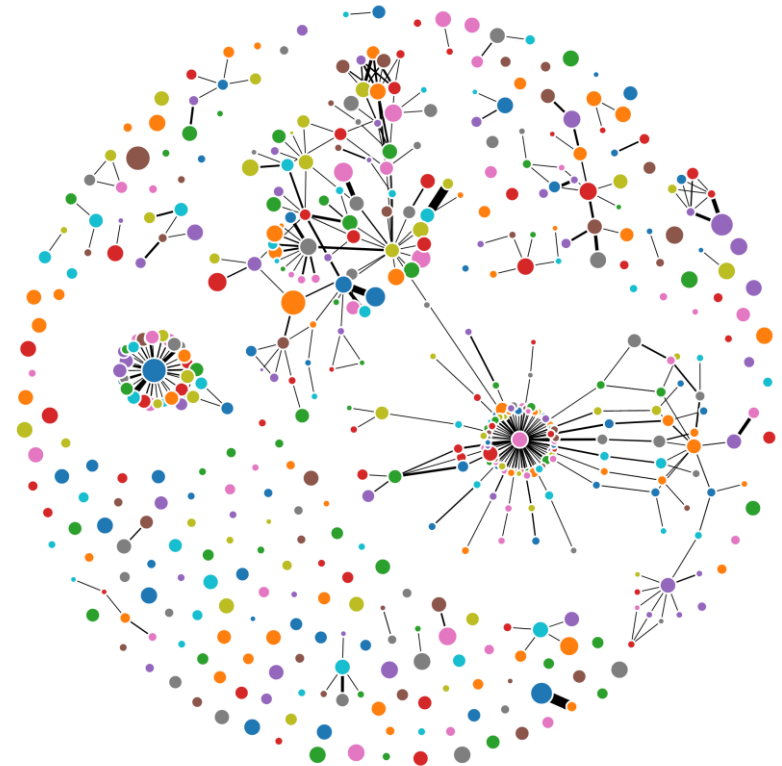


Using secondary datasets for researchers under a legal deposit framework

Jason Webber

UKWA Goals

- Capture and Preserve the UK web space
- Support access to the collection
- Enable research



UKWA Collections

1. Licenced (Full open access)
2. Legal Deposit (Library access only)
3. JISC Domain dataset (...kind of full access)

Access Paradox

Cymraeg



Home

Topics and Themes

Save a UK website

About Us

Contact Us

Clear all filters

Access: Viewable online

Document type (Include):

"Web Page" x

Accessing Content ?

Viewable Online (11,941,152)

At libraries (24,583,363)

Domain ?

gov.uk (4,779,398)

cwgc.org (892,987)

britishfuture.org (585,311)

[+ Show more](#)

Document Type ?

"first world war"

Search

Tips/Notes for using the UK Web Archive

Search results: 11,941,152 results for ""first world war""

Sort by

Items per page

1 2 3 4 5 Next >

Janus: First World War (1914-1918)

[http://janus.lib.cam.ac.uk/db/node.xsp?id=CV%2FSubject%2FFirst%20World%20War%20\(1914-1918\)](http://janus.lib.cam.ac.uk/db/node.xsp?id=CV%2FSubject%2FFirst%20World%20War%20(1914-1918))

First World War (1914-1918) Fish Fishing Industry Flags Folklore Food See later --> Search Janus

Date collected: 2013-05-13

Events for World War One Audit of Surviving Remains PROJECT PROJE

The Size Problem

Cymraeg



Home

Topics and Themes

Save a UK website

About Us

Contact Us

Clear all filters

Access: Viewable online

Document type (Include):

"Web Page" x

Accessing Content ?

Viewable Online (1,045,343)

At libraries (2,285,636)

Domain ?

britishfuture.org (583,248)

ladyadventurer.co.uk (84,879)

cwgc.org (58,233)

[+ Show more](#)

Ypres

Search

Tips/Notes for using the UK Web Archive

Search results: 1,045,343 results for "Ypres"

Sort by

Items per page

1 2 3 4 5 Next >

Cardiff High School - Ypres and the Somme 2013

<http://www.cardiffhigh.cardiff.sch.uk/gallery/?pid=2&gcatid=1&albumid=40>

Lottery School Lottery Ypres and the Somme 2013 Breaking News Photo Gallery Oriol ffotograffau News
Date collected: 2014-12-03

Battle Book Of Ypres - The National Archives Bookshop



How to reach researchers within the law?

