

# Creating an archive of European Union law for Brexit

Tom Storrar, The National Archives  
Chris Doyle, MirrorWeb

# Introduction

T

When the UK leaves the European Union, the UK's European Communities Act (1972) will cease to have effect meaning that over 60 years of EU law will no longer apply.

However, in order to avoid legal chaos on exit, the European Union (Withdrawal) Act (2018) allows for continuity of EU derived law post-exit by retaining all applicable EU law in the UK statute book.

T

# Introduction

The National Archives (UK) is responsible for publishing legislation - we do this on [legislation.gov.uk](https://www.legislation.gov.uk)



[legislation.gov.uk](https://www.legislation.gov.uk)

The European equivalent of [legislation.gov.uk](https://www.legislation.gov.uk) is called **EUR-Lex**



**EUR-Lex**

[Access to European Union law](https://eur-lex.europa.eu/)

EUR-Lex is a complex and unique website but a website nonetheless. We decided to use our web archiving capability to **support our wider work and to built a new collection.**

# European Union (Withdrawal) Act, Sch. 5 para.1

- (1) The Queen's printer **must** make arrangements for the publication of—
  - (a) **each relevant instrument that has been published before exit day by an EU entity**, and
  - (b) the relevant international agreements.
- (1) In this paragraph—

**“relevant instrument”** means—

- (a) an EU regulation,
- (b) an EU decision, and
- (c) EU tertiary legislation; and

**“relevant international agreements”** means—

- (a) the Treaty on European Union,
- (b) the Treaty on the Functioning of the European Union,
- (c) the Euratom Treaty, and
- (d) the EEA agreement.

# European Union (Withdrawal) Act, Sch. 5 para.1

- (3) The Queen's printer **may** make arrangements for the publication of—
  - (a) any decision of, or expression of opinion by, the European Court, or
  - (b) any other document published by an EU entity.
- (3) The Queen's printer may make arrangements for the publication of anything which the Queen's printer **considers may be useful** in connection with anything published under this paragraph.
- (4) This paragraph does not require the publication of—
  - (a) anything repealed before exit day, or
  - (b) any modifications made on or after exit day.

**and...**

s.2(1) EU-derived domestic legislation, as it has effect in domestic law immediately before exit day, continues to have effect in domestic law on and after exit day.

s.3(1) Direct EU legislation, so far as operative immediately before exit day, forms part of domestic law on and after exit day.



EU Reg.

Current

Ex-EU Reg.



UK Amendment

EU Reg.

