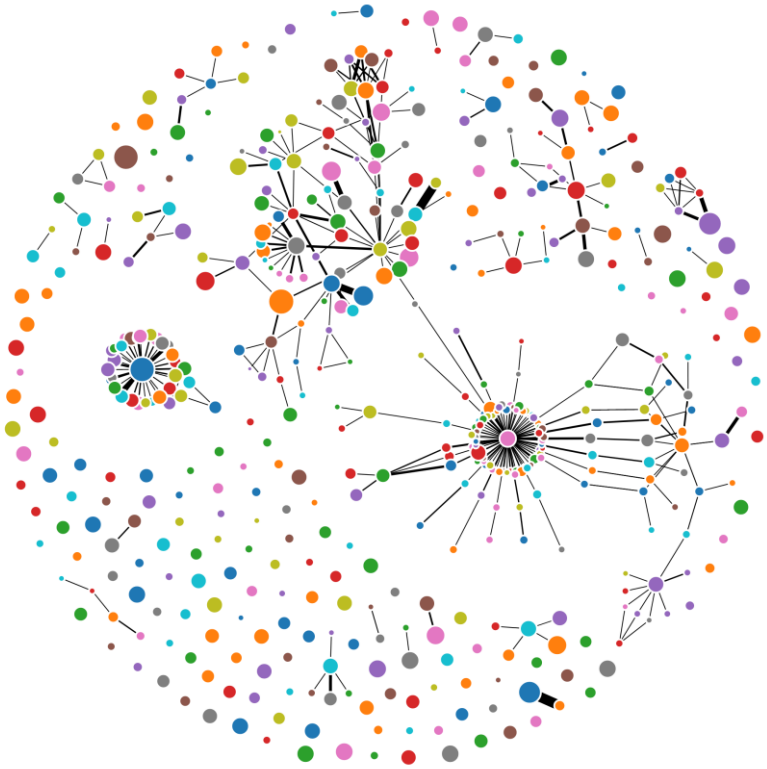
Using secondary datasets for researchers under a legal deposit framework

Jason Webber

UKWA Goals

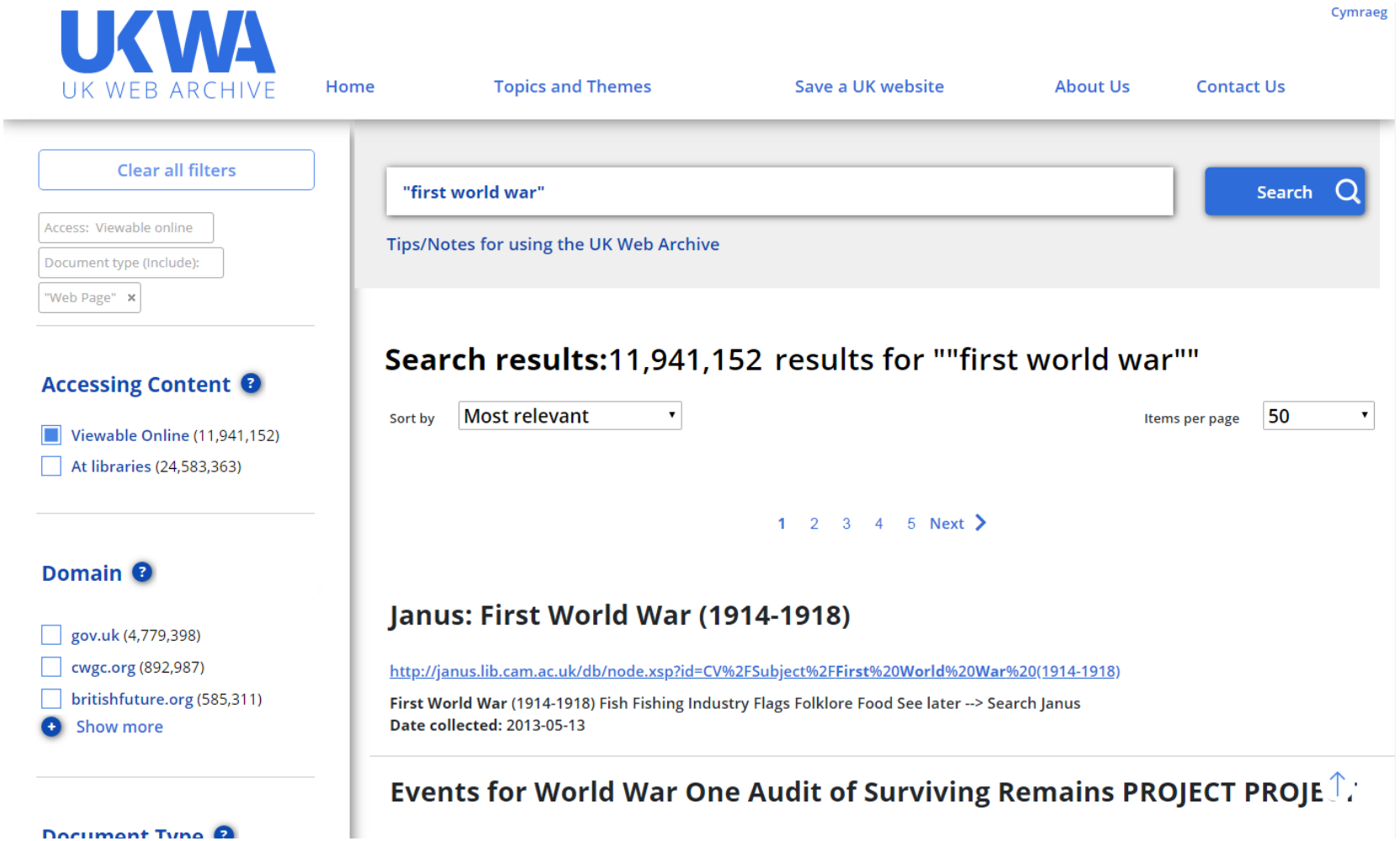
- Capture and Preserve the UK web space
- Support access to the collection
- Enable research



