THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Can Common Crawl Reliably Track Persistent Identifier (PID) Use Over Time

Can Common Crawl Reliably Track
Persistent Identifier (PID) Use Over Time?

Henry S. Thompson
University of Edinburgh
Edinburgh, United Kingdom
ht@inf.ed.ac.uk

TONG Jian*
University of Edinburgh
Edinburgh, United Kingdom
s1615354@sms.ed.ac.uk

ABSTRACT

We report here on the results of two studies using two and four monthly web crawls respectively from the Common Crawl (CC) initiative between 2014 and 2017, whose initial goal was to provide empirical evidence for the changing patterns of use of so-called persistent identifiers. This paper focusses on the tooling needed for dealing with CC data, and the problems we found with it. The first study is based on over 10^{12} URIs from over $5x10^9$ pages crawled in April 2014 and April 2017, the second study adds a further $3x10^9$ pages from the April 2015 and April 2016 crawls. We conclude with suggestions on specific actions needed to enable studies based on CC to give reliable longitudinal information.

KEYWORDS

temporal web analytics, persistent identifier, Common Crawl, Uniform Resource Identifier, longitudinal web crawl analysis, Digital Object Identifier

1 INTRODUCTION

The history of efforts to meet the demand for so-called 'persistent identifiers' (PIDs) for use on the Web is complicated, with many alternative offerings and much debate about the meaning of persistence and how to go about ensuring it. We take no position in that debate here, beyond the observation that the demand for PIDs shows no signs of abating, and that there has been a more-or-less general acknowledgement over the last 5–10 years that to be successful in the context of the Web a PID scheme must define and support a mapping from PIDs in the scheme to 'actionable' identifiers. In practice this has meant specifying a purely syntactic procedure for converting a PID into an http(s): URI using a domain owned and operated by the proprietors of the scheme. An HTTP request for such 'actionable' URIs will typically result in a redirection to the then-current location of the identified resource.

The Digital Object Identifier scheme [11], managed by the International DOI Foundation (IDF) [16], was an early adopter of this approach, and DOIs are now in widespread use, particularly in scientific journals, where their use is actually mandated by a number of major publishers. The mapping for DOIs to actionable https: URIs is simple: For example a DOI for a journal article written in the form of a URI such as doi:... is mapped (client-side) to https://doi.org/...[1]. In response to an HTTP request for that URI, the server at doi.org (operated on behalf of IDF by the Corporation for National Research Initiatives (CNRI) [5]) will respond with a redirect to the appropriate http(s): URI from the actual publisher of the article. We call the three forms involved the 'original' (e.g. doi: or info:hdl), the 'actionable' (e.g. https://doi.org/... and variants thereof or http://hdl.handle.net/...) and the 'locating'. Note that *none* of these is strictly speaking a PID as such: that's what comes *after* the doi: or https://hdl:handle.net/.

The success of this approach has overcome a significant barrier to the adoption of PIDs in general: to date there has been no significant move towards support for any of them *as URIs* in web browsers or PDF viewers. That is, if you try to use doi://10.1000/182 or info:hdl/20.1000/100 as a link (for example, as the value of the **href** attribute of an HTML **A** element), it will not work. But you *can* use them as the link text of an **A** element, and put the actionable form (https://doi.org/10.1000/182 and http://hdl.handle.net/20.1000/100 respectively) in the **href** attribute, and *that* will work just fine.

That's the good news. The less good news is that the use of redirection from the actionable form to the locating form means that when someone follows a link such as those in the previous paragraph, it's the *locating* form that appears in the address bar of their browser, and is thus the form they may well copy and paste into an email to a colleague or their own reading list. But this undermines the fundamental value proposition of the original ('persistent') form: that it is not vulnerable to all the things that cause http: URIs to fail over time.

Our goal in the work reported here was to quantify the growth over time in actual usage of the three forms, to see not only how good the good news was, but also whether there was cause to worry about the less good news: are locating forms 'leaking' into public use?