Non-Print Legal Deposit Act
2003 prohibits access outside
of UK Legal Deposit Library
control.

Secondary data?

Facts about the collection that cannot be used to recreate or reconstruct the originals.

data.webarchive.org.uk

What Secondary data is available?

- Format profiles
- Geoindex
- Host links
- Crawled URL index

data.webarchive.org.uk

What Secondary data is available?

UKWA Open Data Home GitHub UKWA

home

UK Web Archive Open Data

JISC UK WEB DOMAIN DATA SET (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](#)
[Big Data: Demonstrating the Value of the UK Web Domain Dataset for Social Science Research](#)
[SCAPE: Scaleable Preservation Environments](#)

RELATED WORK
[Common Crawl](#)
[httparchive.org](#)

In order to facilitate research, and so we might better understand and preserve the UK's web history, the UK Web Archive has decided to make a number of data and API services available for general use. We also make a few example tools available, showing how the open data might be used, and these are hosted in this [GitHub repository](#).

We hope that by making these datasets available, the broader community will find interesting ways to re-use, explore and visualise the contents of our web archive. We are keen to work with any interested parties to exploit these datasets, and understand what other derivative or summary data would be of interest.

Datasets, Tools & APIs

Open Datasets


In general, we can't provide remote bulk access to the primary datasets listed above (although bulk access can be arranged for particular projects - see below). This is mainly because the conditions underwhich we hold the content do not permit it, but also because the data sets are very large and so providing bulk downloads is not practical.

However, secondary datasets, composed of metadata that describes simple facts about the content, can be made available under open terms.

For all the details, follow the links on the left, or look through the README files and code in [the GitHub repository](#).

Tools

We also make a few tools available, which illustrate how the open datasets might be used. These are hosted on GitHub, and you should feel free to fork our repository or download our tools using the links above. Pull requests containing new or improved tools are welcome!



