‘The Least Cool Club in the World’

Building Capacity to Deal with Challenging Crawls at the UK Government Web Archive’s ‘RegEx Club’

THE NATIONAL ARCHIVES

Everyone hates crawler traps

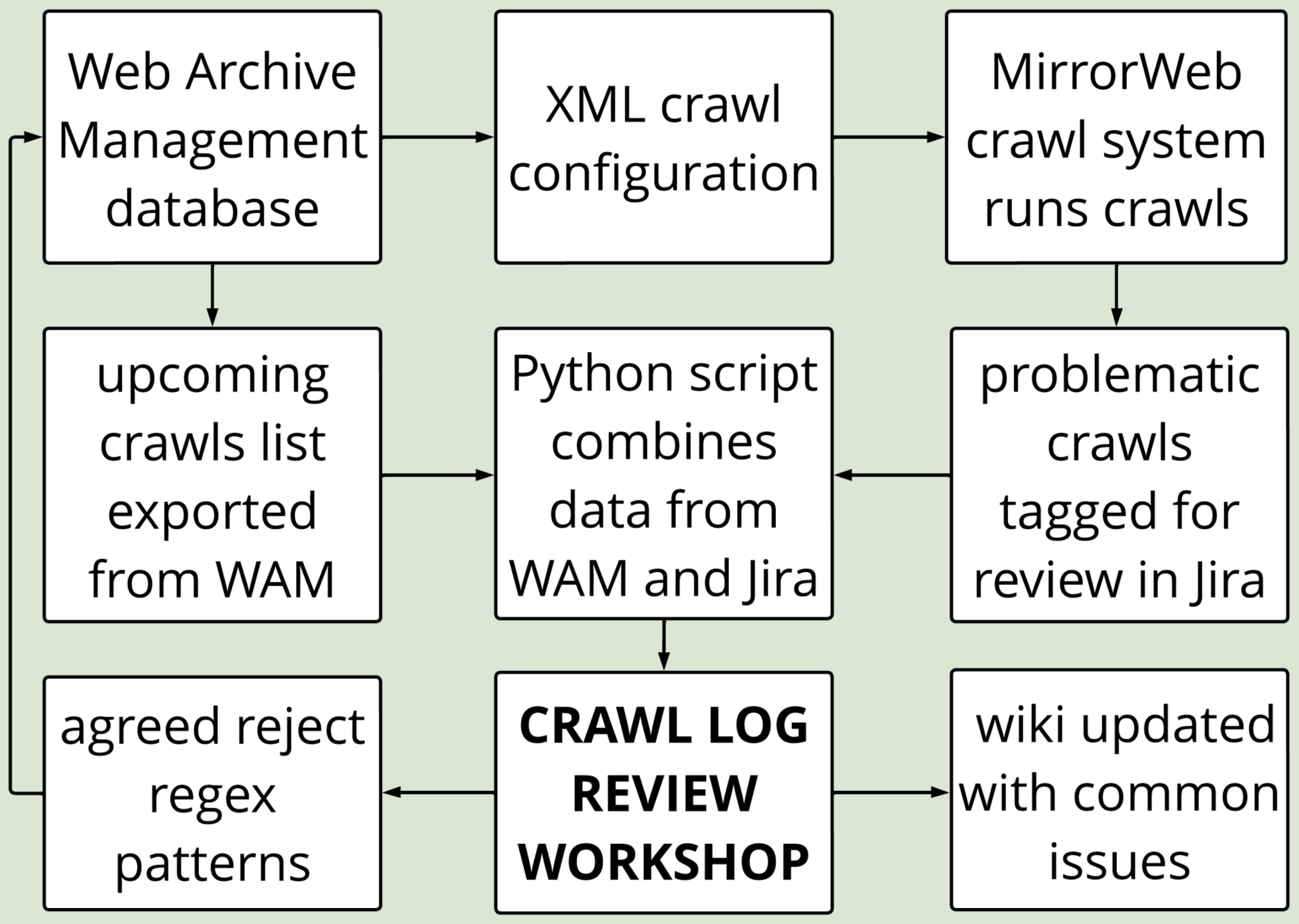
- They waste time, compute and storage.
- They capture useless or out-of-scope content.
- Force stopping problematic crawls might make us miss important content.

But limiting crawl scope is scary!

- Crawler traps are complicated – there are myriad causes and effects.
- What if we miss something important?
- RegEx is arcane wizardry!

