Methodological reflections about establishing a corpus of the archived web: The case of the Danish web from 2005 to 2015

Niels Brügger, Aarhus University, DK
Ditte Laursen, Royal Danish Library
Janne Nielsen, Aarhus University, DK

Work in progress, please not quote without contacting us first for an updated version.

1. Introduction

The aim of this paper is to discuss how a delimited subset — a corpus — can be identified in a web archive, and what impact different approaches to doing this has on the resulting corpus and thus on the research performed with the corpus. The point of view is that of the researcher who has to use the corpus for further studies. Thus, technical issues are only included where they can inform the discussion of the researcher's use.

We focus our discussion of the establishing of a corpus through two different but partly overlapping studies. The two studies use two different methods in creating a corpus. However, both studies are based on the holdings of Netarkivet, (the Danish national web archive), they have a historical approach, and they are focused on the decade 2005-2015. Moreover, the two studies are concerned with similar research questions, ie. size of domains and number of file types. For this reason, the two methods for creating a corpus can be discussed based on results about the same research questions.

The first study, a mapping of the development of the national Danish web archive Netarkivet from 2005 to 2015 was made in connection to the archive's 10th anniversary in 2015 (cf. Laursen & Møldrup-Dalum, 2017). The second study is ongoing and aims at mapping and analysing the historical development of the national Danish web from 2005 to 2015 as it has been archived in Netarkivet. This study intends to 'probe' the Danish web in a number of ways by answering questions such as: How big is the entire web domain? How many of each file type are there? What is the size of web domains? In addition, hyperlink structures as well as the content on the web pages are to be studied (cf. Brügger, 2017 for an outline of the project). Thus the two studies differ regarding their overall scope. The first was made to generate knowledge about Netarkivet and its development, whereas the latter is conducted to provide knowledge about the Danish web's evolution. Since the two studies were initially not made with the present study in mind some adjustments have to be made to be able to compare them (cf. below).

2. Approaches to establishing a corpus

With a view to analysing the historical development of Netarkivet and of the Danish web year by year, temporal 'slices' have to be identified each year. In other words: one corpus needs to be established per year, and if the corpus is to be used in a research project as the latter mentioned above it has to be as close as possible to the web as it looked in the past, given the available source material in the web archive.

This paper aims at investigating two approaches to establishing such an annual corpus in a web archive, based on different data sets and different handlings of them: (all data sets and what is understood by 'filtered' is explained below):

- an unfiltered crawl.log corpus
- a filtered full-text index corpus

The data sets and what is understood by 'filtered' will be explained below.

2.1 Establishing a corpus, why and how?

The present study shall illustrate the general challenge of identifying parts of a web archive to be included in a corpus. We assume that only very few researchers would like to study an entire web archive or web collection, and therefore there is a general need to be able to identify and delimit which parts of a web archive shall be studied. This study can be considered an extreme case because of its size, but the insights are expected to apply for smaller corpora as well. And even in cases where an entire collection is studied there may still be a need to filter the collection by removing versions of 'the same' web entity (cf. below about the filtered broad crawl).

It is worth reflecting on what exactly a corpus in a web archive can look like: is it a collection of ARC/WARC files? Is it a collection of files, extracted from ARC/WARC files? Or is it rather an index of pointers to the web objects in the web archive that have to be included in the study? In the present context we conceive the corpus as the latter of these three. Thus, the aim is to establish a file with a complete index of the addresses of all the web objects in the web archive to be included in the corpus, and with this 'master key' the web archive can be unlocked in different ways, depending on what exactly shall be studied. The workflow of establishing a corpus to fit the analytical purpose is illustrated in figure 1.

