

Interim statistics report

**Second Quarter 2019/2020**

# UK WEB ARCHIVE

Covering months: July, August, and September



# Table of contents

Introduction	01
Curation	02
continued	03
Scope	04
Open Access Licences	05
Usage (Open UKWA)	06
continued	07
Reading Rooms Statistics	08
Page views	09
Number of searches across LDLs	10
Number of searches and distinct searches	10
HDFS Storage	11
continued	12
Notes	12
continued	13
continued	14

# Introduction

This is the second Web Archiving Statistics Quarterly Report for 2019/2020.

It is our intention to distribute this report quarterly (July, October, January, and April) with a more comprehensive report at the end of the financial year.

The Hadoop Distributed File System (HDFS) statistics have also been included thanks to Andrew Jackson's new reporting tool that enables us to analyse the size of the UK Web Archive in more detail.

The format of the report is always in development so please do feedback comments to Nicola Bingham, Helena Byrne or Carlos Rarugal.

## First Report: July

April, May, June

## Second Report: October

July, August, September

## Third Report: January

October, November, December

## Fourth Report: April

January, February, March

Lead Web Curator:	Nicola.bingham@bl.uk
Web Curator:	Helena.byrne@bl.uk
Assistant Web Archivist:	Carlos.rarugal@bl.uk

# Curation


Below shows how many Targets (Titles) were created in ACT in the first quarter of the 2019/2020 reporting year, broken down by agency.

The ACT (Annotation Curation Tool) is the web curation software used by subject specialists across the UK Legal Deposit Libraries, as well as invited external partners, to curate websites and build special collections.

Within ACT, users create Target Records to highlight specific websites, adding basic metadata and setting the archiving frequency of individual websites.

A Target Record usually defines a “website” but can describe anything from a web page, to a sub section of a website, to several URLs grouped together. Archiving frequency depends on factors such as the rate of change of the website and its importance to a particular special collection.

## Total created per month

	July	August	September
 British Library	801	803	947
 National Library of Scotland	455	502	475
 National Library of Wales	145	372	207
 Bodleian Libraries Oxford University	25	32	33
 Cambridge University Library	0	1	0
 Trinity College Dublin	27	0	0

## Cumulative total: July 2019 to September 2019



2551



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

1432



724



Bodleian Libraries  
UNIVERSITY OF OXFORD

90

**600** YEARS  
1416-2016 | CAMBRIDGE  
UNIVERSITY  
LIBRARY

1



TRINITY  
COLLEGE  
DUBLIN

27

# Scope

Web archiving is carried out under the auspices of Legal Deposit Legislation and as such websites are only archived if they can be determined to be UK in scope. To do this, we run three automated checks:

- 1) Search for a .uk top level domain name
- 2) Run a geo-ip look up to determine the location of a server
- 3) Check against the WHO-IS registration database.

Where a website fails to meet any of these three criteria, additional, manual checks, such as locating a published postal address, are carried out by curators.

The table below shows the number of Targets falling into each category. The figures in this table are cumulative totals.

Targets that do not meet Legal Deposit (LD) criteria cannot be scoped in without an additional permission from the website publisher. They remain on the system as an indication of the content that the curator wanted to select and in case the status of the website can be verified by other means.

## Targets in ACT according to LD criteria

	July	August	September
UK Domain	<i>44,892</i>	<i>46,295</i>	<i>47,470</i>
UK GEO IP	<i>N/A</i>	<i>N/A</i>	<i>N/A</i>
UK Postal Address	<i>16,497</i>	<i>16,662</i>	<i>16,894</i>
Via correspondence	<i>2003</i>	<i>2013</i>	<i>2030</i>
Professional judgement	<i>18,763</i>	<i>19,313</i>	<i>20,083</i>
Targets in ACT that do not meet LD Criteria	<i>183</i>	<i>183</i>	<i>183</i>

UK GEO IP reporting tool is currently unavailable

# Open Access Licence

## Open Access Licences

**Licence Requests** - number of emails generated from ACT requesting permission for open access to archived websites.

**Licences Granted** - number of open access licences received. These figures are for all the LDLs combined.

## Open Access Licences

	July	August	September
Licence requests	172	360	353
Licences granted	51	48	35

## Open Access Licences



\*2019-2020 figure represents licence requests from April 2019 till present (September 30th)



# Usage

The “Open UK Web Archive” is the term given to [www.webarchive.org.uk](http://www.webarchive.org.uk), below are the monthly usage metrics.