For concrete evidence we used the Common Crawl sample of HTML pages on the Web [3], the only large-scale public source of evidence readily available to us. This turned out to be challenging in a number of respects, to the extent that although our results are interesting, problems with the CC data mean that they may not accurately reflect the actual situation. In what follows we will first

*SURNAME forename

[1] doi: is not (yet) a registered URI scheme, but often used as if it were one

describe the work as such, and then discuss the ways in which the CC data fell short of what we think is required for reliable analysis.

Note on terminology Although most of the PIDs in various forms (original, actionable or locating) found during our studies were DOIs, we will be careful hereafter to use 'PID' when we mean *anything* recognised as a form of persistent identifier, and 'DOI' for the subset thereof which are some form of DOI.

2 PRIOR WORK AND OTHER SOURCES OF INFORMATION

An excellent overview of the space of PIDs and arguments for their use, only slightly dated, can be found in [21]. The IDF's views on the need for PIDs and their goals for DOIs is described in [1].

The IDF occasionally update their "Key Facts" page [12] which currently says that

- [DOIs are] Currently used by well over 5,000 assigners, e.g., publishers, science data centres, movie studios, etc.
- Approximately 148 million DOI names assigned to date
- Over 5 billion DOI resolutions per year

The leading issuer of DOIs for publications is CrossRef [7], who publish regularly-updated statistics about membership numbers, DOIs registered, etc. [8]

The leading issuer of DOIs for research data (as opposed to publications) is DataCite [9] who similarly publish statistics of the number of data-specific DOIs issued, cited etc. [10]

The only longitudinal study for PID usage we are aware of is [25]. They processed approximately 1.8 million scholarly articles published between 1997 and 2012, drawn from arXiv.org, Elsevier journals and PubMed Central, yielding a total of 2.2 million URIs. Of these there were

- 397,412 actionable-form DOIs (all using `dx.doi.org`)
- 505,657 "should-be-DOIs"

Their results are difficult to compare to ours, not only because they were looking at a disjoint set of years, but also because they didn't actually *look up* the actionable-form DOIs they found and then tabulate the occurrences of the resulting locating-form URIs. Instead they used "a list of hash values of publisher [domain names] provided by CrossRef. If the hash of a [domain name] of an extracted reference matches a hash in CrossRefâĂŹs list, a reference is [considered to be a should-be-DOI]." This was because their goal, as the name suggests, was to identify references that *could* have been DOIs, because the publisher was a CrossRef member and so would have assigned a DOI to the article in question. This is not quite the same goal as ours, which was to measure the ratio of actionable-form to locating-form PIDs *for the same individual article*.

3 MATERIALS

3.1 First study

Our first study, of the use of PIDs in all three forms, compared usage in April 2014 with that in April 2017, based on the Common Crawl sample of HTML pages on the Web [3] for those months. Table 1 gives basic size information for this sample.

The difference between the "URIs crawled" and "Pages retrieved" columns in this table, particularly for 2014, signal a problem with the same URI being retrieved multiple times. Although the crawl always

Table 1: Crawl size for first study [24]

Crawl month	URIs crawled	Pages retrieved	Dup URI %age
2014-04	1,718,646,762	2,641,371,316	34.9%
2017-04	2,907,715,349	2,942,930,482	1.2%

Table 2: Duplicate page estimates for first study [24]

Crawl month	Pages retrieved	Digests	Dup pages %age
2014-04	2,641,371,316	2,250,363,653	14.8%
2017-04	2,907,715,349	2,915,114,582	0.9%

starts with a unique set of URIs and does not follow page-internal links, redirects to URIs in the initial set occur surprisingly often, giving rise to duplication in some cases. The "Duplicate URI %age" column in Table reftab:t1 reports this, as *estimated* by subtracting the ratio of the URI to Page columns from 1. Detecting instances of this problem and not including pages from the duplicates has improved considerably between 2014 and 2017, as can be seen from the convergence of the "URIs crawled" and "Pages retrieved" columns and the big drop in the duplicate URI percentage estimate.