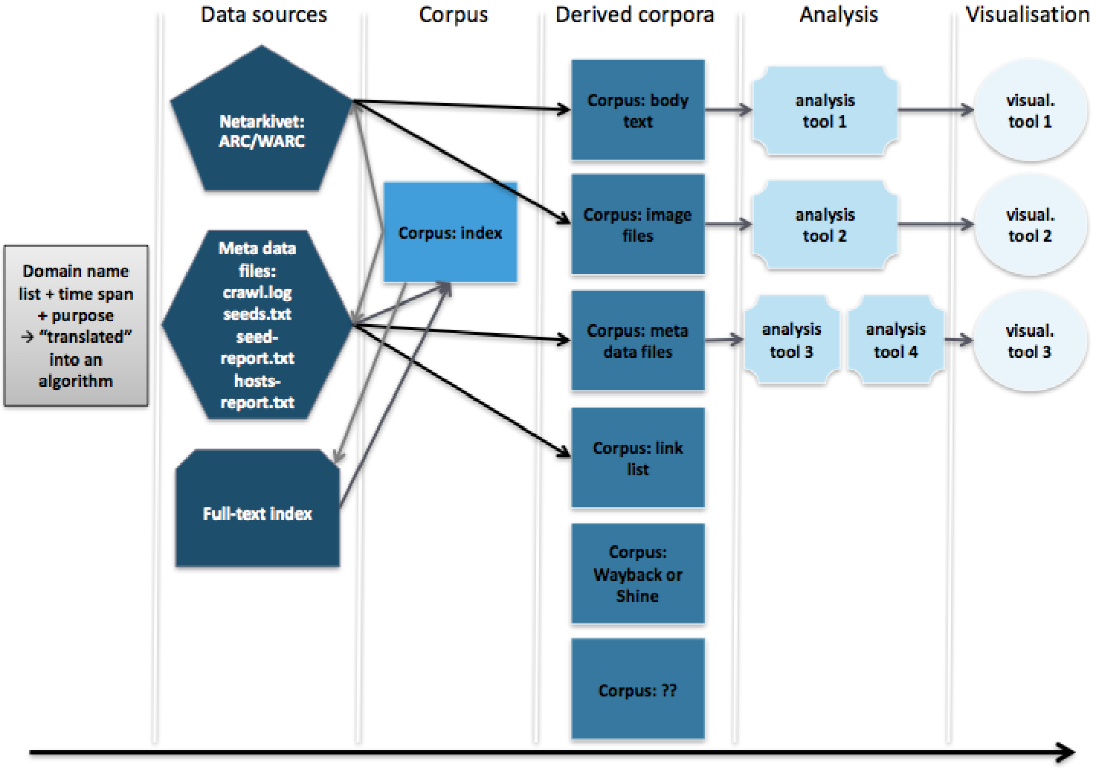

Figure 1: Workflow for web archive analysis, including the establishing of a corpus index file and derived corpora.

If one wants to perform analyses of hyperlinks only hyperlink information is needed, if one wants to study the text or the images only the text or images are needed, or if one sets out to calculate the number of specific file types only this information is necessary. In each of these analytical cases the corpus index file can be used to generate what we term a ‘derived corpus’: a list of hyperlinks, files with the text from the web or image files, or statistics about a specific topic. As indicated these derived corpora will have different formats, depending on the type of analysis to be performed, and once they are to be analysed this can be done by the use of analytical tools that can handle this format, e.g. Gephi for hyperlink analysis, text and image analysis tools, or statistical tools such as R. Some of these analytical tools also include visualisation tools to be used either during the analysis or to present the results. In this paper only the first step, the creation of the corpus index file, is discussed. However, this is also the most important step, because the corpus index file, together with the documentation about its creation, constitutes an index that can later be used by other researchers to make other derived corpora than the ones initially planned. Thus, the corpus index file also constitutes a preparation of the web archive for future research.

2.2 Analytical design and data used

With a view to investigating how different approaches to the establishing of a corpus affect what will be included in the corpus, and thereby the derived corpora as well as the analytical results we have chosen to compare two rather simple, yet illustrative phenomena: the size of the entire Danish web domain, and the number of specific file types, as they were archived by Netarkivet. In addition, these two points of comparison were chosen because they were the only ones that were both present in the data sets to be compared (as mentioned, none of the datasets were initially made with the present comparison in mind).