Usage statistics are retrieved from Google Analytics, with the following metrics are used as an indication of user activity:

**Sessions** – a period of time a user is actively engaged with the website.

**Users** – each user who has initiated at least one session during the date range.

**Page Views** – the total number of pages viewed. Repeated views of a single page are counted.

**Pages/Session** - the average number of pages viewed in a session.

**New Sessions** – an estimate of the percentage of first-time visits.

## Open UK Web Archive usage

	July	August	September
Sessions	73,926	54,620	57,781
Users	66,871	46,337	49,075
Page views	115,341	96,976	99,739
Pages/Sessions	1.56	1.78	1.73
Average session duration	00:00:49	00:01:15	00:01:12
New users	63,023	42,774	45,649

## Cumulative Open UK Web Archive usage

	2016-2017	2017-2018	2018-2019	2019-2020
Sessions	<i>340,068</i>	<i>298,443</i>	<i>363,709</i>	<i>469,250</i>
Users	<i>292,699</i>	<i>257,058</i>	<i>307,341</i>	<i>400,909</i>
Page views	<i>1,070,160</i>	<i>960,913</i>	<i>854,800</i>	<i>768,094</i>
New users	<i>N/A</i>	<i>245,414</i>	<i>290,503</i>	<i>370,597</i>

Some values have been omitted due to the lack of data

# Reading Rooms






## Generation of Reading Rooms statistics:

When an archived webpage is viewed, the page URL is logged in a web server at the LDL and in the LDL's Wayback server. These logs are regularly transferred onto a centralised Hadoop cluster managed by the BL web archiving team. A MapReduce job is run on the numerous logs and a monthly report is then automatically created and emailed to certain Curators.

**Note 1** on usage: there is no way to separate staff and reader's usage in these reports.







**Note 2** Work is ongoing with the intent of improving the current generation of statistics, any errors that are found will be addressed and this should be reflected in future reports. Previous reports may have included different and/or incorrect figures due to inaccuracies during statistic generation, however these are errors that have been flagged in previous reports.

## User numbers

	July	August	September
 British Library	501	709	231
 National Library of Scotland	42	62	66
 National Library of Wales	12	10	11
 Bodleian Libraries Oxford University	12	23	28
 Cambridge University Library	14	13	28
 Trinity College Dublin	27	20	31

# Reading Rooms

## Recognised page views

	July	August	September
 British Library	1996	3642	684
 National Library of Scotland	345	355	209
 National Library of Wales	2	0	0
 Bodleian Libraries Oxford University	9	20	46
 Cambridge University Library	46	20	20
 Trinity College Dublin	182	0	16







Previously, “Page Views” were reported on, however, after improving upon the reporting of statistics, the figures will be stated as “Recognised page views”. The recognised page views are more accurate because the code that generated these figures has been refactored and errors removed. For example, we previously included a user’s request of resources as page views, which was incorrect as this could have inflated page view numbers.

These are the resources that are now no longer counted: '.css', '.jpg', '.gif', '.ico', '.png', '.js', '.swf', '.ttf', '.woff', '.jpeg', '.svg', '.json', '.mp3'.

