UK Web Archive Statistics

September 2017 Quarterly Report

Table of Contents

1) Introduction

This document replaces our monthly "Web Archiving Statistics" report. Following a review of the monthly report, and in consultation with stakeholders, the new document aims to convey a clearer, more narrative account of our web archiving statistics, reflecting key figures such as growth in our curated titles and usage of the collection.

It is our intention to distribute this report quarterly (June, September and December) with a more comprehensive report at the end of the financial year in March.

The format of the report is still in development; please feedback comments to Nicola Bingham or Helena Byrne.

2) Curation

Figure 1 A below shows how many Targets (Titles) were created in ACT in the past six months, 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.

*Figures 1-4 are from the September 2017 UKWA Statistics spreadsheet

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. Figure 1 B shows the cumulative totals for the number of targets created by each agency.

2.a) Targets created monthly

Figure 1 A

Selection	Apr-17	May-17	Jun-17	Jul-17	Aug-17	Sep-17
Targets (Titles) Created in ACT	**915**	**1818**	**554**	**894**	**525**	**671**
British Library	382	1086	287	317	267	199
National Library of Scotland	118	312	46	133	123	133
National Library of Wales	308	340	166	329	113	221
Bodleian Libraries, Oxford University	102	80	53	115	22	118
Cambridge University Library	0	0	0	0	0	0
Trinity College Dublin	5	0	2	0	0	0

Cumulative Target creation

Figure 1 B

Selection	2017-18 Total	2016-17 Total	2015-16 Total	2014-15 Total	2013-14 Total	Cumulative Total
Targets (Titles) Created in ACT	**5,377**	**11,905**	**14,887**	**10,728**	**9,332**	**52,229**
British Library	2,538	4,191	6,667	6,491	9,190	29,077
National Library of Scotland	865	3,046	2,135	1,638	114	7,798
National Library of Wales	1,477	3,417	2,804	891	21	8,610
Bodleian Libraries, Oxford University	490	1,218	3,169	1,708	7	6,592
Cambridge University Library	0	0	0	0	0	0
Trinity College Dublin	7	25	112	0	0	144

2.b) UK 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 and,
3) Check against the WHO-IS registration database.

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

Figure 2 shows the number of Targets falling into each category. The figures in this table are cumulative totals.

Targets that do not meet 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.

*Figures 1-4 are from the September 2017 UKWA Statistics spreadsheet

Figure 2

Targets in ACT according to LD Criteria	Apr-17	May-17	Jun-17	Jul-17	Aug-17	Sep-17
UK Domain	25940	26737	27007	27,552	27,884	28299
UK GeoIP	21269	21746	21927	22,210	22,405	22639
UK Postal Address	9408	9649	9815	9,955	10,089	10182
Via Correspondence	909	926	956	1,020	1,072	1077
Professional Judgement	9143	9890	9995	10,316	10,438	10615
Target records in ACT that do not meet LD UK criteria	197	197	197	196	196	196

3) Open Access Licences

Figure 3 A

Licences	Apr-17	May-17	Jun-17	Jul-17	Aug-17	Sep-17
Open UKWA						
Licence Requests						
General	540	126	377	317	249	341
Licences Granted	77	60	78	92	82	59

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.

Figure 3 B

Licences	2017-18 Total	2016-17 Total	2015-16 Total	2014-15 Total	2013-14 Total	Cumulative Total
Open UKWA						
Licence Requests	0	0	0	694	471	1,165
General	1,950	2,099	1,137	636	369	6,191
Licences Granted	448	583	497	311	224	2,063

4) Usage

4.a) Open UK Web Archive Usage

Figure 4 A

Open UKWA	Apr-17	May-17	Jun-17	Jul-17	Aug-17	Sep-17
Website Usage						
Sessions	25,212	24,727	19,857	21,493	22,110	21,715
Users	22,019	21,336	16,999	18,290	18,572	18,704
Page Views	81,341	76,292	69,251	74,977	77,152	75,741
Pages/Session	3.23	3.09	3.49	3.49	3.49	3.49
New Sessions	83.41%	80.93%	80.87%	81.25%	80.27%	82.22%

Figure 4 B

*Figures 1-4 are from the September 2017 UKWA Statistics spreadsheet

Open UKWA				
Website Usage	2017-18 Total	2016-17 Total	2014-15 Total	2013-14 Total
Sessions	135,114	340,068	388,037	563,177
Users	115,920	292,699	305,960	402,360
Page Views	454,754	1,070,160	1,090,402	1,533,300

The "Open UK Web Archive" www.webarchive.org.uk, contains permission-cleared websites. Figure 4 A gives the monthly statistics while figure 4 B is the annual total.

Usage statistics are retrieved from Google Analytics under the following criteria:

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.