The collection from which the corpora are to be established is Netarkivet. Netarkivet was founded in 2005 and its activities are based on a legal deposit law from December 2004 which allows to archive web material within the Top Level Domain .dk and websites on other domains aimed at a Danish audience. Netarkivet uses three strategies for web harvesting: broad crawl harvesting, where .dk and Danish materials outside the Top Level Domain are harvested (app. 3 times per year), selective harvesting where frequently updated websites are harvested with shorter intervals (from several times daily to monthly), and, finally, event harvesting, where new websites in relation to an event are harvested.

Since the aim is to create a corpus index file, two data sources that were created by Netarkivet have been used. On the one hand the crawl.logs that are generated during the archiving of the web, and which contains information about the archiving process such as a time stamp, the HTTP response code, the size of the downloaded file in bytes, and the file type. The advantage of this data source is that it is a record of exactly what happened during the archiving process. On the other hand a full-text index (Solr index) of the entire web archive that has been generated since 2014 (Laursen & Møldrup-Dalum 2017). In contrast to the crawl.log the full-text index is made on the basis of what was actually archived and was preserved as WARC-files, and not on what happened during the archiving; in addition, the index has also included metadata of all web objects with a response code 200 (OK), such as metadata from the WARC file as well as information returned from the web server. In addition to these differences between the two datasets they also vary with respect to the so-called deduplication, that is the fact the a web archive can decide not to archive a web object if an identical copy is already in the collection. Since the crawl.log records what takes place during the archiving process, the same file that may be encountered say five times by the web crawler will be calculated five times, even if it is later removed as part of the deduplication process, whereas it will only be calculated one time in the full-text index, since the index is based on what was actually preserved by the web archive and not on what the web crawler encountered. However, not all file types are deduplicated, for instance HTML files are not deduplicated in Netarkivet, but image, video, and audio files are. This difference between the two datasets is important to bear in mind when evaluating the results of our comparison.

Finally, two other pieces of information from Netarkivet are used. First, the harvest ID: every harvest has a name (given by the curators), and is allocated a unique identification number, for instance the harvests named ‘2005-4-10MB’ and ‘2005-4-500MB’ have the harvest IDs ‘25’ and ‘27’. Second, the job ID: all harvests are composed of a series of jobs, each with their unique number; the job IDs are extracted from a job database.

As mentioned above the overall intention of the second study, the study of the development of the Danish web, is to base the study on a corpus that is as close as possible to the web as it looked in the past (as archived in Netarkivet). This means that if a given URL was archived more than one time within the time frame covered by the broad crawl it should only be included once in the corpus. The reason for this aim is that preliminary tests have shown that substantial parts of a given domain name may have been harvested several times during a broad archiving because it has been linked to from other domain names and thus has been crawled as a consequence of the web harvester following hyperlinks. We call these extra versions of a website *by-harvests*, whereas we use the term *main harvest* for the version of a web domain that the archive intended to archive (cf. Brügger, Laursen & Nielsen, 2015); thus the main harvest is the version of a web domain that was initially put into a crawl job whereas a by-harvest is the version of the same web domain that was accidently archived in other jobs, because the web domains in these jobs linked to it.

These extra versions of a web domain cannot be considered duplicates (and are therefore not deduplicated) since they are (probably) not identical, rather they are versions of the same (partly or in total), and since it can take between 2-4 months to perform a broad crawl the number of possible by-harvests can be quite high. If more versions of the same web entity are included in some cases but not in all the corpus this will as such bias the analytical results, and therefore it is suggested that the by-harvests of all web domains are identified, evaluated, and either included in or excluded from the corpus. If this identification and inclusion/exclusion is not made we consider the corpus ‘unfiltered’, and we consider it ‘filtered’ if this process has been undertaken.

As mentioned above the intention was to investigate two approaches to establishing a corpus:

- an unfiltered crawl.log corpus
- a filtered full-text index corpus

The overall relation between the approaches shall briefly be outlined. All the used data sources — crawl.log and full-text index — have undergone an initial processing made by the archive with a view to enabling the processing of the data, including different forms of data cleaning, and in addition the creation of the full-text index is in itself a form of processing where ARC/WARC files are indexed by the use of dedicated software. In addition to this initial processing, initiated by the web archive, the data relevant for the creation of the corpora went through an additional cleaning process as a preparation for the studies. Following this the corpora were either filtered or not, that is: by-harvests and main harvests were identified and by-harvests should either be included or not. In contrast to the archive's processing the filtering was initiated by a researcher need to have as 'clean' a corpus as possible (cf. above).