data.webarchive.org.uk

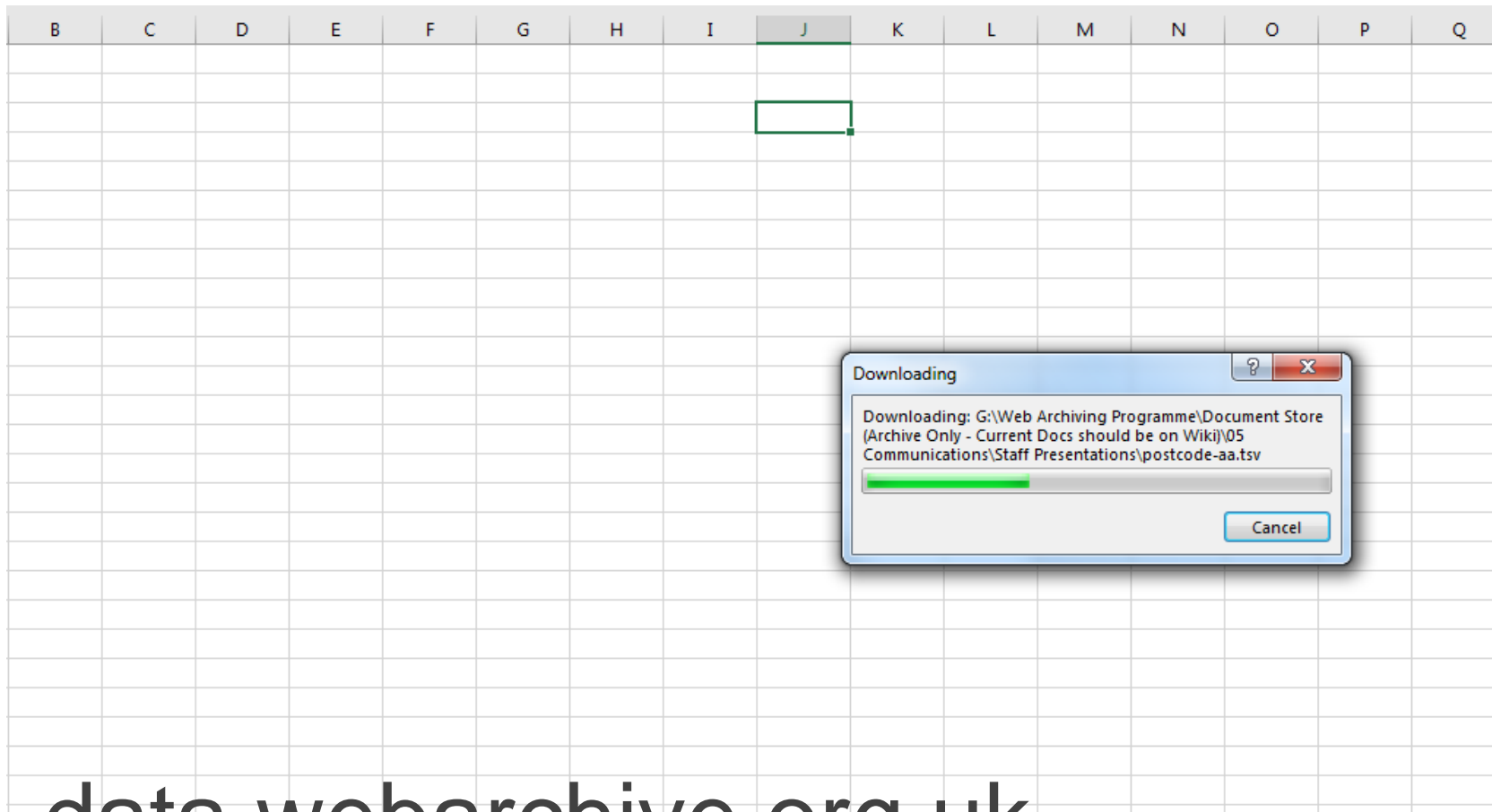
What Secondary data is available?

Index of /datasets/ukwa.ds.2/geo/1996-2010/

../		
east-london.tsv.gz	19-Jul-2012 15:06	195367777
manifest.sha1	22-May-2012 06:44	744
manifest.sha256	18-Dec-2017 11:20	1032
postcode-aa.tsv.bz2	21-May-2012 15:34	879794142
postcode-ab.tsv.bz2	21-May-2012 15:34	410923672
postcode-ac.tsv.bz2	21-May-2012 15:35	535704896
postcode-ad.tsv.bz2	21-May-2012 15:35	558723154
postcode-ae.tsv.bz2	21-May-2012 15:35	631921635
postcode-af.tsv.bz2	21-May-2012 15:36	649547036
postcode-ag.tsv.bz2	21-May-2012 15:36	755675083
postcode-ah.tsv.bz2	21-May-2012 15:37	740100032
postcode-ai.tsv.bz2	21-May-2012 15:37	694866562
postcode-aj.tsv.bz2	21-May-2012 15:38	664782309
postcode-ak.tsv.bz2	21-May-2012 15:38	508020513
postcode-al.tsv.bz2	21-May-2012 15:39	1539549873
summary-count-year-postcode.tsv.gz	19-Jul-2012 16:05	56962848

data.webarchive.org.uk

Waiting....



data.webarchive.org.uk

What do you get?