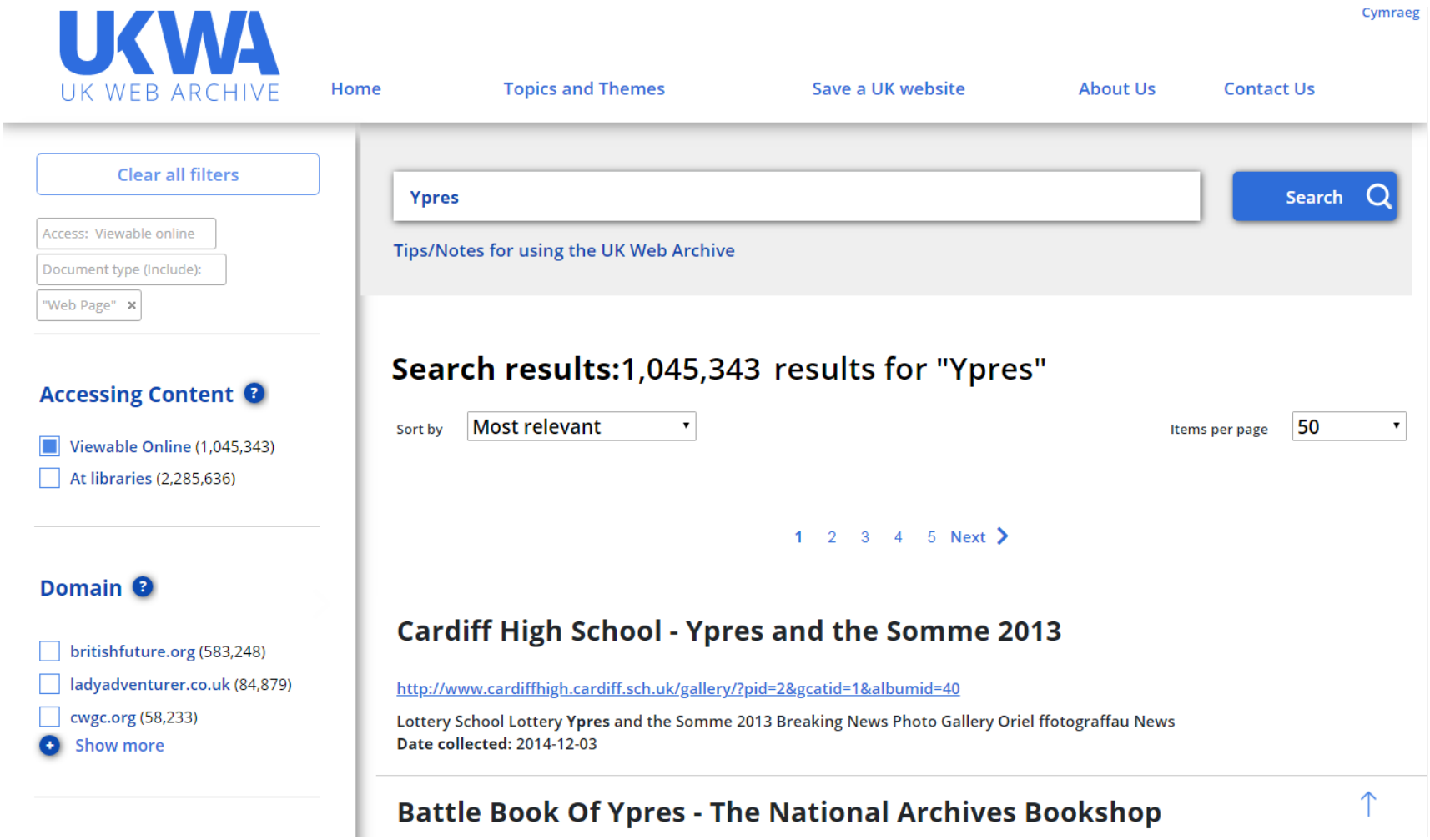
UKWA Collections

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

Access Paradox



The Size Problem



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.

