Creating an archive of European Union law for Brexit

Tom Storrar, The National Archives
Chris Doyle, MirrorWeb

The National Archives

MirrorWeb

Introduction

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.

The National Archives

MirrorWeb

Introduction

The National Archives (UK) is responsible for publishing legislation - we do this on **legislation.gov.uk**

The European equivalent of legislation.gov.uk is called **EUR-Lex**

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**.

The National Archives

T

MirrorWeb

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—

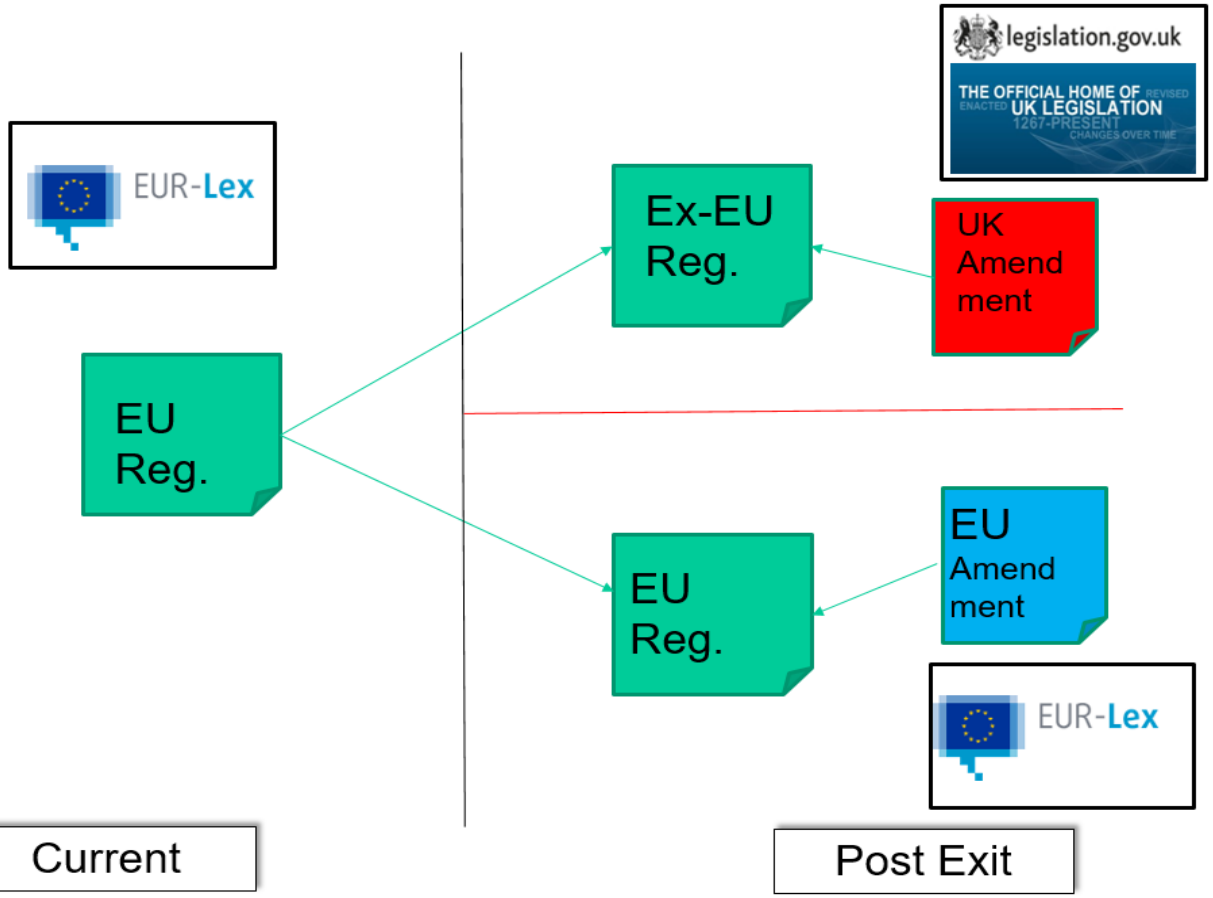
(a) anything repealed before exit day, or

(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.