EU Amendment



Post Exit

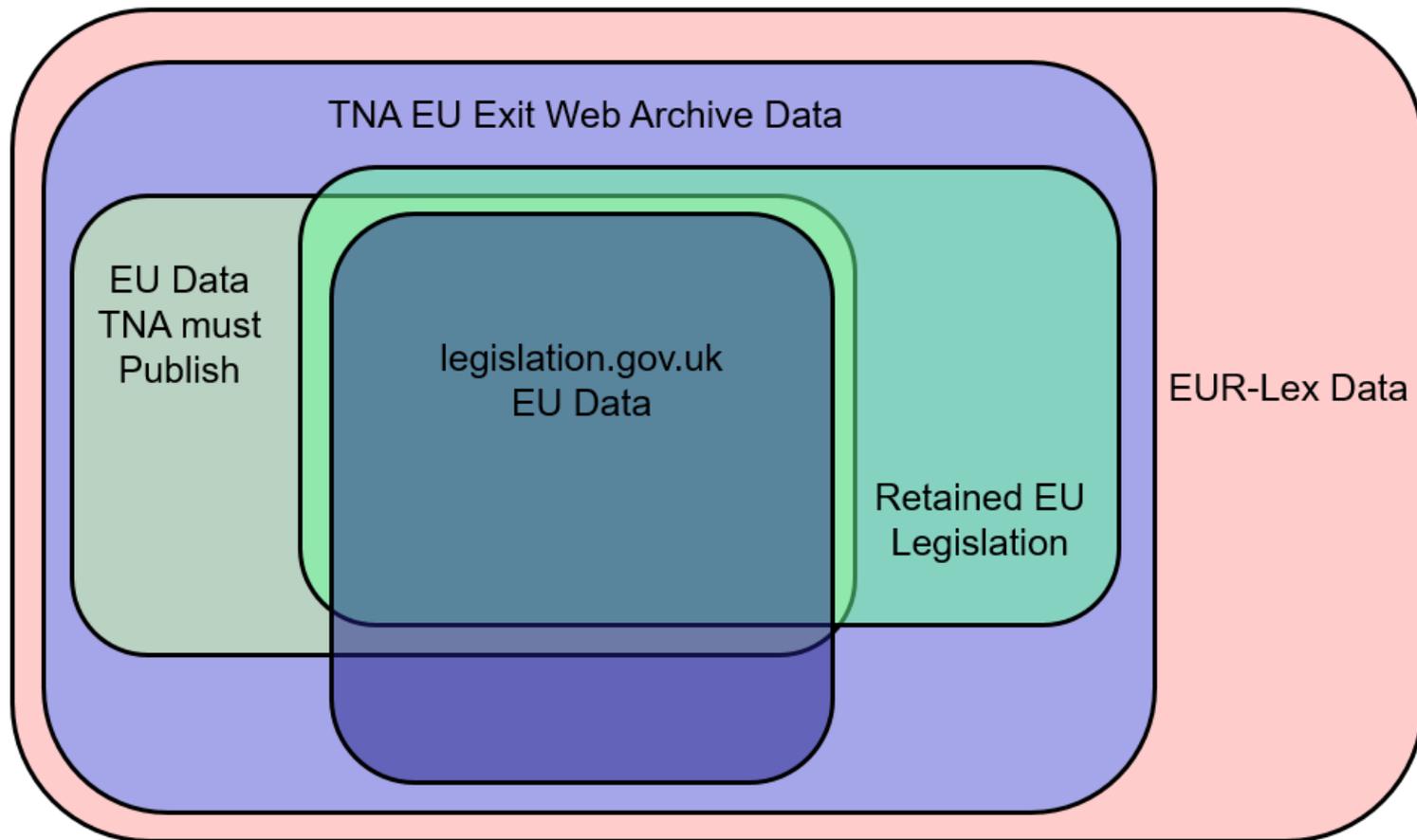
# The two services

In order to meet our duty, we have deployed a two-prong approach:

- 1) **legislation.gov.uk** - the home of UK legislation. This holds all UK law and will hold all EU-derived law. On this platform amendments can be applied post-exit.



- 2) **EU Exit Web Archive** - the comprehensive and official UK reference point for EU law as it stood at EU exit. Serves as:
  - A backstop: harvest more data than we have to harvest (e.g. Case Law)
  - A tool for supporting legal certainty: “showing our working” - provenance - contextualisation for legislation.gov.uk
  - A research service: what was the legal landscape at the time of exit?



# Project schedule

Capture scope is all HTML, PDF and XML formats in English, French and German.

Phase 1: two full captures of everything

Phase 1a: May to August 2018

Phase 1b: September to December 2018

Phase 2: incremental captures of newly published and modified content (ongoing)

Phase 3: continued incremental captures + archive access development and launch

Phases 2 & 3 run in parallel

# Building the archive #1 - Generating seeds 1

Our approach is data-driven which means that data we extract through querying the site's APIs leads to the construction of specific lists of seeds. We do no “conventional” crawling.

