

Interim statistics report

**First Quarter 2019/2020**

# UK WEB ARCHIVE

Covering months: April, May, and June 2019



# 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

# Introduction

This is the first 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 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

	April	May	June
 British Library	836	963	604
 National Library of Scotland	746	335	466
 National Library of Wales	430	300	162
 Bodleian Libraries Oxford University	6	18	13
 Cambridge University Library	2	0	112
 Trinity College Dublin	1	0	0

## Cumulative total: April 2019 to June 2019



# 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

	April	May	June
UK Domain	42,354	43,106	43,991
UK GEO IP	23,600	N/A	N/A
UK Postal Address	15,852	16,134	16,338
Via correspondence	1972	1991	1996
Professional judgement	17,328	18,013	18,387
Targets in ACT that do not meet LD Criteria	179	181	182

UK GEO IP reporting tool is currently broken but will be addressed in the next W3ACT update

# Open Access Licence

## Open Access Licences

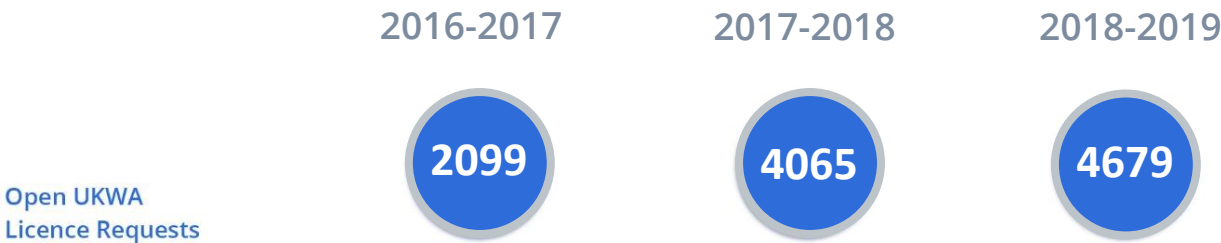
**Licence Requests** - number of emails generated from ACT requesting permission for open access to archived websites.

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

## Open Access Licences

	April	May	June
Licence requests	328	183	76
Licences granted	83	51	42

## Open Access Licences





# 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

	April	May	June
Sessions	84,227	66,374	65,948
Users	69,885	55,379	57,983
Page views	124,703	110,331	110,673
Pages/Sessions	1.48	1.66	1.68
Average session duration	00:01:04	00:01:11	00:00:59
New users	63,941	50,395	54,420

## Cumulative Open UK Web Archive usage

	2016-2017	2017-2018	2018-2019
Sessions	340,068	298,443	363,709
Users	292,699	257,058	307,341
Page views	1,070,160	960,913	854,800
New users	N/A	245,414	290,503

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** Previous statistics that have been reported have recently come under scrutiny due to perceived reporting errors; errors were proven and the current statistics from April 2019 reflect a revised and improved approach to log analysis.

## User numbers

		April	May	June
	British Library	459	659	723
	National Library of Scotland	40	16	87
	National Library of Wales	0	3	42
	Bodleian Libraries Oxford University	5	6	79
	Cambridge University Library	1	4	69
	Trinity College Dublin	0	0	57

# Reading Rooms







## Page views

	April	May	June
 British Library	1537	3789	2873
 National Library of Scotland	420	139	946
 National Library of Wales	0	7	120
 Bodleian Libraries Oxford University	364	52	447
 Cambridge University Library	1	14	109
 Trinity College Dublin	0	0	58

# Reading Rooms







## Number of searches across LDLs

Number of search terms across LDL Reading Rooms

	April	May	June
 British Library	150	43	26
 National Library of Scotland	1	1	62
 National Library of Wales	0	0	1
 Bodleian Libraries Oxford University	4	0	1
 Cambridge University Library	0	0	0
 Trinity College Dublin	5	0	1