The National Archives

MirrorWeb

The two services

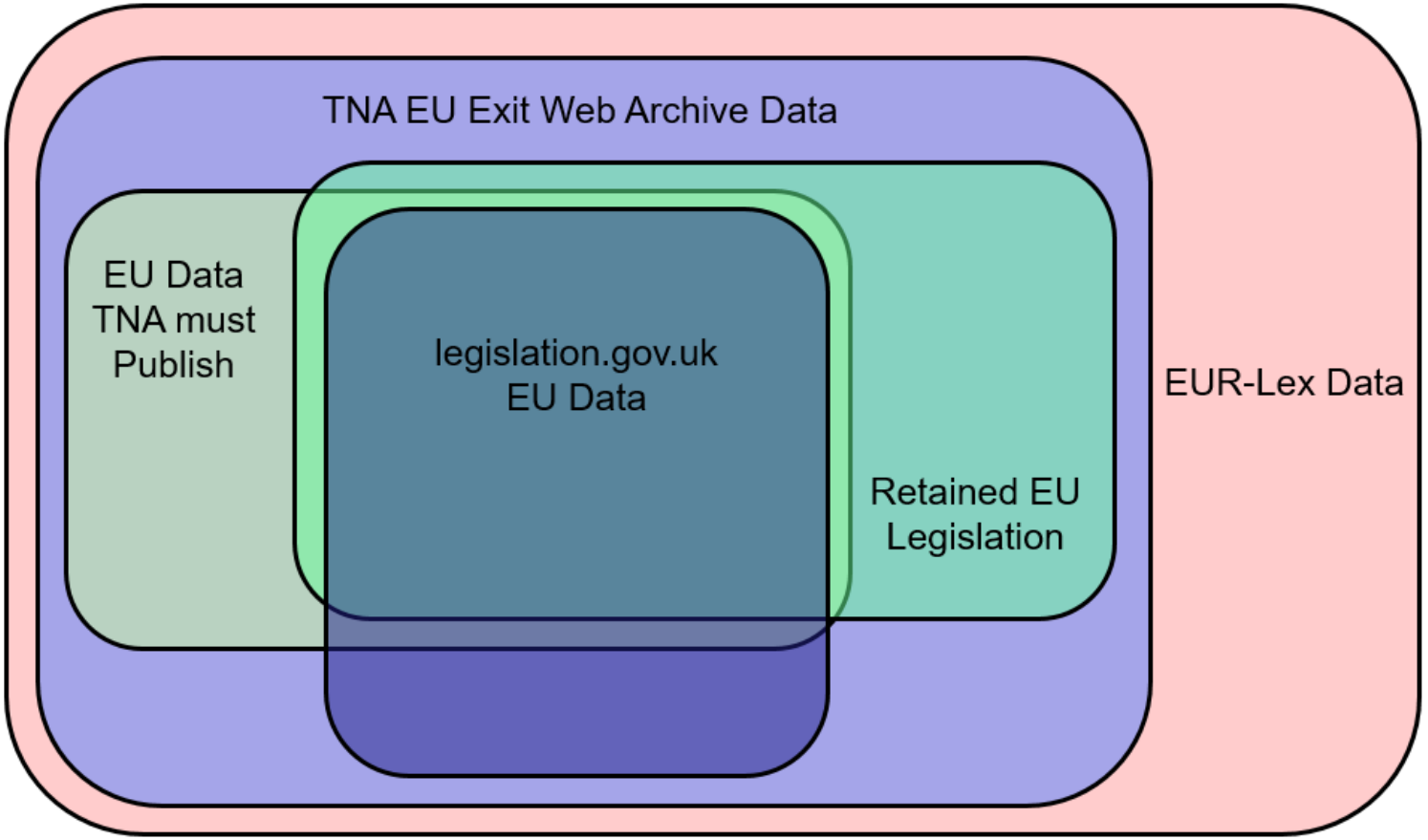
In order to meet our duty, we have deployed a two-prong approach:

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?