An unfiltered crawl.log corpus

In 2015 the national Danish web archive Netarkivet celebrated its 10th anniversary, and in relation to this event a keynote talk about the development of Netarkivet was given at a conference (the first RESAW conference in Aarhus, June 2015). This keynote was later expanded to become a book chapter (Laursen & Møldrup-Dalum, 2017).

Description of the establishing of the corpus:

1. Aim: to map the development of Netarkivet from 2005 to 2015.
2. Based on: the first broad crawl from each year, each year including what Netarkivet terms 'step-1' and 'step-2', that is all websites were harvested to a certain limit ('step 1'), and then websites larger than the limit in step 1 were harvested again with a higher limit ('step 2'); in addition, the special harvests of 'Very big sites' as well as 'Ministries and Departments' closest in time to the step-1 and step-2 harvests.
3. Established by: analysis of crawl.logs; by using simple unix commands, java programs, descriptive statistical analysis in R, R for visualising and validity.
4. Result: a corpus file of 5,5GB and 50 million lines, including information about sizes, MIME types, and HTTP response codes.
5. The following will be used in this paper: sizes and MIME types (called file types). Only .dk domains are included in size.
6. Documented in: Møldrup-Dalum, 2015, Laursen & Møldrup-Dalum 2017.

A filtered full-text index corpus

In November 2015 the authors of the present paper should present a paper about the historical development of the Danish web at a conference (ECREA, Prague). As part of that paper some initial findings in the form of statistics about the Danish web were to be presented (the paper is documented in Nielsen, Brügger, & Laursen, 2016).

Description of the establishing of the corpus:

1. Aim: to provide statistics about the development of the Danish web, and to investigate to what extent it was possible to provide the needed statistics by querying a full-text index.
2. Based on: the same broad crawl of each year as the one used in the unprocessed broad crawl above (identical harvest IDs), each year including 'step-1', 'step-2'. However, a filter query was used to set a filter for only the .dk domain..
3. Established by: using the built-in statistics module of the index server to query all the domain names of a broad crawl to get the needed information; harvest and job IDs extracted from the database to a list of domains crawled, and domain names compiled in an annual csv-file, that was then sorted/converted/queried in full-text index by domain name/job ID; the results

were output as a JSON-file, imported into R, and exported to an Excel spreadsheet.

4. Result: a file including information about number of objects found (numFound), content types (content_type_norm), minimum/maximum size of objects in bytes (min and max), total size of objects (sum), and average size of objects (mean).
5. The following will be used in this paper: total size of objects, and number of objects by content type.
6. Documented in: Have, 2017.

3. Results

As the following comparisons will show, the results differ depending on the method used for creating the corpus. The differences between the results from using an unfiltered crawl.log corpus and an filtered full-text index corpus are obvious, and they attest to the challenges of creating a corpus from archived web materials.

In this section of the paper, we will describe the factors that we assume can help explain the differences in the results but it is important to stress that at this point in time we have not been able to verify that it is exactly these factors that cause the differences, nor what the relationships between the different factors are or how these may change over time.

What we do have are two different studies so we can compare the size of the Danish domain and the number of files types, respectively, in each study. The outcome suggests that some of the factors influencing the results are the same in the two studies while some are different due to the different methods applied.
We also have two different measurement points to compare the results and what we see is that some of the factors influencing the results in relation to the two measurement points are the same and some are different. But not only this, the same factors might influence the results in different ways in different corpora - and maybe even within the years that we are studying within one corpus. We will explain this in more detail following the figures showing the results of the comparisons.

3.1. Domain sizes of the .dk domain

The first figure below shows the size of the Danish web as seen in the results from the broad crawl corpus and the full-text corpus. It is important to keep in mind that the first study is based on what was harvested, while the second was based on what is in the archive and has been indexed in the full-text index but other factors need to be included in order to try to describe the differences.

