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

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

Sort by: Most relevant

Janus: First World War (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

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

Cardiff High School - Ypres and the Somme 2013
http://www.cardifffhigh.cardiff.sch.uk/gallery/?pid=2&gcatid=1&albumid=40
Lottery School Lottery Ypres and the Somme 2013 Breaking News Photo Gallery Oriel 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?

data.webarchive.org.uk
What Secondary data is available?

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

- east-london.tsv.gz
- manifest.sha1
- manifest.sha256
- postcode-aa.tsv.bz2
- postcode-ab.tsv.bz2
- postcode-ac.tsv.bz2
- postcode-ad.tsv.bz2
- postcode-ae.tsv.bz2
- postcode-af.tsv.bz2
- postcode-ag.tsv.bz2
- postcode-ah.tsv.bz2
- postcode-ai.tsv.bz2
- postcode-aj.tsv.bz2
- postcode-ak.tsv.bz2
- postcode-al.tsv.bz2
- summary-count-year-postcode.tsv.gz

data.webarchive.org.uk
Waiting....

data.webarchive.org.uk
What do you get?

<table>
<thead>
<tr>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007102152122/<a href="http://www.thebathweb.co.uk:80/mtree/Carpenters/">http://www.thebathweb.co.uk:80/mtree/Carpenters/</a> SE23 3EP</td>
<td></td>
</tr>
<tr>
<td>2007102152122/<a href="http://www.thebathweb.co.uk:80/mtree/Carpenters/">http://www.thebathweb.co.uk:80/mtree/Carpenters/</a> SE23 3EP</td>
<td></td>
</tr>
<tr>
<td>2007102152122/<a href="http://www.thebathweb.co.uk:80/mtree/Carpenters/">http://www.thebathweb.co.uk:80/mtree/Carpenters/</a> SE23 3EW</td>
<td></td>
</tr>
<tr>
<td>2007102152122/<a href="http://www.thebathweb.co.uk:80/mtree/Carpenters/">http://www.thebathweb.co.uk:80/mtree/Carpenters/</a> SE23 3EW</td>
<td></td>
</tr>
<tr>
<td>2007102152122/<a href="http://www.thebathweb.co.uk:80/mtree/Carpenters/">http://www.thebathweb.co.uk:80/mtree/Carpenters/</a> SE23 3HA</td>
<td></td>
</tr>
<tr>
<td>2007102152122/<a href="http://www.thebathweb.co.uk:80/mtree/Carpenters/">http://www.thebathweb.co.uk:80/mtree/Carpenters/</a> SE23 3HA</td>
<td></td>
</tr>
<tr>
<td>2007102152122/<a href="http://www.thebathweb.co.uk:80/mtree/Carpenters/">http://www.thebathweb.co.uk:80/mtree/Carpenters/</a> SE23 3HA</td>
<td></td>
</tr>
<tr>
<td>2007102152122/<a href="http://www.thebathweb.co.uk:80/mtree/Carpenters/">http://www.thebathweb.co.uk:80/mtree/Carpenters/</a> SE23 3HA</td>
<td></td>
</tr>
<tr>
<td>2007102152122/<a href="http://www.thebathweb.co.uk:80/mtree/Carpenters/">http://www.thebathweb.co.uk:80/mtree/Carpenters/</a> SE23 3HF</td>
<td></td>
</tr>
<tr>
<td>2007102152122/<a href="http://www.thebathweb.co.uk:80/mtree/Carpenters/">http://www.thebathweb.co.uk:80/mtree/Carpenters/</a> SE23 3HF</td>
<td></td>
</tr>
<tr>
<td>2007102152122/<a href="http://www.thebathweb.co.uk:80/mtree/Carpenters/">http://www.thebathweb.co.uk:80/mtree/Carpenters/</a> SE23 3HF</td>
<td></td>
</tr>
<tr>
<td>2007102152122/<a href="http://www.thebathweb.co.uk:80/mtree/Carpenters/">http://www.thebathweb.co.uk:80/mtree/Carpenters/</a> SE23 3HF</td>
<td></td>
</tr>
<tr>
<td>2007102152122/<a href="http://www.thebathweb.co.uk:80/mtree/Carpenters/">http://www.thebathweb.co.uk:80/mtree/Carpenters/</a> SE23 3HF</td>
<td></td>
</tr>
<tr>
<td>2007102152122/<a href="http://www.thebathweb.co.uk:80/mtree/Carpenters/">http://www.thebathweb.co.uk:80/mtree/Carpenters/</a> SE23 3HF</td>
<td></td>
</tr>
<tr>
<td>2007102152122/<a href="http://www.thebathweb.co.uk:80/mtree/Carpenters/">http://www.thebathweb.co.uk:80/mtree/Carpenters/</a> SE23 3HF</td>
<td></td>
</tr>
<tr>
<td>2007102152324/<a href="http://www.thebathweb.co.uk:80/mtree/Home/Carpet_and_Rugs/">http://www.thebathweb.co.uk:80/mtree/Home/Carpet_and_Rugs/</a> SE23 3HN</td>
<td></td>
</tr>
<tr>
<td>2007102152324/<a href="http://www.thebathweb.co.uk:80/mtree/Home/Carpet_and_Rugs/">http://www.thebathweb.co.uk:80/mtree/Home/Carpet_and_Rugs/</a> SE23 3HN</td>
<td></td>
</tr>
<tr>
<td>2007102152324/<a href="http://www.thebathweb.co.uk:80/mtree/Home/Carpet_and_Rugs/">http://www.thebathweb.co.uk:80/mtree/Home/Carpet_and_Rugs/</a> SE23 3HN</td>
<td></td>
</tr>
<tr>
<td>2007102152324/<a href="http://www.thebathweb.co.uk:80/mtree/Home/Carpet_and_Rugs/">http://www.thebathweb.co.uk:80/mtree/Home/Carpet_and_Rugs/</a> SE23 3HN</td>
<td></td>
</tr>
<tr>
<td>2007102152324/<a href="http://www.thebathweb.co.uk:80/mtree/Home/Carpet_and_Rugs/">http://www.thebathweb.co.uk:80/mtree/Home/Carpet_and_Rugs/</a> SE23 3HN</td>
<td></td>
</tr>
<tr>
<td>2007102152324/<a href="http://www.thebathweb.co.uk:80/mtree/Home/Carpet_and_Rugs/">http://www.thebathweb.co.uk:80/mtree/Home/Carpet_and_Rugs/</a> SE23 3HN</td>
<td></td>
</tr>
<tr>
<td>2007102152324/<a href="http://www.thebathweb.co.uk:80/mtree/Home/Carpet_and_Rugs/">http://www.thebathweb.co.uk:80/mtree/Home/Carpet_and_Rugs/</a> SE23 3HN</td>
<td></td>
</tr>
</tbody>
</table>
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!