Some values have been omitted due to the lack of data

## Number of distinct searches across LDLs

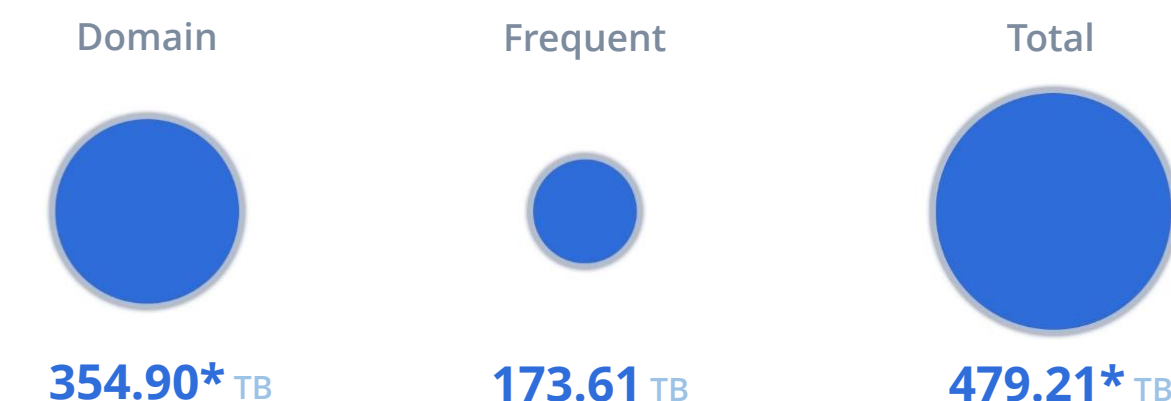
	April	May	June
 British Library	25	19	20
 National Library of Scotland	1	1	22
 National Library of Wales	0	0	1
 Bodleian Libraries Oxford University	2	0	1
 Cambridge University Library	0	0	0
 Trinity College Dublin	5	0	1

# 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

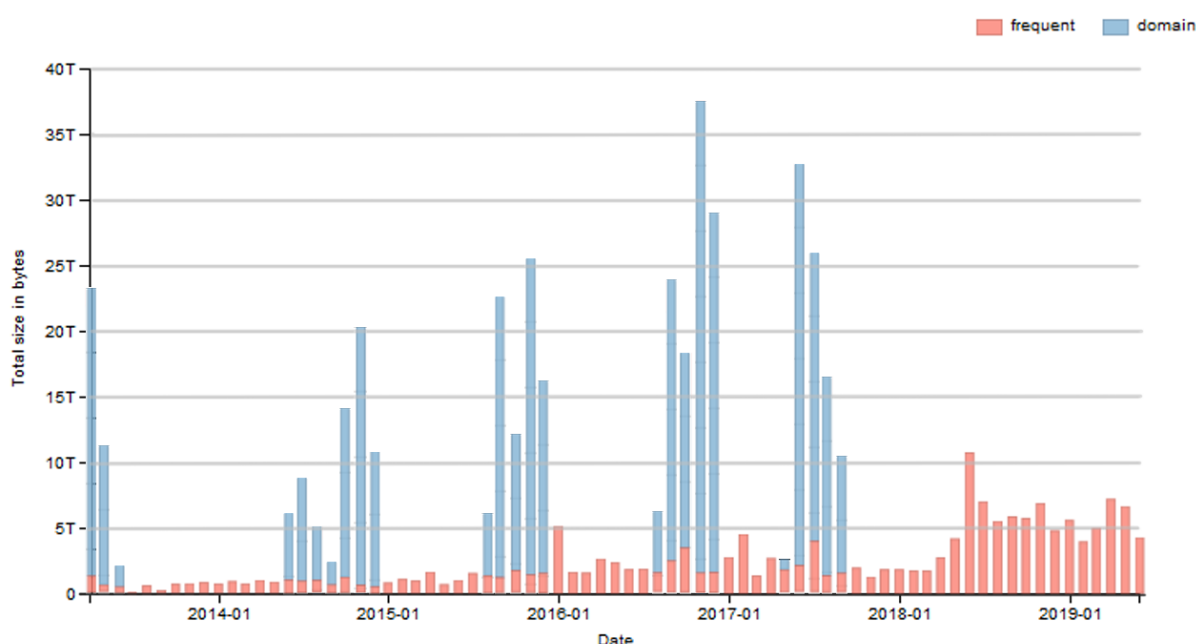
Year	Domain	Frequent	Total
2013	34.31 TB	5.25 TB	39.66 TB
2014	62.37 TB	9.75 TB	72.12 TB
2015	75.72 TB	14.45 TB	90.17 TB
2016	104.66 TB	26.96 TB	131.62 TB
2017	77.75 TB	26.35 TB	104.10 TB
2018	N/A	58.38 TB	N/A
2019	N/A	32.48 TB	N/A

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. \* Figure does not include 2018 Domain Crawl as it has yet to be fully processed.

# 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 2018 is not present in the graph above, this is due to the ongoing processing of said dataset. It has yet to be indexed and processed accordingly and therefore we are not yet able to give an exact figure to the size and the amount of data that was captured per month during the crawling period.

### Statistics

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.

Below is a overview of how our systems currently record user activity and where errors may have been introduced along the processing pipeline.

The issue with statistics centres around usage within Legal Deposit Library 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 the log lines look like.

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

The accuracy of statistics have been improved by debugging and refactoring the MapReduce scripts that processes the logs; this means that the logs will be processed more accurately and there should be fewer errors. This will be monitored and if anomalies are spotted, then please do contact the relevant people.