The National Archives

MirrorWeb



The National Archives

MirrorWeb

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

The National Archives

T

MirrorWeb

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

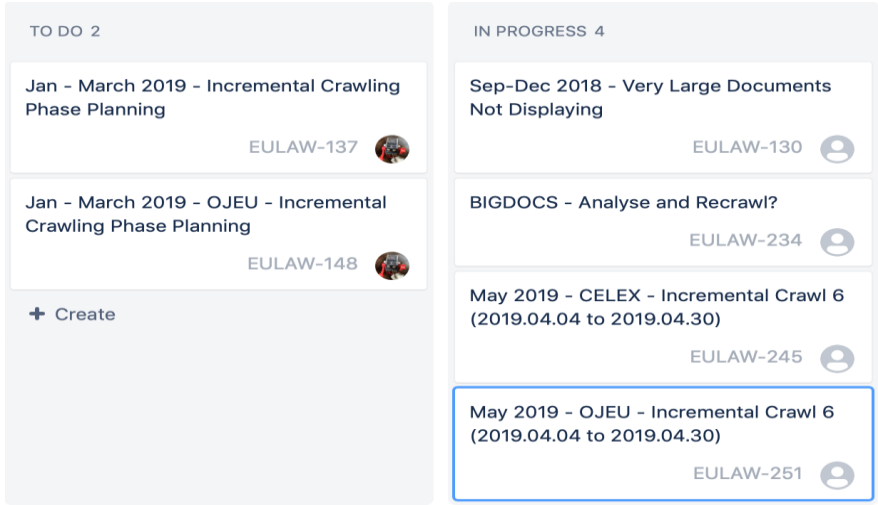
Meaning that each complete phase consists of **~10 million seeds**

The National Archives

MirrorWeb

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 ...



The National Archives

c

MirrorWeb

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

```
disposition.delayFactor=1
disposition.delayFactor=1
disposition.minDelayMs=500 <!--shold be 500-->
disposition.maxDelayMs=1000 <!--should be 1000-->
queueAssignmentPolicy.parallelQueues=5
crawlController.maxToeThreads=5
metadata.operatorContactUrl=http://www.gov.uk
metadata.userAgentTemplate=Mozilla/5.0 (Windows NT 5.1) AppleWebKit/
warcWriter.prefix=MW-TNAEULAW-CELEX-CRAWL-AUG-P3-387
warcWriter.template=${prefix}-${timestamp14}-eur-lex.europa.eu-${ser
        </value>

    </property>

</bean>

<bean id="longerOverrides" class="org.springframework.beans.factory.

    <property name="properties">

        <props>

            <prop key="seeds.textSource.value">
https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=DD:1972I:FULL
https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=DD:1972I:TOC
https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:1997:340:F
https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:1997:340:T
https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2003:218:F
https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2003:218:T
https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2003:235:F
https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2003:235:T
https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2011:058:F
https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2011:058:T
```

The National Archives

MirrorWeb

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

```
sha1:LOCV7Q66ESWR5XCPNFNWHACW4XDIAC42 - -
2019-05-04T21:02:29.569Z   302       377 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2013:194:TOC
sha1:H4RWHH6KR7EP6EO2EOAUVNQK2ERICSIT - -
2019-05-04T21:02:29.570Z   302       379 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2013:200:FULL
sha1:LSD7SFSRWZFHZGOH3Z3QJHMNCA2OF2SN - -
2019-05-04T21:02:29.573Z   302       377 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2013:192:TOC
sha1:WIR3Q56PNMNGXNN67YYDCOVTSSJXM37E - -
2019-05-04T21:02:29.576Z   302       379 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2013:197A:TOC
sha1:GNWRFGERRHX6VJIKJLVRIIY7XP6NKNV2 - -
2019-05-04T21:02:29.620Z   302       377 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2013:199:TOC
sha1:G4EPBY5RXZROKR7SL57ZNSHWATUQ7GDI - -
2019-05-04T21:02:30.654Z   302       379 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2013:200A:TOC
sha1:GAWT4USB6EDOVBMJDJE2B7R5JWHB5M4G - -
2019-05-04T21:02:30.654Z   302       379 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2013:193:FULL
sha1:RFSNTIDFNTAPFCLB7DSPEPIDNSMFTN4H - -
2019-05-04T21:02:30.656Z   302       381 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2013:195A:FUL
sha1:BUMEIXRBXUGFLDUDM3UFRIH2EJFXYSDO - -
2019-05-04T21:02:30.657Z   302       379 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2013:198:FULL
sha1:34CMQYVVROBBRZWTEE6GTIJH5TWVU7GM - -
2019-05-04T21:02:30.704Z   302       379 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2013:202:FULL
sha1:P4SMYD2EOKKWN6TWNHUMRVCPPXI4LIDF - -
2019-05-04T21:02:31.736Z   302       381 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2013:193A:FUL
sha1:L3Z4PWWONKY6IQVJ5MWY6MAUZ44PALSI - -
2019-05-04T21:02:31.742Z   302       377 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2013:202:TOC
sha1:CK7WT5U642LBLWJIDEZ43CFLN57VM7ML - -
2019-05-04T21:02:31.744Z   302       379 https://eur-lex.europa.eu/legal-content/DE/AUTO/?uri=OJ:C:2013:195A:TOC
```