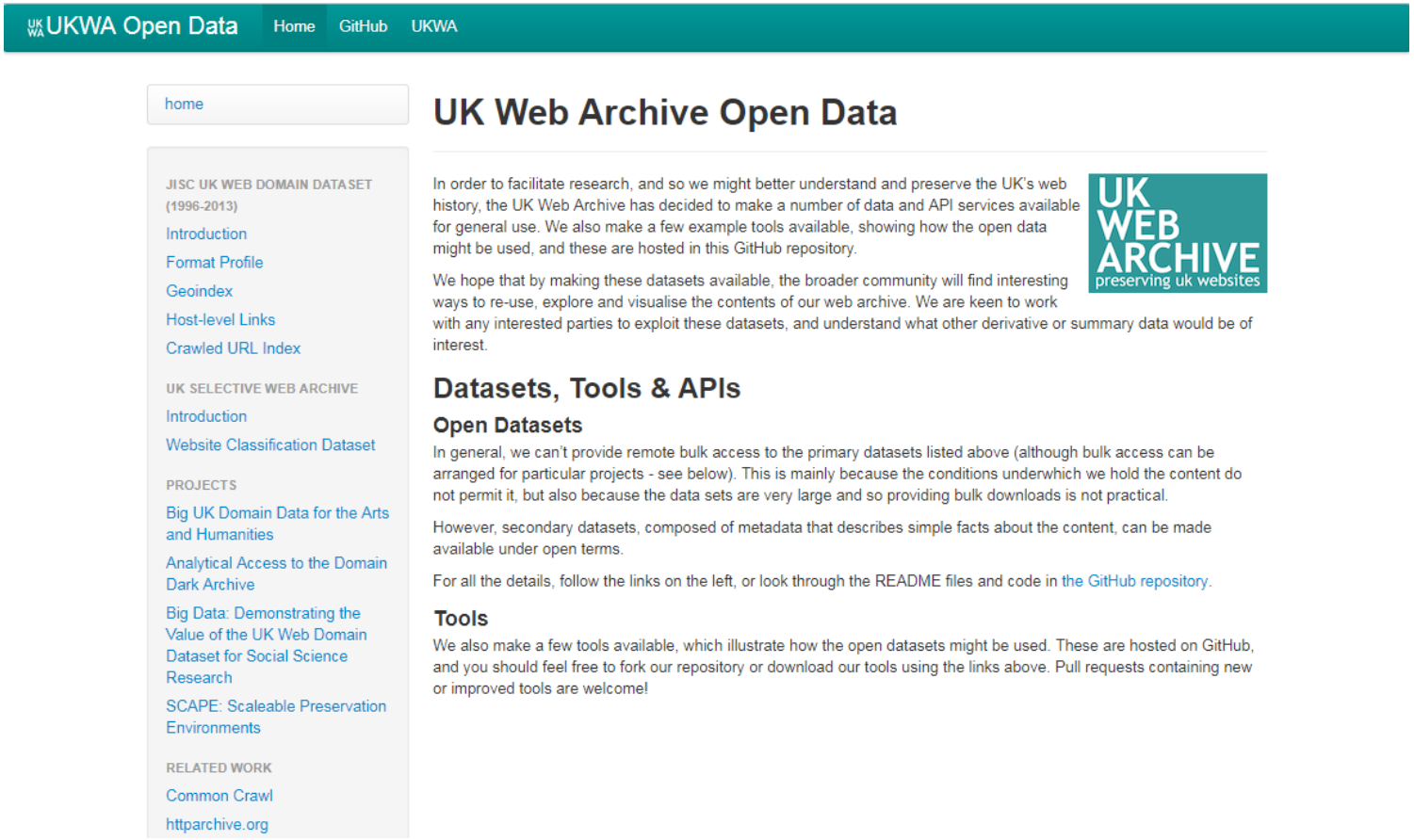
data.webarchive.org.uk

What Secondary data is available?

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