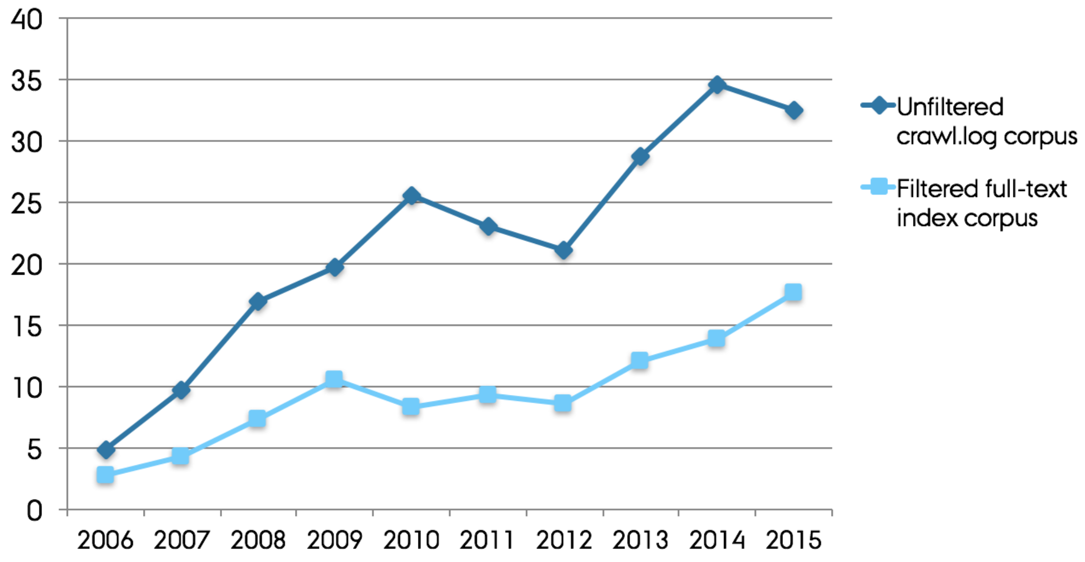

Figure 2. Size of the Danish web domain in terabytes.

- Both crawllog and index are ascending
  - possible explanation is the growth of the Internet
- Crawllog is higher than index
  - possibly due to dedublication, exclusion of non -dk, exclusion of by-harvest, index errors, exclusion of redirects, exclusion of reponse codes (only 200-299 included)
- Crawllog is more than twice the height
  - no known explanation
- Crawllog is ascending two times
  - one possibly due to introduction of global filters (cf. Laursen & Møldrup-Dalum 2017)

As is obvious from the chart, the Danish web grows larger over time.
appears to be much larger in the broad crawl corpus than in the full-text corpus. We can see that if we use the crawl.log as basis for our study our results show a Danish web that is approximately double the size of the Danish web as seen in the results using the full-text index. However, this is not surprising as we did in fact expect that there would be some important differences due to the way the corpus is created.

The reason why the Danish web is much smaller in the full-text corpus (only half the size of the Danish web in the broad crawl corpus) is probably mainly because the full-text index does not include more than one version of the files that have been deduplicated. Moreover, the full-text index corpus includes only .dk domains, while the broad crawl corpus includes other domains such as .net, .com, .org, which are considered part of the Danish web. Other reasons are exclusion of by-harvest, index errors, exclusion of redirects, exclusion of reponse codes (only 200-299 included).

Another point worth mentioning is that looking at the chart, the drop from 2010 to 2012 in results from the broad crawl corpus stands out, especially since this is not mirrored in the results from the full-text index, at least not in 2011. The decrease can probably be explained, at least partly, by the introduction of global filters to avoid crawler traps (as described in Laursen & Møldrup-Dalum, 2017). However, this assumption has not yet been tested.

Thus, the differences are (among other possible reasons) a result of how the full-text index handles some of the technical and curatorial choices that have been made by the technicians and curators of Netarkivet.