We address these issues by working collaboratively in ‘RegEx club’

- Crawl Log Review (aka ‘RegEx Club’) is a bi-weekly workshop, where we examine the logs of problematic crawls in small groups and collectively decide how to fix them, usually with reject RegEx patterns.
- The workshop format helps us share knowledge of crawler traps and RegEx, and come to a consensus on the best way to improve each crawl.
- This is an iterative process: sometimes we don’t get the fix right first time, and sometimes ‘fixing’ one trap can reveal other problems. In that case the crawl is flagged again and we refine our approach next time.

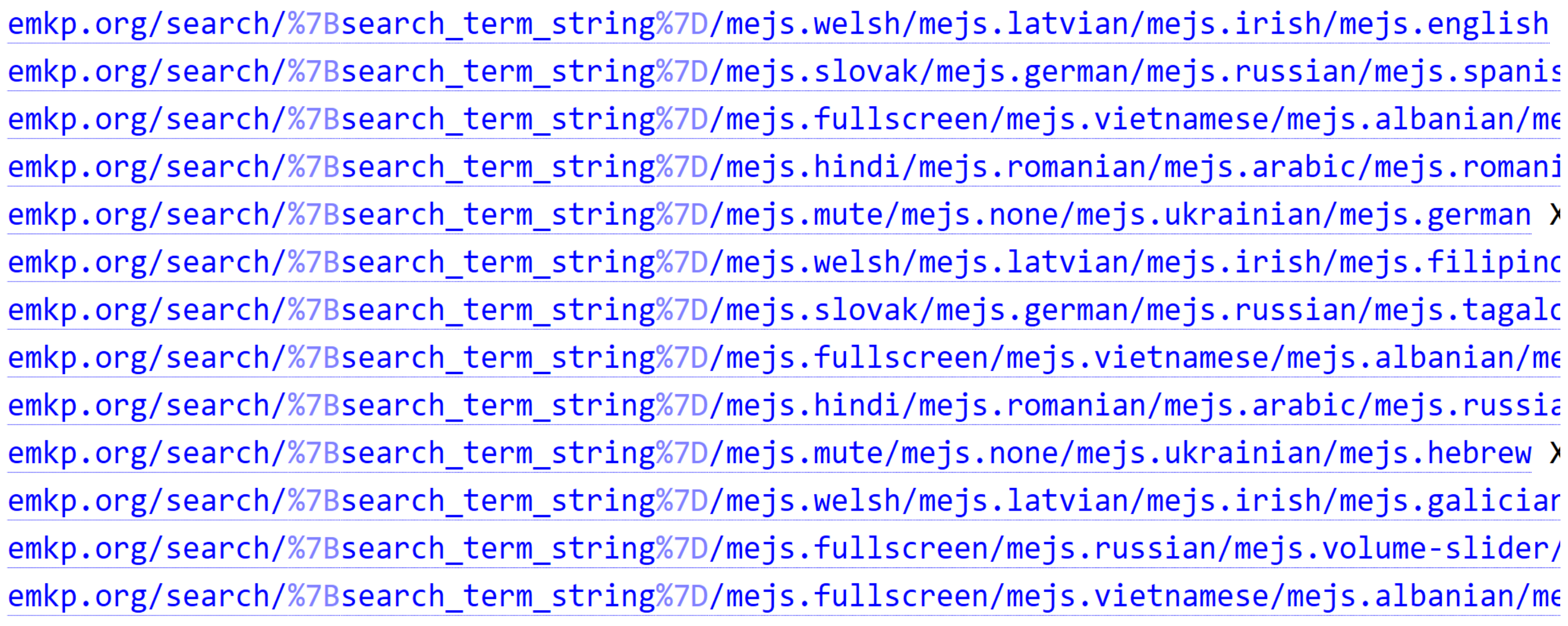


Web Archive Management database
XML crawl configuration
MirrorWeb crawl system runs crawls
upcoming crawls list exported from WAM
Python script combines data from WAM and Jira
problematic crawls tagged for review in Jira
agreed reject regex patterns
CRAWL LOG REVIEW WORKSHOP
wiki updated with common issues

The process looks complicated, but it’s really all about getting the information we need and then working on it together. You don’t need to do all this to start your own RegEx Club!