This has two key advantages:

- 1) It allows us to focus on the content we know we need
- 2) It allows high-quality QA of this content to be performed quickly

EUR-Lex presents its data using a number of **identifiers**, which can be combined with **URI patterns** to generate the seeds. The two central identifiers for the archive are:

- CELEX: most documents on EUR-Lex have one and are a basic unit of the data
- OJ: the identifier for the Official Journal of the European Union

# Building the archive #1 - Generating seeds 2

**CELEX** - e.g. CELEX:62000CC0049 <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:62000CC0049&from=EN>

20 patterns x 405,000 IDs = 8.1 million seeds

**OJ** - e.g. OJ:L:2016:170

<https://webarchive.nationalarchives.gov.uk/eu-exit/20190315005131/https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L:2016:170:FULL&from=EN>

33 patterns x 40,000 IDs = 1.4 million seeds

Meaning that each complete phase consists of **~10 million seeds**

# Building the archive #2 - Crawling

- I received seed lists via tickets on a Kanban board within JIRA
- Each list of CELEX and OJ had to be crawled against a defined set of patterns
- OJ
  - `eur-lex.europa.eu/legal-content/EN/AUTO/?uri=OJ:[SERIES:YEAR:NUMBER]:FULL`
  - `eur-lex.europa.eu/legal-content/EN/AUTO/?uri=OJ:[SERIES:YEAR:NUMBER]:TOC`
  - `eur-lex.europa.eu/legal-content/EN/ALL/?uri=OJ:[SERIES:YEAR:NUMBER]:FULL`

- `eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:[SERIES:YEAR:NUMBER]:FULL&from=EN ...`

TO DO 2

Jan - March 2019 - Incremental Crawling Phase Planning  
EULAW-137

Jan - March 2019 - OJEU - Incremental Crawling Phase Planning  
EULAW-148

+ Create

IN PROGRESS 4

Sep-Dec 2018 - Very Large Documents Not Displaying  
EULAW-130

BIGDOCS - Analyse and Recrawl?  
EULAW-234

May 2019 - CELEX - Incremental Crawl 6 (2019.04.04 to 2019.04.30)  
EULAW-245

May 2019 - OJEU - Incremental Crawl 6 (2019.04.04 to 2019.04.30)  
EULAW-251

# Building the archive #2 - Crawling

- Using a group of large servers within AWS running Heritrix
- Each pattern was broken up into groups of between 75K and 100K seeds
- Due to the relative frailties of the live site and the importance of not impeding its performance we crawled these
  - Very politely and slowly
  - With zero hops (meaning only the seed list would be captured)
- Once crawled, the same Heritrix server was used to create the CDX index records for each crawl

```
49 disposition.delayFactor=1
50 disposition.delayFactor=1
51 disposition.minDelayMs=500 <!--should be 500-->
52 disposition.maxDelayMs=1000 <!--should be 1000-->
53 queueAssignmentPolicy.parallelQueues=5
54 crawlController.maxToeThreads=5
55 metadata.operatorContactUrl=http://www.gov.uk
56 metadata.userAgentTemplate=Mozilla/5.0 (Windows NT 5.1) AppleWebKit/
57 warcWriter.prefix=MW-TNAEULAW-CELEX-CRAWL-AUG-P3-387
58 warcWriter.template=${prefix}-${timestamp14}-eur-lex.europa.eu-${ser
59     </value>
60
61     </property>
62
63 </bean>
64
65 <bean id="longerOverrides" class="org.springframework.beans.factory.
66
67     <property name="properties">
68
69         <props>
70
71             <prop key="seeds.textSource.value">
72 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=DD:1972I:FULL
73 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=DD:1972I:TOC
74 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:1997:340:F
75 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:1997:340:F
76 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2003:218:F
77 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2003:218:F
78 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2003:235:F
79 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2003:235:F
80 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2011:058:F
81 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2011:058:F
82
```

# Building the archive #2 - Crawling

- These WARCs and CDX were then synchronised up to AWS S3 buckets
- All S3 buckets have cross-region replication for resilience
- The same CDX were then resolved to remove revisit records
- The resolved CDX were then put through a Hadoop job using very large servers to create a zipnum index
- The zipnum was then synchronised to a test server in AWS EC2
- This server contains two collections - /qa/ and /2/