3.2. Number of file types on the .dk domain

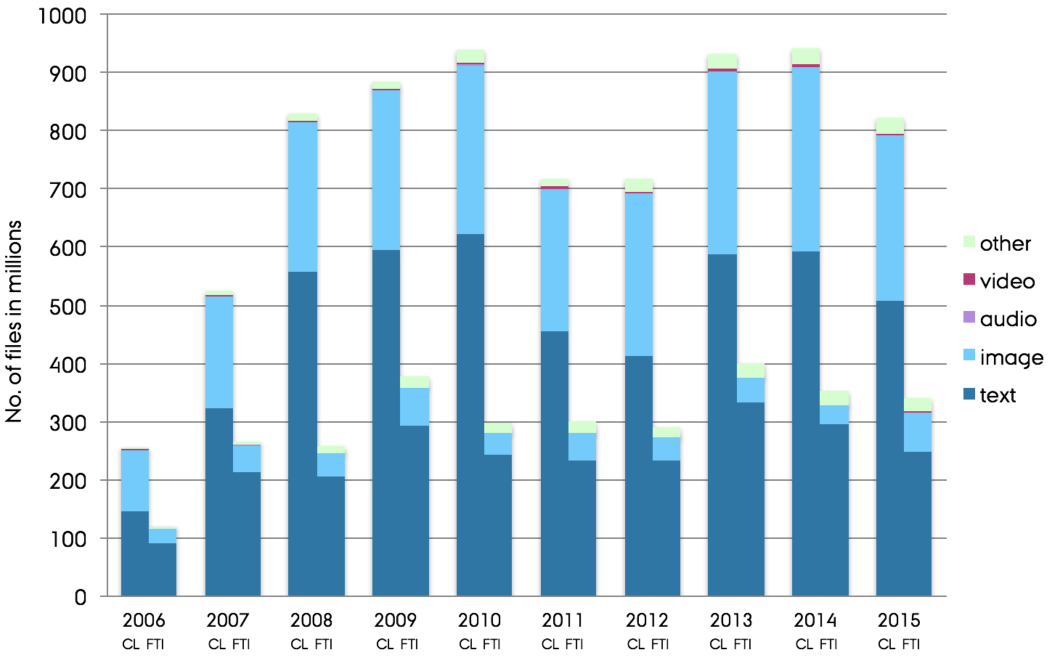

Figure 3. File types compared (number of files in millions)