This duplication does not always mean that duplicate pages have been retrieved – as the crawl takes several weeks to complete the identified page may have changed. A direct estimate of the number of duplicate pages retrieved, based on comparing Hyperloglog digest values, is shown in Table 2. We'll return to the impact this has on our DOI tabulations in the Results section below.

3.2 Second study

Our second study added crawls from April 2015 and 2016, but focused exclusively on URIs using the `doi:` scheme. Table 3 combines the columns from Tables 1 & 2 and includes these additional years.

The Common Crawl makes data from each crawl available in 3 variants of the WARC format [15],[17]:

- WARC for the raw crawl data;
- WAT [18] for computed metadata, including request and response headers and, for responses, link tabulations from HEAD and BODY, using JSON
- WET for plaintext from the BODY

In both studies we worked exclusively with the WAT format, as that contains the link data we were interested in without the additional overhead of the entire HTML response. The number of files, average number of request/response pairs reported and the approximate total compressed WAT file size (in terabytes) is shown in Table 4.

It should be noted that for 2014 and 2017, the number of actual request/response entries recovered from the WAT files was slightly less than the numbers published by Common Crawl: approximately 4 million less in 2014 and 600,000 less in 2017.

Table 3: Crawl number for all four years [24]

Crawl month	URIs crawled	Pages retrieved	Dup URI %age	Digests	Dup pages %age
2014-04	1,718,646,762	2,641,371,316	34.9%	2,250,363,653	14.8%
2015-04	1,934,559,347	2,115,818,059	8.6%	1,910,978,257	9.7%
2016-04	1,335,046,923	1,335,046,923	0.0%	1,211,048,216	9.3%
2017-04	2,907,715,349	2,942,930,482	1.2%	2,915,114,582	0.9%

Table 4: Sizes for the second study

Crawl month	WAT file count	pages per file	Total size (TB)
2014-04	44488	59373	17
2015-04	38609	54801	14
2016-04	22200	60137	9
2017-04	64700	45486	19

4 METHODS

4.1 First study

For the first study we wanted to check every link from the body of each crawled HTML page, which meant downloading around 110,000 WAT-format files totalling around 36TB in (compressed) size.

We achieved this by streaming about 1/10th of the data each night, divided over approximately 100 machines that were detected as idle in one of several student labs. Each machine tabulated summary counts for approximately 100 WAT files each night, taking 4–6 hours. These were uploaded to a central machine and merged. The process was slightly different in 2014 and 2017: only in 2017 did we look for PIDs in their locating form, as explained below.

4.1.1 2014 crawl.

There were 44488 WAT files to be processed, containing information from a total of 2,534,229,771 pages. For each page the WAT file contains three JSON objects, one each for information about the crawl, the HTTP request and the HTTP response. We extracted the latter, and from it the following three components:

- `Envelope/WARC-Header-Metadata/WARC-Target-URI` (a string)
- `Envelope/Payload-Metadata/HTTP-Response-Metadata/ Headers/Content-Type` (a string)
- `Envelope/Payload-Metadata/HTTP-Response-Metadata/ HTML-Metadata/Links` (an array, see below)

For each page we accumulated counts for

- the target URI scheme (always `http:` or `https:`)
- the target URI host (strictly speaking the 'authority' per RFC3986 [20])
- the Content-Type header

The contents of the .../Links component array are each an object with at least the following contents:

```
{ "path": [quasi-XPath, e.g. "A@/href", "IMG@/src",
           "FORM@/action"],
  "url":  [absolute or relative URI],
```

[other optional properties per path]}

For the value of the "url" property of each entry in this array we accumulated counts for

- the target URI scheme (possibly absent)
- the target URI host (possibly absent)
- if the host was one of a list of actionable PID resolvers (see below), the number of times the whole URI (normalised) appeared in the Links array