data.webarchive.org.uk

What Secondary data is available?



data.webarchive.org.uk

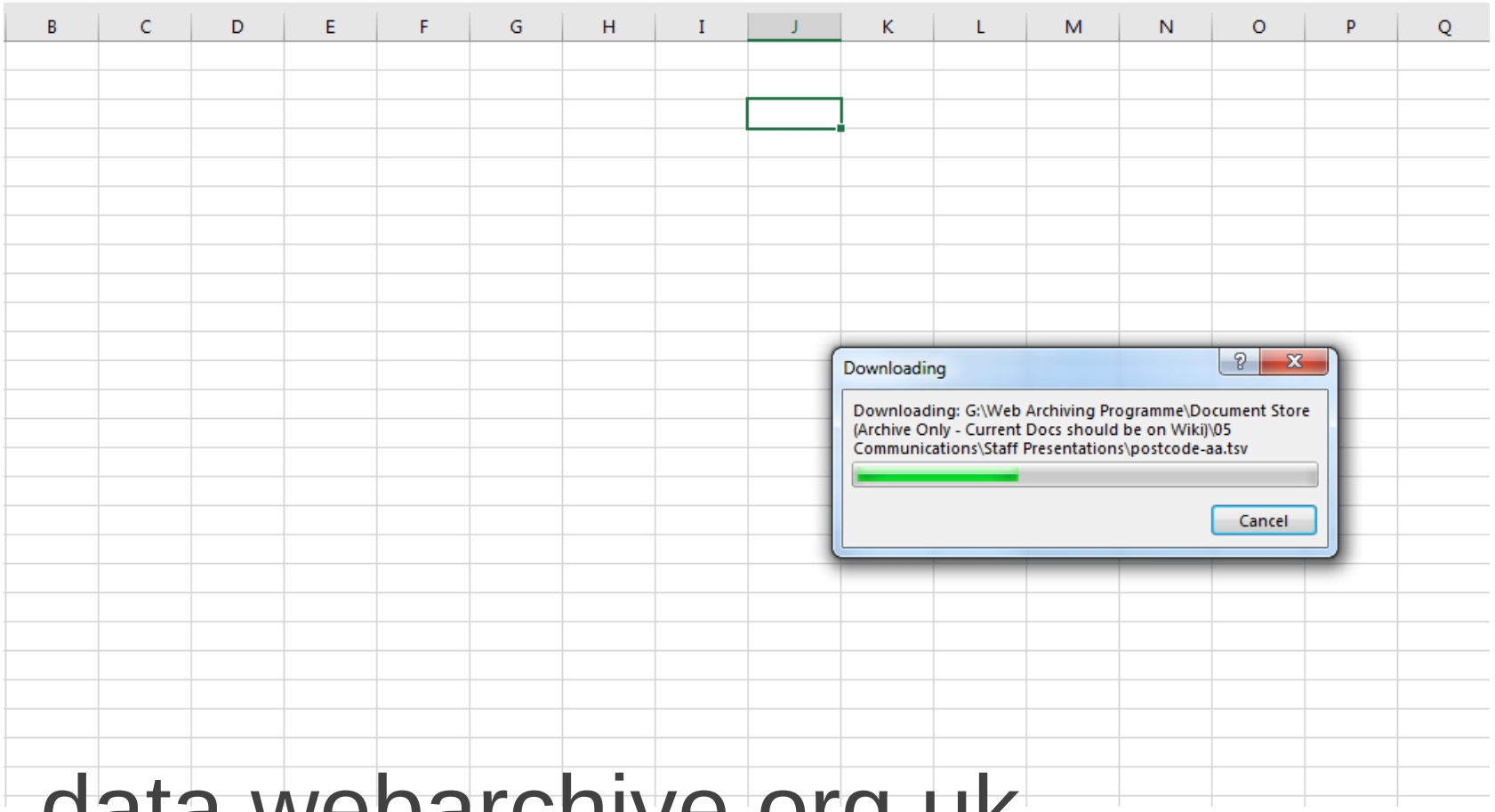
What Secondary data is available?