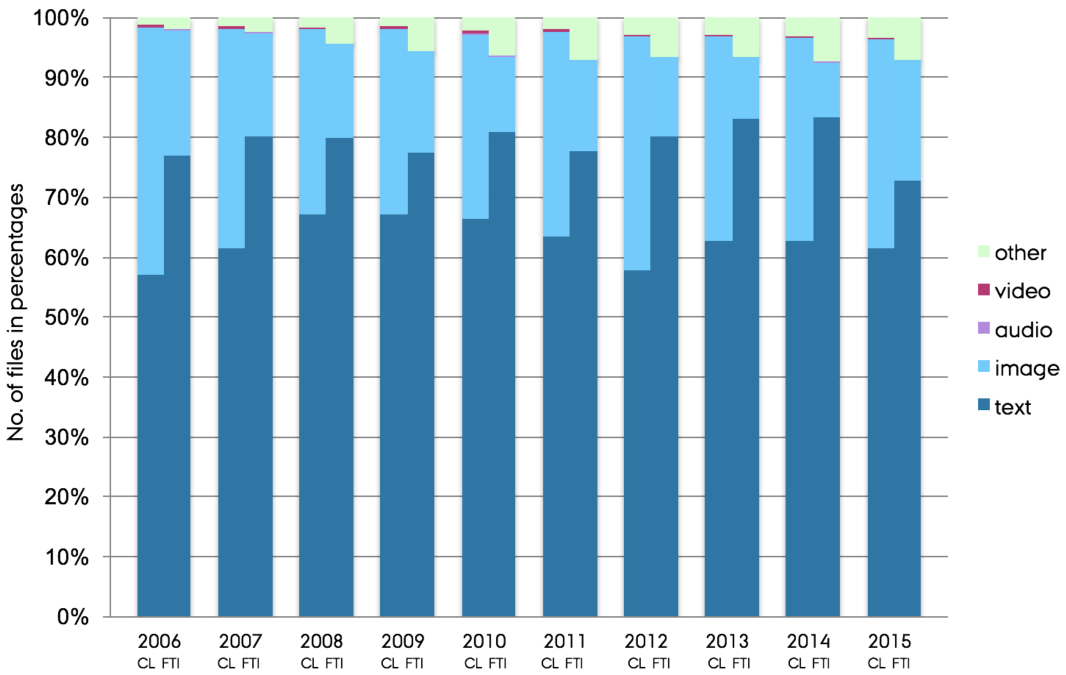

Figure 4. File types compared (number of files in percentages)

- Both crawllog and index are ascending
  - possible explanation is the growth of the Internet
- Crawllog is higher than index
  - possibly due to dedublication, exclusion of non -dk, exclusion of by-harvest, index errors, exclusion of redirects, exclusion of reponse codes (only 200-299 included)
- Crawllog is more than twice the height
  - no known explanation
- Crawllog is ascending two times
  - one possibly due to introduction of global filters (cf. Laursen & Møldrup-Dalum 2017)
- More images
  - possible explanation is no deduplication of html
- Very little audio and video, possibly a bit more in crawllog
  - possible explanation is deduplication

4. Concluding remarks

Results are somewhat alike but differ depending on the method used for creating the corpus – the two methods offer different views into the development of the Danish web and can complement and inform each other.

Some of the factors influencing the results are the same in the two methods, ie. cleaning, filtering, selection, normalizing – however, these factors might influence the results in different ways in different corpora depending on analytical and technological approaches.

Need for insight and documentation not only in how the archive has been created, but also in how different views (ie. crawllog view, index view) offer different biases.

5. Next steps

There is an inbuilt bias in the archive where some domains or some parts of domains are harvested more than once; you get more versions of 'the same' from different points in time. Therefore we would like to explore results from a filtered crawl.log corpus (the third method below). The filtering will ensure that only the domain material from the main harvest including , potentially, material from the by-harvest is included.

Data source and filtering Corpus	Crawl.log	Full-text index (ARC/WARC)	Filtered
Unfiltered crawl.log corpus	✓		
Filtered full-text index corpus		✓	(✓)
Filtered crawl.log corpus	✓		✓

Table 1: Data sources and filterings of the three corpora.

6. References

Brügger, N. (2017). Probing a nation's web domain: A new approach to web history and a new kind of historical source. In G. Goggin & M. McLelland (Eds.), *The Routledge Companion to Global Internet Histories*, pp. 61-73. New York/Abingdon: Routledge.

Brügger, N., Laursen, D., Nielsen, J. (2015). Studying a nation's websphere over time: Analytical and methodological considerations. Paper presented at The International Internet Preservation Consortium (IIPC) General Assembly 2015, 27 April, Palo Alto.

Have, U.K. (2017). *Report: Probing the Danish Web using the full-text index*, 2. version. Aarhus: NetLab. (unpublished)

Laursen, D., & Møldrup-Dalum, P. (2017, forthcoming). Looking back, looking forward: 10 years of development to collect, preserve, and access the Danish web.

In N. Brügger (Ed.), *Web 25: Histories from the first 25 years of the World Wide Web*, pp. 207-226. New York: Peter Lang Publishing.

Møldrup-Dalum, P. (2015). *Report on data mining the Netarkiv for its 10-years birthday*. Aarhus: Statsbiblioteket. (unpublished)

Nielsen, J., Brügger, N., Laursen, D. (2016). Mapping the past of a national web domain: The development of the Danish web 2005-2015. Paper presented at The European Communication Research and Education Association (ECREA), the 6th European Communication Conference, 12 November, Prague.

Acknowledgements

We would like to thank Netarkivet's and NetLab's IT developers Per Møldrup-Dalum, Toke Eskildsen, Thomas Egense, and Ulrich Karstoft Have for their invaluable help with this paper.