The resolvers we watched for were as follows:

```
doi.org, dx.doi.org, dx.medra.org
hdl.handle.net
n2t.net
```

The normalisation of the Link URIs involved

- removing spurious whitespace apparently arising from issues with the Common Crawl process itself;
- replacing both percent-encoded and HTML entity-encoded character forms

The Links array data is our primary concern in this paper. As the individual processor results were merged, the individual per-page tabulations enabled us to produce the following summary tabulation:

- The frequency of `http:` and `https:` URI schemes in both the crawled URI set and of (none), `http:`, `https:` and many other URI schemes in the Link URI set
- In particular, the frequency of `doi:` and `info:` in the Link URI set
- The frequency of the five resolvers in the Link URI set
- The frequency of each actionable URI in the Link URI set

(Note that only a handful of actionable-form URIs appeared in the crawled URI set)

For all but the first (URI schemes in general) frequency tabulations, we have both type and token frequency.

4.1.2 2017 crawl.

For the April 2017 crawl, we added two additional tabulations:

- Document frequency for actionable URIs, that is, the number of pages in which each appears, regardless of how many times
- For each URI in the Link set which is the locating form of an actionable URI in the 2014 Link set, type, token and document frequency

The latter counts were tabulated by taking all the 2014 actionable forms, issuing HTTP HEAD requests for them and noting the Location response header which came back (iterating and accumulating until a 200 response was achieved). This succeeded more that 99%

of the time, and a Bloom filter was constructed from the results, which then allowed us to check every Link URI as we processed the 2017 Link URIs.

These counts are restricted to the 2017 appearances of the locating forms of 2014 actionable forms, because we didn't have time to do two passes over either the 2014 data or the 2017 data.

For both years, the final step was to extract the PID itself (that is, the path part of the actionable-form URI, regardless of URI scheme, redirection server hostname or query parameters) and merge the counts across all the actionable-form URIs with the same PID. Unless otherwise noted, these are the counts reported in the Results section below.

4.2 Second study

The second study aimed to fill in the gap between 2014 and 2017, but at a much lower level of detail. It simply counted original-form DOI occurrences in the HTML head (in `link` and `meta` elements) as well as the body.

The scale of the 4 years' data is shown above in Table 4. This study was actually a pilot study to determine whether using 100 8-core computers with better bandwidth via Microsoft's Azure facility[2] would significantly increase throughput, and did in fact allow for one month's crawl data to be processed in about 6 hours, an improvement of about a factor of 8 over the first study.

As in the first study, only the 'response' JSON object was processed, extracting 3 components:

- Envelope/Payload-Metadata/HTTP-Response-Metadata/ HTML-Metadata/Links (as in study 1)
- Envelope/Payload-Metadata/HTTP-Response-Metadata/ HTML-Metadata/Head/Link (an array)
- Envelope/Payload-Metadata/HTTP-Response-Metadata/ HTML-Metadata/Head/Metas (an array)

Each member of the Metas array is an object, where the ones of interest had the following contents:

```
{ "name":    [the META element's name attribute]
  "content": [the META element's content attribute]}
```

and we counted objects where the "content" property was an original-form DOI.

Likewise for the Link array, where we care about

```
{ "rel":  [the LINK element's rel attribute]
  "href": [the LINK element's href attribute]}
```

and counted ones where the "href" property was an original-form DOI.

In contrast to the first study, all that was tabulated were occurrence counts per page of any original-form DOI, counts for the different DOIs themselves were not kept, so the net results were just three totals, first per WAT file, then after merging, per month.

Finally a very small sample, just 645 WAT files from April 2014 (1.5% of the total), was processed looking only at the Metas array, to count the different values of "name" whose "content" was an original-form DOI.