The National Archives

C

MirrorWeb

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)
```

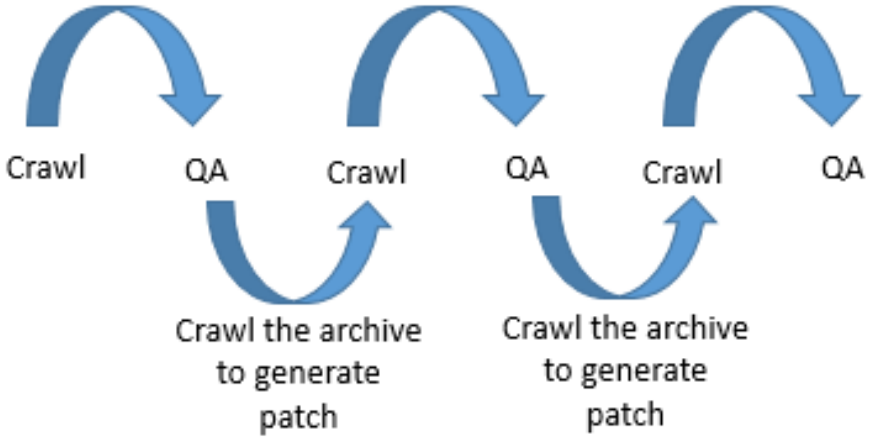
The National Archives

MirrorWeb

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;



The National Archives

MirrorWeb

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.

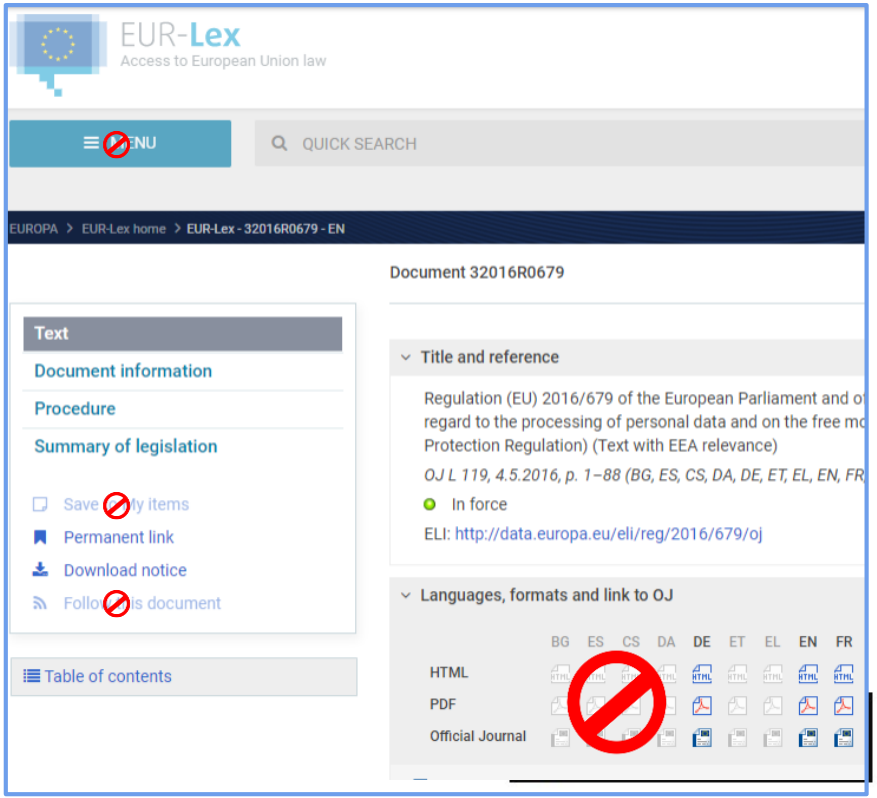
This manifests itself in three main ways:

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

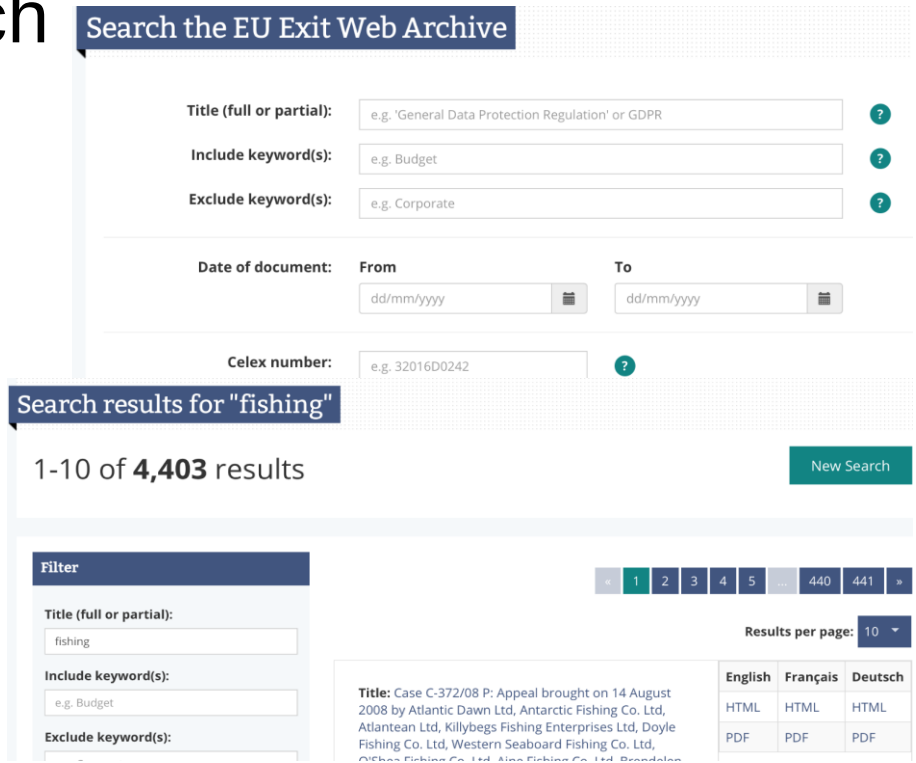