Our favourite crawler traps

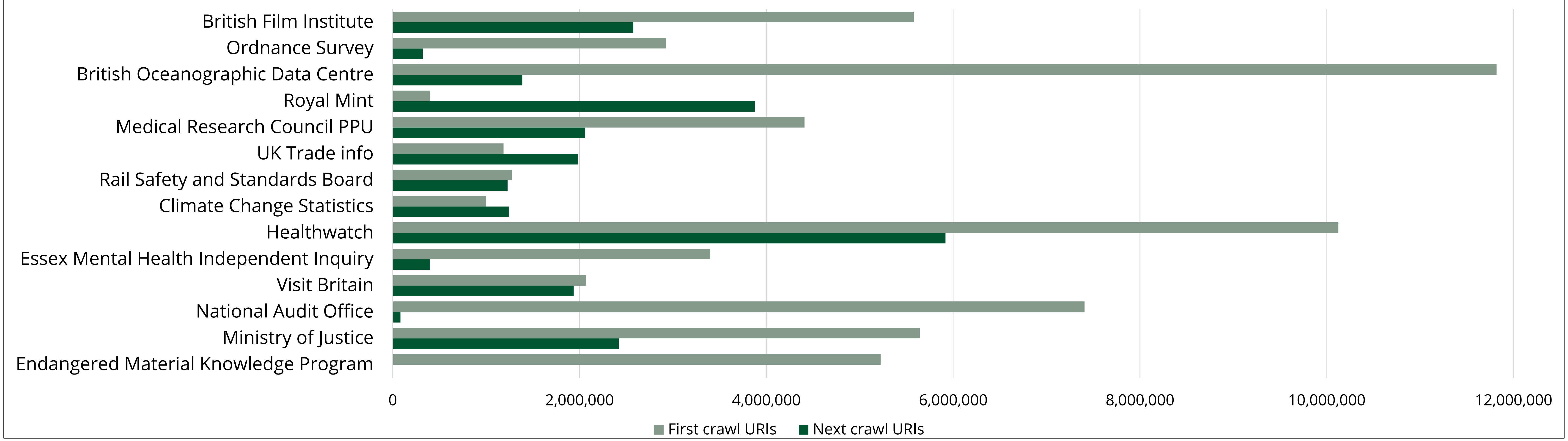
- Search facets – infinite fun combinations!
- Plugins like mejs.js – which generates thousands of repetitive URLs with different language options.
- Erroneous search strings – where the crawler repeatedly inserts plugin or script-related content into a site’s search facility.
- Trackers & social media – a particular issue with browser-based crawlers.



```
emkp.org/search/%7Bsearch_term_string%7D/mejs.welsh/mejs.latvian/mejs.irish/mejs.english
emkp.org/search/%7Bsearch_term_string%7D/mejs.slovak/mejs.german/mejs.russian/mejs.spanis
emkp.org/search/%7Bsearch_term_string%7D/mejs.fullscreen/mejs.vietnamese/mejs.albanian/me
emkp.org/search/%7Bsearch_term_string%7D/mejs.hindi/mejs.romanian/mejs.arabic/mejs.romani
emkp.org/search/%7Bsearch_term_string%7D/mejs.mute/mejs.none/mejs.ukrainian/mejs.german
emkp.org/search/%7Bsearch_term_string%7D/mejs.welsh/mejs.latvian/mejs.irish/mejs.filipinc
emkp.org/search/%7Bsearch_term_string%7D/mejs.slovak/mejs.german/mejs.russian/mejs.tagalc
emkp.org/search/%7Bsearch_term_string%7D/mejs.fullscreen/mejs.vietnamese/mejs.albanian/me
emkp.org/search/%7Bsearch_term_string%7D/mejs.hindi/mejs.romanian/mejs.arabic/mejs.russia
emkp.org/search/%7Bsearch_term_string%7D/mejs.mute/mejs.none/mejs.ukrainian/mejs.hebrew
emkp.org/search/%7Bsearch_term_string%7D/mejs.welsh/mejs.latvian/mejs.irish/mejs.galiciar
emkp.org/search/%7Bsearch_term_string%7D/mejs.fullscreen/mejs.russian/mejs.volume-slider/
emkp.org/search/%7Bsearch_term_string%7D/mejs.fullscreen/mejs.vietnamese/mejs.albanian/me
```

The ‘mejs’ crawler trap – in this example we were able to reduce future crawls by over 5,200,000 URIs.

Number of URIs before and after Crawl Log Review



Jake Bickford, The National Archives (UK) – jake.bickford@nationalarchives.gov.uk