H	I	J	K	L	M
		20071102151717/http://www.wwt.org.uk:80/article/217/502/toad_hall_to_open_at_slimbric	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		

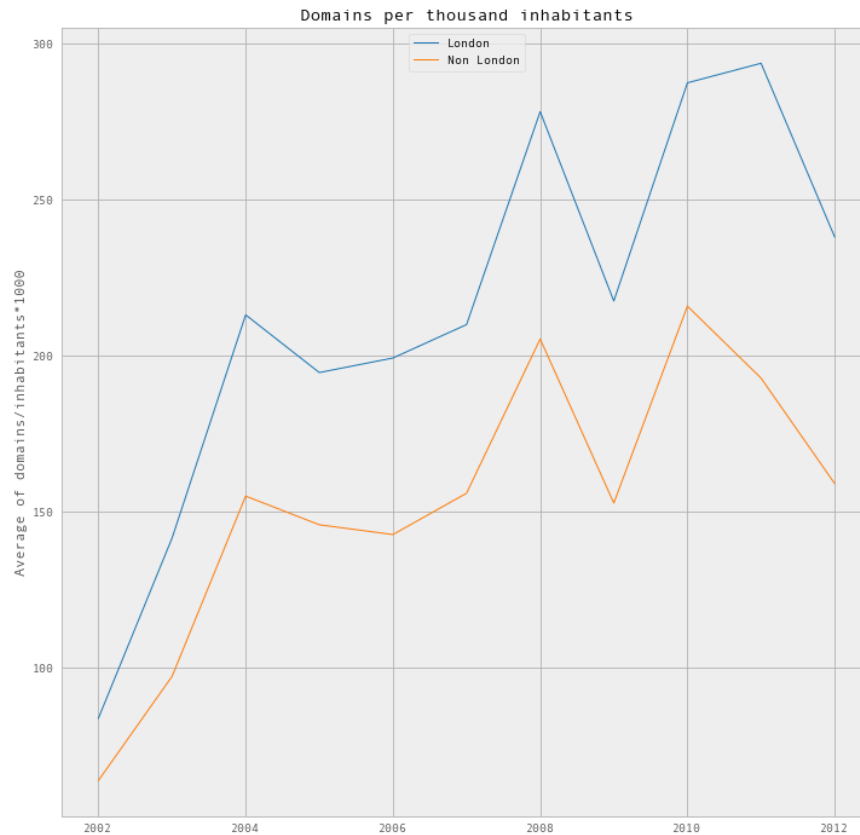
data.webarchive.org.uk

Case Study 1 – Geographical research

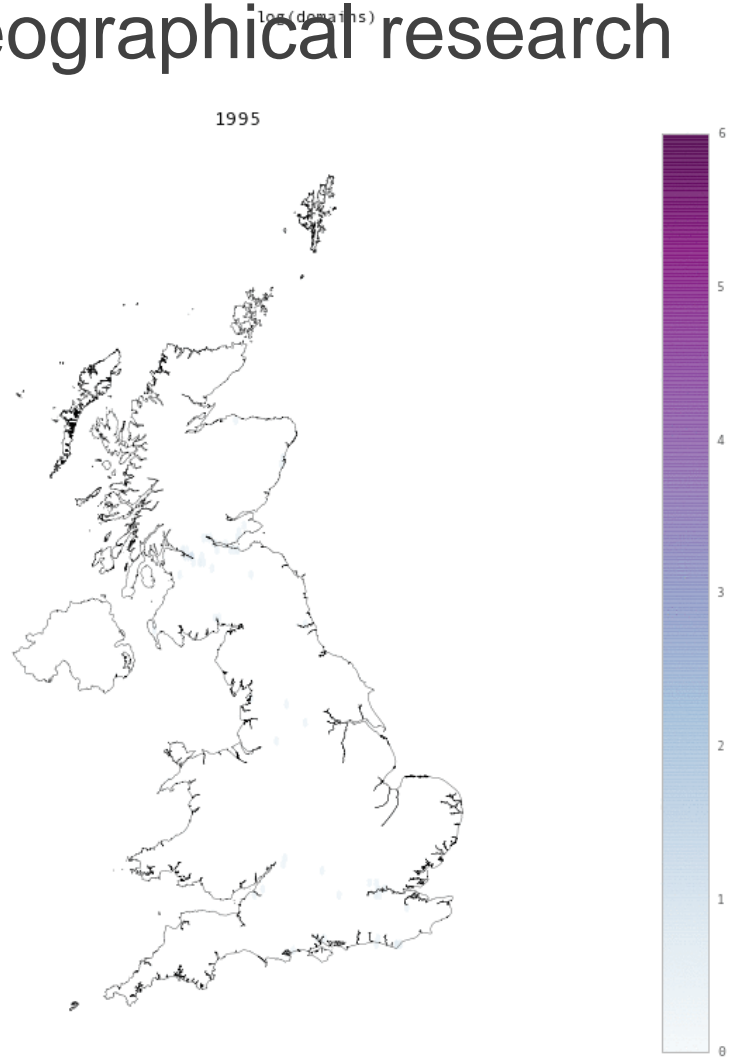
“If you are a quantitative social scientist there are few things more fascinating than free, under-utilised, quirky and easy to download data that also fits well the narrative of 'big data'.”