4.b) Reading Rooms

Figure 4 C

LD UKWA	Apr-17	May-17	Jun-17	Jul-17	Aug-17	Sep-17
Users	**211**	**192**	**247**	**152**	**87**	**0**
British Library	151	158	219	139	85	0
National Library of Scotland	59	28	12	11	2	0
National Library of Wales	0	2	15	0	0	0
Bodleian Libraries, Oxford University	1	1	0	2	0	0
Cambridge University Library	0	3	1	0	0	0
Trinity College Dublin	0	0	0	0	0	0
Page Views	**1408**	**1439**	**1768**	**1122**	**848**	**0**
British Library	1166	1364	1641	1059	794	0
National Library of Scotland	235	68	48	39	54	0
National Library of Wales	0	2	78	0	0	0
Bodleian Libraries, Oxford University	7	2	0	24	0	0
Cambridge University Library	0	3	1	0	0	0
Trinity College Dublin	0	0	0	0	0	0
Searches (across LDLs)	50	17	25	10	14	47

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 transferred onto a Hadoop cluster managed by the BL web archiving team on a regular basis. A MapReduce job is run on Hadoop that gathers the data, summarises it and generates a report. However, there have been some issues with transferring the data from the logs which is why in previous reports there were

*Figures 1-4 are from the September 2017 UKWA Statistics spreadsheet

no reading room figures in the statistics and there may have been some inaccuracies. The technical team is working with an external contractor to develop a new version of Hadoop and to stabilize our usage reporting.

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

Note 2 on usage: *The way the Reading Room statistics are generated are currently under review and more details on the status of this review will be given in the annual report due in March 2018.*

Most visited URLs in Reading Rooms

The most visited URLs in reading rooms with the number of visits to that URL in brackets.

British Library

April: http://www.blackpool-unitarians.org.uk/home/home.html (23)
May: http://imgcdn.mediaplex.com/0/19476/universal.html?page_name=british_library_homepage&Homepage=1&mpuid= (31)
June: http://holybritain.co.uk/Britain_s_Holiest_Places.html (33)
July: http://www.findmypast.co.uk/includes/blank.html (32)
August: http://www.amdigital.co.uk/home (56)
September: N/A

National Library of Scotland

April: https://twitter.com/DeidreBrock (9)
May: http://www.scotcourts.gov.uk/ (4)
June: http://www.educationscotland.gov.uk/learningteachingandassessment/approaches/index.asp (10)
July: http://www.scotlawcom.gov.uk/ (4)
August: http://www.eqtr.com/twitterendum/thumbnail.html (4)
September: N/A

National Library of Wales

April: N/A
May: http://www.llgc.org.uk/cy/casgliadau/nbsp/gweithgareddau-cadwraeth/ (1)
June: http://edge.sharethis.com/share4x/index.c2670ee4b52a2b88a02bd11172df6393.html (6)
July: N/A
August: N/A
September: N/A

Bodleian Libraries, Oxford University

*Figures 1-4 are from the September 2017 UKWA Statistics spreadsheet

April: http://www.oxfordshirehealtharchives.nhs.uk/contact.htm (3)
May: http://www.heraldscotland.com/news/13211149.Centenary_of_the_1916_Easter_Rising_in_Ireland_to_be_commemorated_in_Scotland/ (2)
June: N/A
July: http://www.belfasttelegraph.co.uk/img/fonts/opensans-bold-webfont.woff (4)
August: N/A
September: N/A

Cambridge University Library

April: N/A
May: N/A
June: N/A
July: N/A
August: N/A
September: N/A

Trinity College Dublin

April: N/A
May:http://www.huffingtonpost.co.uk/june-sun/why-i-love-britain_b_2172995.html?utm_hp_ref=uk (1)
June: http://www.1916rising.com/ (1)
July: N/A
August: N/A
September: N/A

Searches (Across LDLs)

Number of search terms and most popular search terms across LDL Reading Rooms

April: Distinct search terms = 30
Most popular search term = Guatemala (5)

May: Distinct search terms = 13
Most popular search term = Wallace (4)

June: Distinct search terms = 15
Most popular search term = gilh+firefox (5)

July: Distinct search terms = 8
Most Popular search term = NHS (2)

August: Distinct search terms = 8
Most popular search term= baillieu (4)

*Figures 1-4 are from the September 2017 UKWA Statistics spreadsheet

September: Distinct search terms= 30

Most popular search term =

Calman+AND+Commission+AND+%22Scottish+Devolution%22 (4)

Kubrick (3)

bananas (1)

caerphilly (1)

5) Size of Collection

Update on Domain Crawling Activity:

The 2017 Domain Crawl started in May. At the time of writing, September 2017, we have crawled c. 70TB of data. A full report will be released at a future date once the crawl is completed and the results are analysed.

Previous Domain crawls

Figure 5

	2013	2014	2015	2016
Size TB	31	57		
Hosts million	4			
URLs billion	1.9	2.5		

*Figures 1-4 are from the September 2017 UKWA Statistics spreadsheet