_____
[2]See 6.3

Table 5: Link URI counts for first study

Crawl month	Link URIs		Actionable Link URIs		Distinct PIDs
	Total	Corrected	URIs	Ratio	
2014-04	$299x10^9$	$194x10^9$	30,445,532	0.00016	5,369,831
2017-04	$620x10^9$	$613x10^9$	37,913,544	0.00006	12,659,694

Table 6: Shared vs. one-year-only PIDs in the first study

	2014	not 2014
2017	3,354,906	9,304,788
not 2017	2,014,925	0

Table 7: PID scheme and resolver counts from distinct actionable-form PIDs

When	DOIs	handles	other
2014 only	1,656,913	357,997	15
2014 and 2017	2,914,930	439,969	7
2017 only	7,383,189	1,914,395	7,204

5 RESULTS AS SUCH

In this section we present the results as if the data they are derived from gave reliable evidence. Discussion of reasons to fear this may not be the case and suggestions for what to do about this are given in section 6

5.1 First study

Counts for all Link URIs, the actionable-form URI subset thereof and distinct PIDs extracted from those, as tabulated across April 2014 and 2017, are shown in Table 5.

Two columns are shown for the total number of Link URIs: The first column is the actual number we found, the second is adjusted downwards for the estimated degree of duplication, as reported above in Table 1. This correction is not needed for the actionable URI and PID columns (see section 6), but is given here as it is used for the ratios given in the Actionable Link URIs Ratio column.

The overlap between the sets of URIs crawled in April 2014 and April 2017 is low (estimated at 7%) and for the responses (pages) themselves even lower (estimated at 0.8%) [6] However the PID numbers have a much higher overlap: the union of the two years contains only 14.7 million PIDs – the details are given in Table 6. This suggests that the overlap PIDs are *very* popular, as they are not just persisting from 2014 to 2017, but their second appearance is in a *different* set of pages.

As mentioned earlier, the actionable-form PIDs that we looked for can be divided on the basis of the domain name used to identify their resolving proxies: `doi.org`, `dx.doi.org` and `dx.medra.org` for DOIs, `hdl.handle.net` for handles and `n2t.net` for ARKs and other PIDs (we didn't explore these in any detail). The numbers in each category are shown in Table 7.

Table 8: Locating-form URIs in 2017 for 2014 actionable-form PIDs

	Distinct	Total
Actionable found in 2014	5,369,831	12,642,054
Retrieved locating form	5,315,129	
Locating found in 2017	413,397	1,202,610
Ratio	8%	10%

Table 9: Growth of `doi:` usage for second study

	Body links		Head `meta`	
year	n	per mil pg	n	per 10K pg
2014	1893	0.72	731938	2.77
2015	1410	0.67	727167	3.44
2016	1440	1.08	410603	3.08
2017	3550	1.21	459328	1.56

The relative recency of the arrival of the `n2t.net` resolver on the scene is clearly evident here.

Finally the locating-form leakage question is addressed in Table 8, which gives the number of locating-form URIs retrieved for the actionable-form URIs found in 2014.

There were 12+ million actionable-form URIs found in 2014, from which 5+ million distinct PIDs were extracted, almost all of which successfully yielded locating URIs. Of these around 400,000 (8%) by type count, or 1.2 million (10%) by token count, occurred in body links in the the 2017 crawl. There is of course no way to tell whether these usages arose from the kind of leakage scenario discussed in the Introduction, or whether they were found and used independently of the antecedent actionable-form URI, but either way this is a large enough number to be of some concern.

5.2 Second study

Adding data for April 2015 and 2016 allows us to track the growth of `doi:` use in HTML body and head links (distinguishing between `meta` and `link` elements). In head `link` elements we found no (!) uses of the `doi:` form in April 2014 or 2015, and only 2 in April 2016 and 2017, so the data in Table 9, with graphs in Figure 5.2, only report on the numbers for use in body links and head `meta` elements.

It's interesting to see that a small number of original-form `doi:` URIs are appearing as e.g. `A/@href`, and that this usage is slowly increasing. It's certainly not obvious *why* anyone would do this: It would be necessary to look at the complete HTML pages to make sense of this.