# Reading Rooms







## Number of searches across LDLs

Number of search terms across LDL Reading Rooms

	July	August	September
 British Library	8	17	11
 National Library of Scotland	19	14	1
 National Library of Wales	0	0	0
 Bodleian Libraries Oxford University	2	0	0
 Cambridge University Library	1	0	0
 Trinity College Dublin	7	0	6

Some values have been omitted due to the lack of data

## Number of distinct searches across LDLs

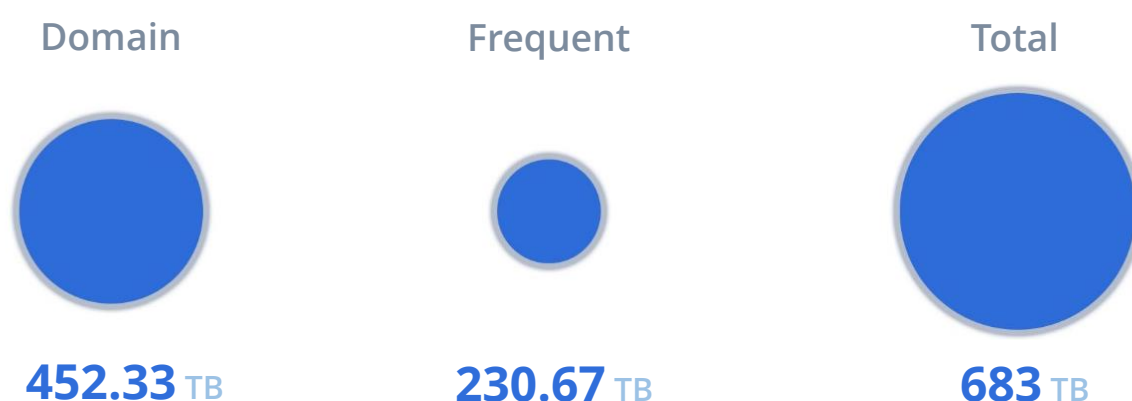
	July	August	September
 British Library	7	8	6
 National Library of Scotland	7	8	9
 National Library of Wales	0	0	0
 Bodleian Libraries Oxford University	2	0	0
 Cambridge University Library	1	0	0
 Trinity College Dublin	3	0	4

# HDFS Storage

The following statistics are generated based on the contents of the Hadoop Distributed File System (HDFS) that we use to store our data.

## Non-Print Legal Deposit totals

This section only includes archival content i.e. WARCs (either normal content or 'viral WARCs' containing material that appears to contain computer viruses), crawl logs and any additional archival package material.



## Yearly totals

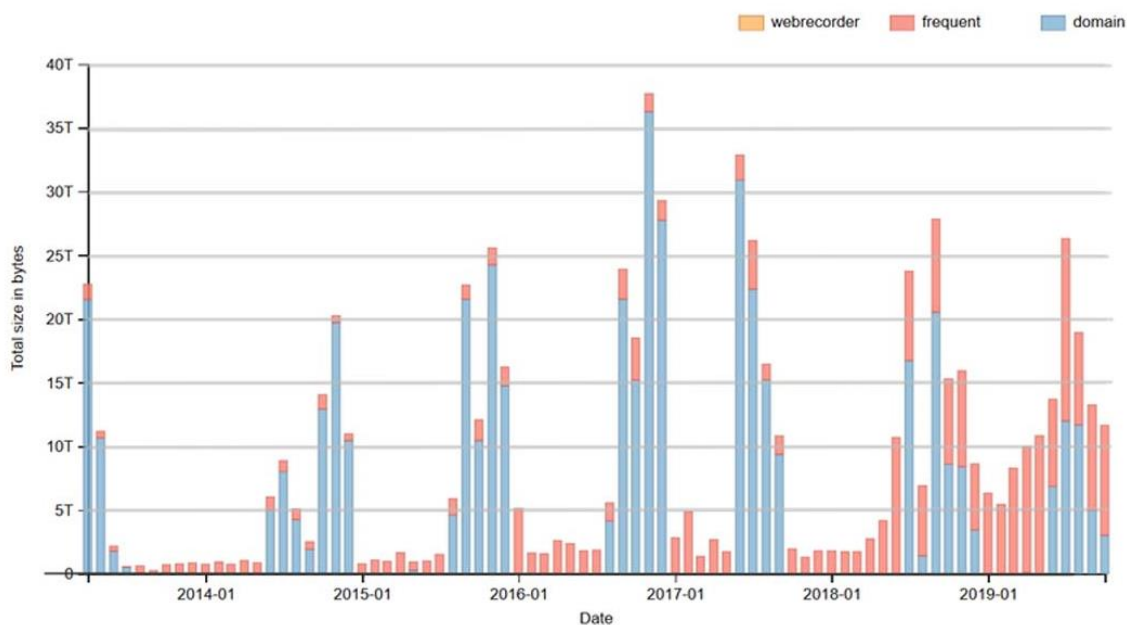
Year	Domain	Frequent	Webrecorder	Total
2013	34.31 TB	5.25 TB	-	39.56 TB
2014	62.14 TB	9.69 TB	-	71.83 TB
2015	75.73 TB	14.37 TB	-	90.1 TB
2016	104.83 TB	26.73 TB	491.40 MB	132.56 TB
2017	77.81 TB	26.71 TB	-	104.52 TB
2018	59.00 TB	61.93 TB	-	120.93 TB
2019	38.51* TB	85.99 TB	1.14 GB	124.5 TB