- The QA collection is emptied at the end of each round of incremental crawls
- 2 is a copy of the full collection to date
- Both are used to undertake AUTO and manual QA and crawler checks run by The National Archives team

```
sh1:LOCV7066ESWR5KCPNFNWHACW4XDIAC42 -- --
2019-05-04T21:02:29.569Z 302 377 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2013:194:TOC
sh1:H4RWHH6KR7EP6E02EOAUVNQK2ERICISIT -- --
2019-05-04T21:02:29.570Z 302 379 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2013:200:FULL
sh1:LSD7SFSRWZFHZGGOH3Z3QJHMCA2OF2SN -- --
2019-05-04T21:02:29.573Z 302 377 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2013:192:TOC
sh1:WIR3Q56RNMGN67TYDCOVSSJXM37E -- --
2019-05-04T21:02:29.576Z 302 379 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2013:197A:TOC
sh1:GNWRFGERRHX6VJIKJLVRIY7XP6KNV2 -- --
2019-05-04T21:02:29.620Z 302 377 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2013:199:TOC
sh1:G4EPBY5RXZROKR7SL57ZNSHWAUQ7GDI -- --
2019-05-04T21:02:30.654Z 302 379 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2013:200A:TOC
sh1:GAWT4USB6EDOVBMJDJE2B7R5JWHB5M4G -- --
2019-05-04T21:02:30.654Z 302 379 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2013:193:FULL
sh1:RFSMTIDFVWVPCJL7DEPEPIDNSMPTN4H -- --
2019-05-04T21:02:30.656Z 302 381 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2013:195A:FUL
sh1:BUMEIXRBXUGFLDUM3UFRIH2EJFYXSDO -- --
2019-05-04T21:02:30.657Z 302 379 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2013:198:FULL
sh1:34CMQVVR0BBRWEE6GTIJH5TWVU7GM -- --
2019-05-04T21:02:30.704Z 302 379 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2013:202:FULL
sh1:P4SMYD2EOKW6TNRHUMRVCPXKI4LIDF -- --
2019-05-04T21:02:31.736Z 302 381 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2013:193A:FUL
sh1:L3Z4PWNKYG6IQVJSMWY6MAUZ44PALSI -- --
2019-05-04T21:02:31.742Z 302 377 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2013:202:TOC
sh1:CK7WTSU642LBLEWJDEZ43CFLNS7VM7ML -- --
2019-05-04T21:02:31.744Z 302 379 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2013:195A:TOC
```

# Building the archive #3 - AUTO QA

- This system was developed by Liam Freeman, one of MirrorWeb's finest developers
- After a lot of trial and error - the focus of which was to give 100% confidence that everything we had captured was in scope and was presented **exactly** as it is on the original site, Liam developed this:
  - Construct the URLs - Each CELEX or OJ Number was split into the different URLs for each language and each format type
  - Push to Queue - These URLs were pushed to a queue for processing
  - Hit the Link - A request was made to the link
  - Handle the Response - The response was passed through a process to figure out which of the many scenarios it represented
  - Push to a File - Every result was logged and categorised for assessment