Standard PYWB playback

With EEWA modifications



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



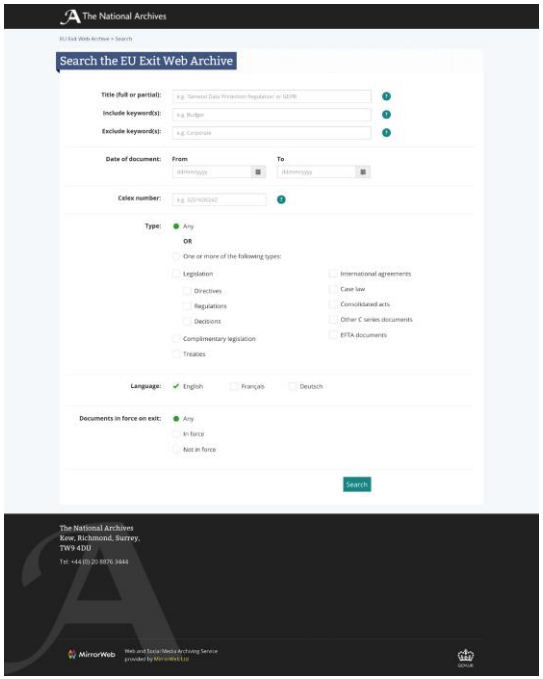
The National Archives

MirrorWeb

Developing Access - Search

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

After



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/ in the coming months...

The National Archives

T

MirrorWeb

Thank you!

webarchive@nationalarchives.gov.uk

The National Archives

T

MirrorWeb