Emmanouil Tranos, University of Birmingham and Alan Turing Fellow

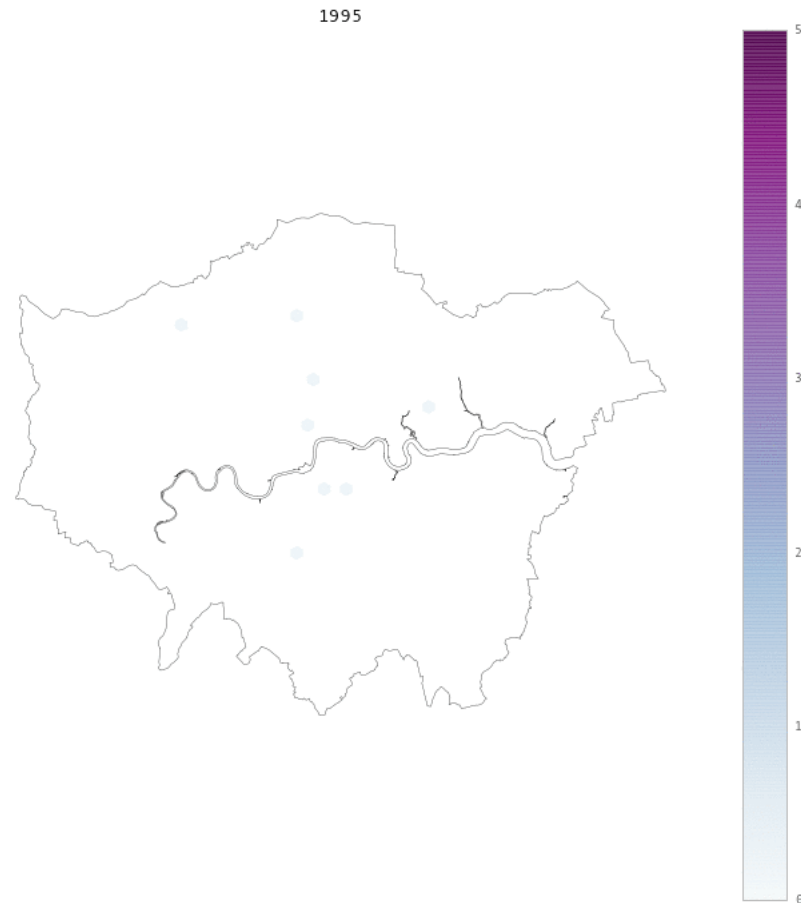
Case Study 2 – Geographical research



Case Study 2 – Geographical research



Case Study 2 – Geographical research



Case Study 2 – Geographical research

Hypothesis

“The availability of internet content of local interest can attract people online in order to access and take advantage of the potential on-line opportunities such as accessing local products and services. *The first results seem to support our hypothesis.*”

[BL blog post by Emmanouil Tranos](#)

Case Study 2 – What is a national web?

“Understanding the limitations of the ccTLD as a proxy for the national web: lessons from cross-border religion in the Northern Irish web sphere.”

Peter Webster, Historian and Consultant

Case Study 2 – What is a national web?

peterwebster.me/2019/04/04/understanding-national-web-domains/

peterwebster.me/2018/04/26/ukcreationism/

@pj_webster

Next steps 1

Make it easier!

Give some training?
Workshop?

Next steps 2

Spread the word and
get involved!

@UKWebArchive

data.webarchive.org.uk