# Building the archive #3 - AUTO QA

- The results for these checks are then placed on a ticket on the Kanban board and are recrawled by me.
- The recrawls generally consist of
  - Page crawls that received the message 404: This request could not be processed
  - Pages that received a 200 code but contain the text: **This document cannot be displayed due to its size** - “big docs” - along with a link to download the document. As we crawl with zero hops, these documents are not captured and must be archived individually.
  - Occasional 5xx errors
- Armed with the list provided, I wrote a little python script that would loop through the urls and output a list of download urls
- These were crawled very slowly and with huge timeout margins.
- Post “big docs” crawl, I would compare the archive size from the log to the live size - as a sanity check

```
import requests
import re
f = open('list.txt', 'a+') #creates a text file called list.txt that is appendable.
for line in open('C14b - from auto qa - 404sCD.txt'): # reads it line by line

    r = requests.get(line) #opens each line
    html = r.text #gets all the page content

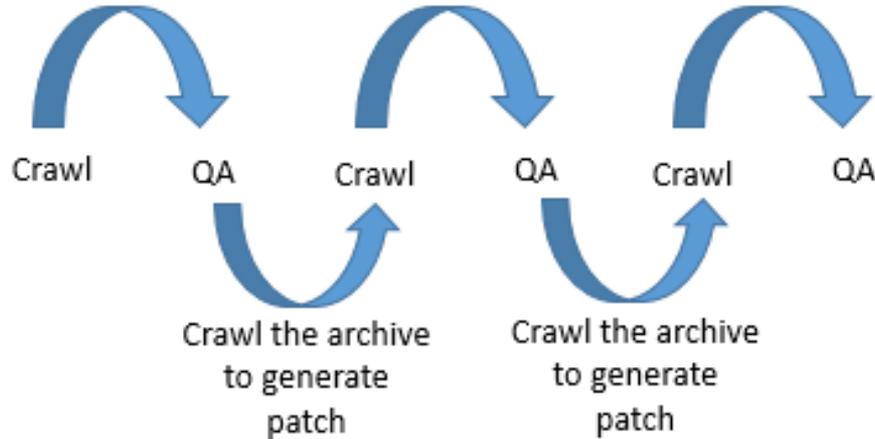
    # use re.findall to get all the links - basic regex used here

    links = re.findall('"(.*/publications.europa.eu/resource/cellar.*)"', html)
```

# Building the archive #4 - Further QA & Discovery

One of the cons of data-driven archiving is how restricted it is. If something is not explicitly in a list it will not be captured. This applies to css, js, image and font files but also to other document content that we cannot reliably predict up front to bring it in to the data-driven process.

Therefore, we have developed our QA process to find and capture these missing files;



# Developing Access

As with our UK Government Web Archive, this archive will be made public.

Data-driven archiving leads naturally to data-driven access, in which the same restrictions applied at capture time make themselves felt in access.

This manifests itself in three main ways:

- 1) A need to customise access rules
- 2) A need to customise presentation, e.g. to deactivate non-functioning links
- 3) A strong preference for search as the default access route

# Developing Access - Front End Customisations

## Standard PYWB playback

The screenshot shows the standard EUR-Lex interface. At the top left is the EUR-Lex logo with the tagline 'Access to European Union law'. Below it is a blue 'MENU' button and a search bar labeled 'QUICK SEARCH'. A dark blue breadcrumb trail reads 'EUROPA > EUR-Lex home > EUR-Lex - 32016R0679 - EN'. The main content area is titled 'Document 32016R0679'. On the left, there is a sidebar with sections: 'Text' (highlighted), 'Document information', 'Procedure', and 'Summary of legislation'. Below these are links: 'Save to My items', 'Permanent link', 'Download notice', and 'Follow this document'. At the bottom left is a 'Table of contents' button. The main content area has a 'Title and reference' section with the text: 'Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) (Text with EEA relevance)'. It includes the reference 'OJ L 119, 4.5.2016, p. 1–88 (BG, ES, CS, DA, DE, ET, EL, EN, FR...)' and a green dot indicating it is 'In force'. Below this is the ELI link: 'http://data.europa.eu/eli/reg/2016/679/oj'. A 'Languages, formats and link to OJ' section shows a grid of language icons (BG, ES, CS, DA, DE, ET, EL, EN, FR) and format icons (HTML, PDF, Official Journal).

## With EEWA modifications

The screenshot shows the EUR-Lex interface with EEWA modifications. The layout is similar to the standard version, but with several changes. The 'MENU' button is replaced by a blue button with a red 'X' over it and the text 'NU'. The breadcrumb trail is the same. The main content area is titled 'Document 32016R0679'. The sidebar is present but the 'Table of contents' button is missing. The 'Title and reference' section is identical to the standard version. However, the 'In force' indicator is missing. The 'Languages, formats and link to OJ' section shows the language and format icons, but a large red 'X' is overlaid on the PDF icon, indicating that the PDF download option is disabled. The 'Summary of legislation' section contains links: 'Save to My items' (disabled with a red 'X'), 'Permanent link', 'Download notice', and 'Follow this document' (disabled with a red 'X'). A 'Table of contents' button is located at the bottom left of the main content area.