The much more substantial use in HTML head `meta` elements is, on the other hand, quite plausible, although the drop in 2017 is hard to evaluate without seeing data from the surrounding months.

We did a quick check of around 1.5% of the April 2014 to see which meta tags the doi: URIs were being used with. Table 10 gives the rank-ordered results.

Figure 1: Growth in `doi:` use in body links and in head meta
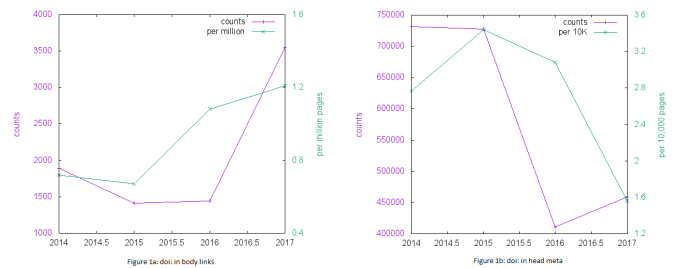
Table 10: What `meta` tags are `doi:` URIs used for?

Tag	Count		Tag	Count
dc.identifier	6548		dcterms.isReferencedBy	4
eprints.id_number	1174		eprints.related_url_url	2
citation_doi	435		keywords	2
dc.Identifier	146		bepress_citation_doi	1
dcterms.isVersionOf	105		dc.citation.spage	1
dc.relation	44		eprints.data	1
dcterms.hasPart	15		eprints.doi	1
dcterms.isPartOf	12		eprints.note	1
			eprints.official_url	1

The vast majority of these are Dublin Core [13] or EPrints toolset [14] tags.

Given the quite large number of doi: URIs showing up in the HTML header as `META/@content`, it's a bit surprising not to find any as `LINK/@uri`.

5.3 Conclusions for DOIs

In summary, the results of the two studies show

- Virtually no use of original-form URIs in head links
- Only small numbers of 1000s of original-form URIs in body links
- Significant, slowly increasing, use (100s of thousands) of original-form URIs as meta-information
- Much larger numbers (millions) of actionable-form URIs in body links
- A 2.5 times increase in the number of distinct DOIs in body links between the 2014 and 2017 crawls
- For about 8% of the actionable-form URIs in the 2014 crawl, the corresponding locating-form URI appears in the 2017 crawl

6 CONCLUSIONS FOR LONGITUDINAL STUDIES

The work reported here cannot be taken as anything more than a starting point, demonstrating that it is possible to extract longitudinal information about URI usage and encouraging others to do so:

neither the numbers nor the trends presented here can be claimed to be reliable. It illustrates the kind of questions we would like to get answers for with respect to one kind of longitudinal study, and the very *lack* of reliable answers to those questions points towards the things we need to do to improve the situation.

In what follows we look at a number of different kinds of problem we encountered and suggest possible remediations.

6.1 Common Crawl itself

Duplication of pages crawled within an individual release is an issue for any use of Common Crawl data. Duplication of pages between releases may be a bug or a feature for longitudinal study, but its existence needs to be taken account of in any case.

A number of discussions of these issues can be found in the Common Crawl forum [4], and it does appear that within-release duplication has been considerably reduced.

However, for our study, and as noted above in the Materials section, the April 2014 data shows a substantial degree of likely duplication at the page level. This is almost entirely due to two sources [23]

(1) Shared error pages;
(2) Same page for distinct URIs differing only in query parameter values.

It seems reasonable to assume that error pages are unlikely to involve much PID use, and the same is true for the kinds of commercially-orientated applications which make heavy use of query parameters. The latter expectation is easy to check empirically, and a quick check of a random sample (3425 actionable-form (a mixture of doi.org and dx.doi.org) URIs from 4 different WAT files from April 2014) confirms this: none of them have query parameters.

There's clearly a pressing need for a careful study of the last 3 or 4 years of CC data, to establish in detail the within- and between-release overlaps, both with respect to content and URI (see also section 6.2). CC's own version of this information [6] covers 2015 onwards, but has not as far as we know been published in a peer-reviewed context or otherwise confirmed.

