive.org.uk/shine](http://www.webarchive.org.uk/shine)

[data.webarchive.org.uk](http://data.webarchive.org.uk)

[www.webarchive.org.uk/blog](http://www.webarchive.org.uk/blog)

[glam-workbench.github.io/](http://glam-workbench.github.io/)

BRITISH LIBRARY

Thank  
you