Hadoop File System (HDFS) statistics have been generated by the UKWA Reports tool; a new tool in development by the technical team to summarise HDFS statistics. \* The 2019 Domain crawl figure will increase as this is the size of data at the time of statistical generation (October 2019). Included are WARCs created by the Webrecorder tool, WARCs that are NPLD compliant.

# HDFS Storage

The following statistics are generated based on the contents of the HDFS file systems we use to store our data.

## Monthly breakdown of frequent and domain crawls



## Notes

### Domain crawls

The Domain crawl for 2019 is ongoing at the time of publishing (October 2019), so the size of the 2019 domain crawl will most likely increase.

### Frequent crawls

Frequent crawls occur non-stop all year round and can occur in parallel with the Domain crawl. The number of websites that have been selected to be crawled frequently are in the range of 90,000; these websites have different frequencies of crawling, from: daily, weekly, monthly, quarterly, six-monthly, and annually.

Webrecorder is a standalone tool used by curators to create high-fidelity web archives, the output of which is the standardised WARC format. These WARC files exist outside of the automated UKWA work pipeline, so are added manually and saved to the HDFS where they are then processed and indexed.

It has been noted for some time that the statistics being reported on LD UKWA usage may be significantly lower than actual user activity. After extensive investigation, Gil Hoggarth has been able to pinpoint the issues regarding the reporting anomalies.

The issue with statistics centres around usage within Legal Deposit Library(LDL) Reading Rooms, so statistics for Open UKWA ([webarchive.org.uk](http://webarchive.org.uk)) are not affected by the same errors.

The user activity is logged in the web servers and that relevant information is processed for useful metrics.

The user activity is recorded within the webserver, the webserver record the usage in logs. These logs can contain thousands and hundreds of thousands of log lines, below is a snippet of what a few log lines look like, where one line equals one complete log line.

For example, the useful information includes requested web resources (indicating page views) and attached to these are session IDs (indicates a unique user).



**How the stats have been improved**

The accuracy of statistics has been improved by debugging and refactoring the MapReduce python scripts that processes the logs; this means that the logs will be processed more accurately and therefore should produce fewer to no errors.

Additionally, recognised page views are now more accurate, resource requests were previously and incorrectly counted (these have now been discounted ['.css', '.jpg', '.gif', '.ico', '.png', '.js', '.swf', '.ttf', '.woff', '.jpeg', '.svg', '.json', '.mp3'])

Usage will be monitored but if any anomalies are spotted, then please do contact the relevant people.