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_slimbri	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		

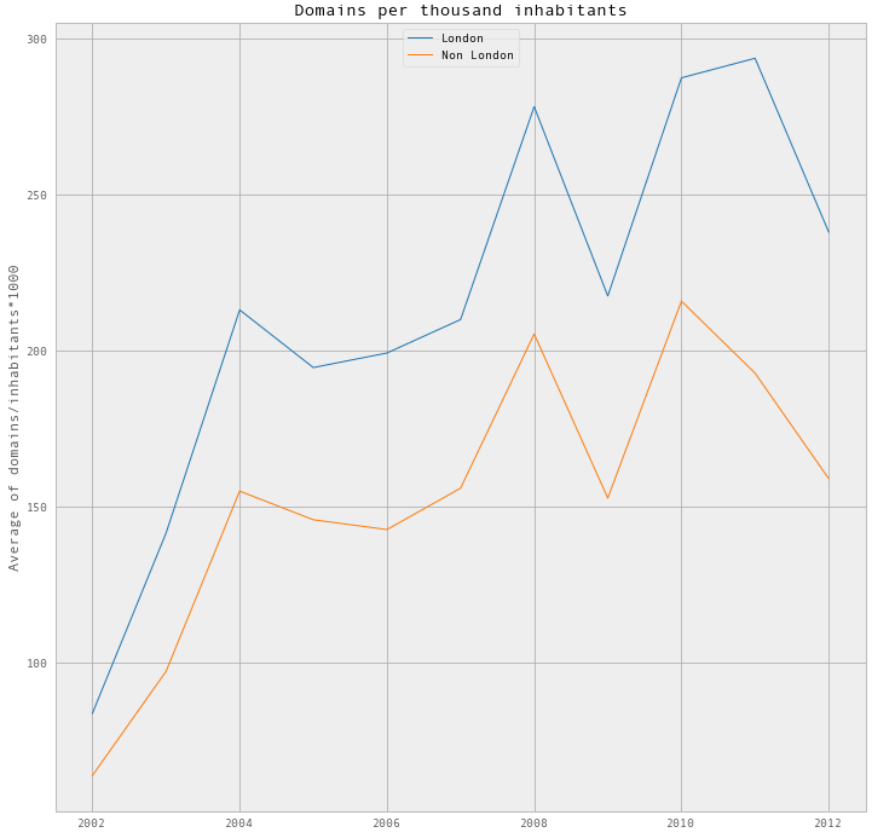
data.webarchive.org.uk

Case Study 1 – Geographical research

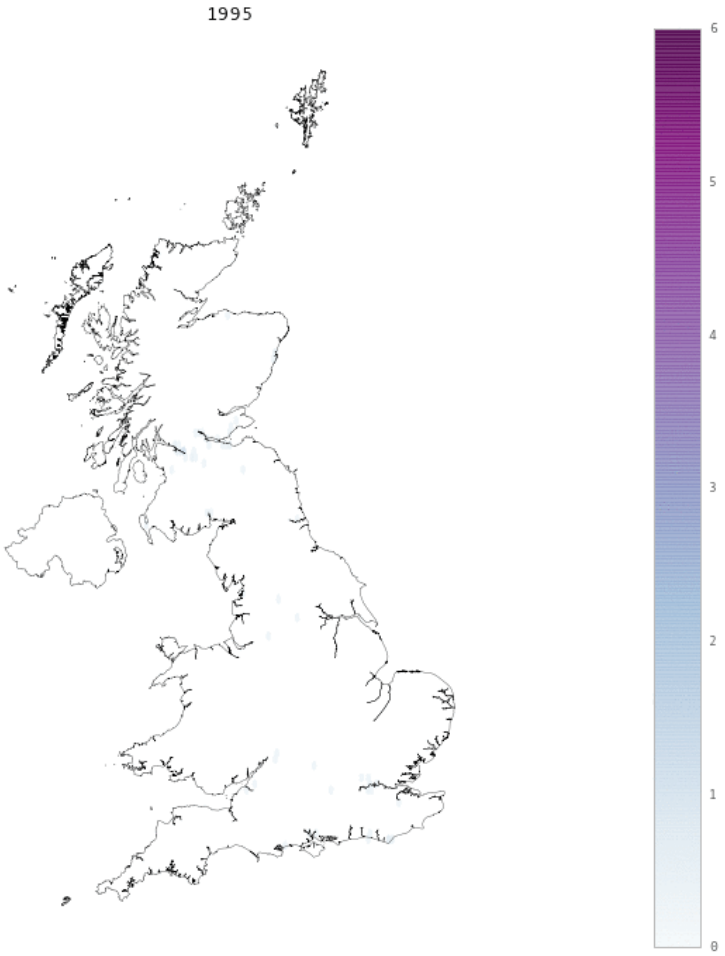
"If you are a quantitative social scientist there are few things