In harvesting URIs from the Links component of the response records in a CC WAT file, we encountered a wide range of low-level problems with format and character encoding. Some of these did *not* occur in the original, in the few cases we checked by hand and could find. Although tedious, a survey of the kinds of errors introduced in the WAT files is needed, to at least document their frequency over time, but also to try to establish which can be detected reliably and, of those detectable, which can be reliably corrected. For those problems which persist in more recent releases, we would hope once alerted to them CC could fix the problem at source going forward.

Correct reporting of links that are found is important, but so is actually reliably *detecting* links—some empirical checking of this would also be a good idea.

The recent rapid growth of personalised responses based on information in the query string of URIs and/or on cookies has serious implications for the 'Identifier' aspect of URIs, and the extent to which responses to URIs which share their `authority` and `path` components but not their `query`. Again, at least some comparison

of CC crawl target URIs both with and without including the `query` component is needed.

For the two CC releases we have checked, that is April 2014 and April 2017, the good news is that the percentage of HTTP vs. HTTPS is close as between URIs crawled and Link URIs seen (around 20 to 1 in 2014, dropping to 3.3 to 1 in 2017, reflecting the success of initiatives such as Encrypt the Web [19]. A more systematic tabulation over all releases is obviously needed before we can tell whether this is a reliable trend or not. How representative a sample the CC HTML is of Web HTML as a whole is unknown, and indeed it's not clear how one would quantify this.

A *much* more serious coverage issue, particularly for the persistence issues we started out to explore, is the lack of anything other than HTML documents in the CC releases. For scholarly publications, which are a major market for PIDs, PDF is the preferred format for publication. Expanding CC to include PDF files clearly would be a major undertaking, but at least some attempt to crawl links from a CC release to PDF files would be useful to get some sense of how much the profile of links found *there* differs from that in the HTML data.

6.2 Versioning and deduplication

Detecting and at least tabulating, preferably eliminating, exact duplicate content is of course important, but for at least some kinds of longitudinal studies, detecting and relating multiple versions of the 'same' content is also important. Detecting similar-but-different content retrieved from different (post-redirection) URIs is obviously non-trivial, as it depends implicitly on some notion of 'sufficient' similarity to count as the same version. Plagiarism detection software has a contribution to make here. Even quite a tight threshold might be very useful: in a way two documents which differ only by a tiny change, say a single spelling correction, is much worse than two identical document, because hash-based methods will find the latter but not the former.

6.3 Scale

At the very least the variability and occasional unreliability of the Common Crawl data means that for improved confidence the usage being studied should be tabulated for *every* month's crawl over a at least a year. With hindsight it would also probably be wise to use crawls from 2015 onwards, as both the effort to remove duplicates and the documentation improve noticeably at that point. This in turn however begins to move the effort involved out of the reach of the kind of *ad-hoc* multiprocessor we assembled and used for the first study. Even the 6-hour turnaround we achieved for the second study still made debugging a tedious and potentially expensive process. We have some speed-ups in mind, but they are unlikely to gain us more than a factor of two or so. Generosity of the sort provided by the donor of cloud resources for the second study will be needed if academic longitudinal studies of Web usage are to reach the levels of reliability and utility we need.

In conclusion, Common Crawl releases since 2015 provide a potential basis for longitudinal studies of HTML web page content and linking, but results have to be treated with caution. A number of gaps in documentation and quality assurance need to be addressed before conclusions based on such studies can be taken as reliable.

ACKNOWLEDGMENTS

REFERENCES

[1] 2015. DOI System and Internet Identifier Specifications. (6 May 2015). Retrieved 2017-11-02 from https://www.doi.org/factsheets/DOIIdentifierSpecs.html

[2] 2016. (17–19 Feb 2016). Retrieved 2018-01-23 from https://github.com/identifier-services/phoibos2/wiki Workshop on Practical Hacking On Identifiers.