# Developing Access - Search

- We developed a solution in house to extract all the metadata from each document
- This was vital due to inconsistent and incomplete xml formats used on the live site
- This was all then indexed and stored in an Elasticsearch cluster
- Development of the front end started with a 4 day sprint between front-end dev and a Python dev with the intent of meeting in the middle

## Search the EU Exit Web Archive

**Title (full or partial):**  ?

**Include keyword(s):**  ?

**Exclude keyword(s):**  ?

**Date of document:** **From**   **To**

**Celex number:**  ?

## Search results for "fishing"

1-10 of **4,403** results

New Search

**Filter**

**Title (full or partial):**

**Include keyword(s):**

**Exclude keyword(s):**

1 2 3 4 5 ... 440 441 »

Results per page: 10 ▾

<b>Title:</b> Case C-372/08 P: Appeal brought on 14 August 2008 by Atlantic Dawn Ltd, Antarctic Fishing Co. Ltd, Atlantean Ltd, Killybegs Fishing Enterprises Ltd, Doyle Fishing Co. Ltd, Western Seaboard Fishing Co. Ltd, O'Flaherty Fishing Co. Ltd, Ainslie Fishing Co. Ltd, Broadbent	<b>English</b>	<b>Français</b>	<b>Deutsch</b>
	HTML	HTML	HTML
	PDF	PDF	PDF

# Developing Access - Search

- A huge number of iterations following extensive UAT and input from UX experts followed
- Before

EU Exit Web Archive - Search

Search the EU Exit Web Archive

Title (full or partial):  
e.g. International

Include keyword(s):  
e.g. Trade

Exclude keyword(s):  
e.g. Finance

Date of document:  
From: \_\_\_\_\_ To: \_\_\_\_\_  
e.g. 1/1/2019 e.g. 31/12/2019

Context number:  
e.g. 1000000000

Type:  
 All  
 DE  
One or more of the following types:  
 Legislation  
 Directives  
 Regulations  
 Decisions  
 Complementary legislation  
 Treaties  
 International agreements  
 Case law  
 Consultation acts  
 Other C series documents  
 EFTA documents

Language:  
 English  
 French  
 Deutsch

Documents in force on exit:  
 All  
 In force  
 Not in force

Search

The National Archives  
Kew, Richmond, Surrey  
TW9 4DU  
Tel: +44 (0)20 896 3000

MirrorWeb Web and Social Media Archiving Service  
powered by Internet2 Ltd

After

The National Archives

EU Exit Web Archive - Search

Search the EU Exit Web Archive

Title (full or partial):  
e.g. Services Data Protection Regulation or GDPR

Include keyword(s):  
e.g. Brexit

Exclude keyword(s):  
e.g. Corporate

Date of document:  
From: \_\_\_\_\_ To: \_\_\_\_\_  
e.g. 1/1/2019 e.g. 31/12/2019

Context number:  
e.g. 1000000000

Type:  
 Any  
 DE  
One or more of the following types:  
 Legislation  
 Directives  
 Regulations  
 Decisions  
 Complementary legislation  
 Treaties  
 International agreements  
 Case law  
 Consultation acts  
 Other C series documents  
 EFTA documents

Language:  
 English  
 French  
 Deutsch

Documents in force on exit:  
 Any  
 In force  
 Not in force

Search

The National Archives  
Kew, Richmond, Surrey  
TW9 4DU  
Tel: +44 (0)20 896 3000

MirrorWeb Web and Social Media Archiving Service  
powered by Internet2 Ltd

# Conclusion

- The archiving approach we have taken has harvested vast numbers of specific URIs efficiently.
- The decisions we made then influence the quality of access we can provide and in the context of aiding legal certainty additional modifications needed to be made.
- Further exploration made through user research (April/May 2019)

In the near future we will launch the archive - check out [nationalarchives.gov.uk/webarchive/](https://nationalarchives.gov.uk/webarchive/) in the coming months...

Thank you!

[webarchive@nationalarchives.gov.uk](mailto:webarchive@nationalarchives.gov.uk)