more fascinating than free, under-utilised, quirky and easy to download data that also fits well the narrative of 'big data'."

Emmanouil Tranos, University of Birmingham and Alan Turing Fellow

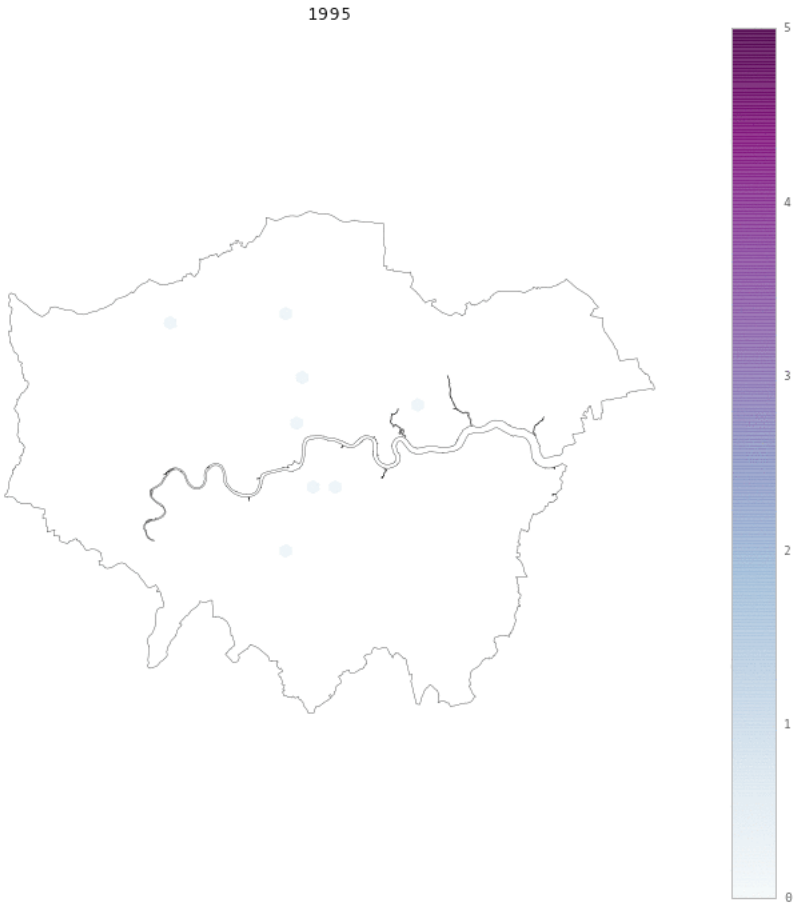
Case Study 2 – Geographical research



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