[3] 2017. Common Crawl. (2017). Retrieved 2017-11-02 from http://commoncrawl.org/

[4] 2017. Common Crawl forum. web forum. (2017). Retrieved 2017-11-02 from https://groups.google.com/forum/#!forum/common-crawl

[5] 2017. Corporation for National Research Initiatives. (2017). Retrieved 2017-11-02 from https://www.cnri.reston.va.us/

[6] 2017. Crawl overlap tabulations. (2017). Retrieved 2017-11-02 from https://commoncrawl.github.io/cc-crawl-statistics/plots/crawloverlap

[7] 2017. CrossRef. (2017). Retrieved 2017-11-02 from https://www.crossref.org/

[8] 2017. CrossRef facts. (2017). Retrieved 2017-11-02 from https://www.crossref.org/dashboard/

[9] 2017. DataCite mission. (2017). Retrieved 2017-11-02 from https://www.datacite.org/mission.html

[10] 2017. DataCite stats. (2017). Retrieved 2017-11-02 from https://stats.datacite.org/

[11] 2017. DOI Handbook. (2017). https://www.doi.org/hb.html

[12] 2017. DOI key facts. (2017). Retrieved 2017-11-02 from https://www.doi.org/factsheets/DOIKeyFacts.html

[13] 2017. Dublin Core. (2017). Retrieved 2017-11-02 from http://dublincore.org/

[14] 2017. EPrints. (2017). Retrieved 2017-11-02 from http://www.eprints.org/uk/

[15] 2017. *Information and documentation – WARC file format.* Retrieved 2017-11-02 from https://www.iso.org/standard/44717.html

[16] 2017. International DOI Foundation. (2017). Retrieved 2017-11-02 from https://www.doi.org/

[17] 2017. So you're ready to get started. (2017). Retrieved 2017-11-02 from http://commoncrawl.org/the-data/get-started/

[18] 2017. WAT format spec. (2017). Retrieved 2017-11-02 from https://webarchive.jira.com/wiki/spaces/Iresearch/pages/13467719/Web+Archive+Metadata+File+Specification Not clear what if any official status this has...

[19] 2018. Encrypt the Web. (2018). Retrieved 2018-01-22 from https://www.eff.org/encrypt-the-web

[20] T. Berners-Lee, R. Fielding, and L. Masinter. 2005. RFC 3986, Uniform Resource Identifier (URI): Generic Syntax. (2005). https://tools.ietf.org/html/rfc3986

[21] Ruth E. Duerr, Robert R. Downs, Curt Tilmes, Bruce Barkstrom, W. Christopher Lenhardt, Joseph Glassy, Luis E. Bermudez, and Peter Slaughter. 2011. On the utility of identification schemes for digital earth science data: An assessment and recommendations. *Earth Science Informatics* 4, 3 (9 2011), 139–160. https://doi.org/10.1007/s12145-011-0083-6

[22] Greg Janée. 2016. Are PIDs effective. personal communication. (18 Feb 2016).

[23] Tom Morris. 2017. personal communication. (30 October 2017). Archived at https://groups.google.com/d/msg/common-crawl/DdEjqaVRwfg/q1qlACsrCAAJ.

[24] Sebastian Nagel. 2017. Understanding Common Crawl. personal communication. (23 October 2017). Archived at https://groups.google.com/d/msg/common-crawl/DdEjqaVRwfg/Y1zUd0dIAwAJ. See also these pages referenced in the email: https://commoncrawl.github.io/cc-crawl-statistics/plots/crawlsize; https://github.com/commoncrawl/cc-crawl-statistics.

[25] Herbert Van de Sompel, Martin Klein, and Shawn M. Jones. 2016. Persistent URIs Must Be Used To Be Persistent. In *Proceedings of the 25th International Conference Companion on World Wide Web (WWW '16 Companion).* International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 119–120. https://doi.org/10.1145/2872518.2889352