The DK-domain: in words and figures

by daily manager of netarchive.dk
Bjarne Andersen
State & University Library
Universitetsparken
DK-8000 Aarhus C
+45 89462165
bja@netarkivet.dk

On July 1, 2005, a new version of the legal deposit law came into effect in Denmark. It meant that the national libraries in Denmark – The Royal Library and The State and University Library – were given the duty and legal authority to collect and preserve the Danish part of the Internet.

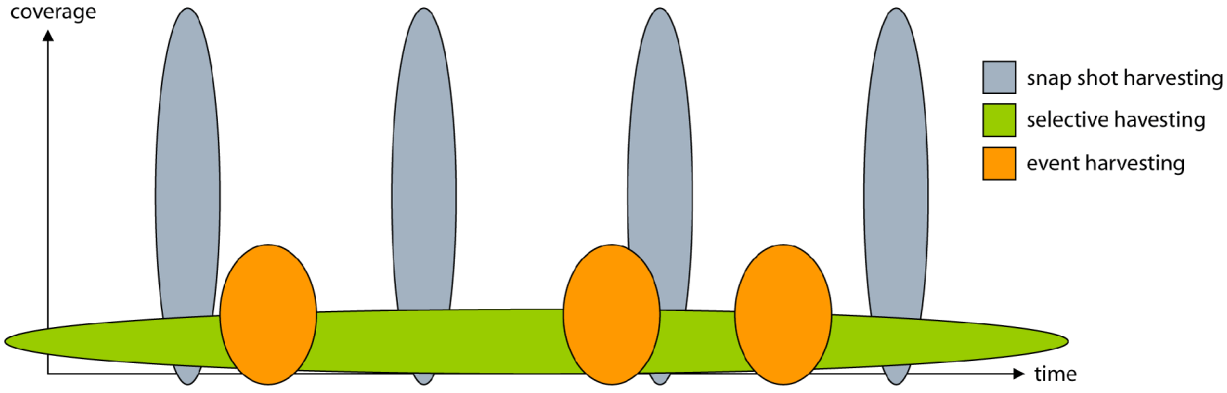
Collecting was initiated in July 2005, and is handled by a virtual center: netarkivet .dk (hereafter the Net Archive), which is run by the two libraries in close cooperation.

Collecting is implemented with traditional webcrawler software, which carries out automatic downloading – so-called "harvesting" – via a three-pronged strategy:

1. Cross-section harvesting of all relevant domains four times yearly.
2. Selective harvesting of approximately 80 domains with greater frequency (e.g. daily).
3. Event harvesting of two to three events annually.

Cross-section harvesting gives a broad view of the Danish part of the Internet. All known domains are downloaded in their entirety, with few restrictions. **Selective harvesting** covers certain websites with greater frequency – potentially as often as once an hour – gi8ving an uninterrupted picture of a small number of particularly important and dynamic websites, while **event-harvesting** combines the two other strategies and collects a larger amount of sites (on the order of 2-4000) more frequently, e.g., daily. In the period of October to November, the Net Archive carried out the first event harvesting in connection with the municipal elections.

The overall strategy can be illustrated in the following way:



This article describes the Net Archive's early experiences with cross-section harvesting, which was implemented in July to October of 2005. This cross-section harvesting covered only .dk-domains, which is why our experiences give a very good view of the DK-domain. This is, as far as we know, the first such comprehensive characterization of the Danish domains.

Conceptual premises

The cross-section harvesting was implemented according to the following premises:

1. Harvesting was initiated based on a complete list of domains from DK-Hostmaster. The list contained ca. 607,000 domain names.
2. At that point in time, the harvesters could not handle domain names which contained Danish characters (æ, ø and å), which is why the actual number of domains we attempted to harvest was ca. 579,000.
3. The harvesters did not respect the robots.txt standard (see below).
4. Collection was implemented with a maximum limit of 5000 objects per domain, to avoid overloading the visited web servers and to avoid so-called crawler traps (see below), among other reasons.

Technical Specifications

Harvesting was implemented using the following technical setup:

1. Using two machines, each with two CPUs and 4GByte of RAM. Each machine ran two copies of the webcrawler software.
2. With the open-source harvester Heritrix[1].
3. Via the national libraries' 100 Mbit net connection on the Danish Research Network.
4. With a bandwidth limit of max. 3MByte/s per harvester-machine and an upper limit of 500KByte/s per domain.
5. Took approximately three weeks' effective harvesting-time, apportioned over the period of July to October 2005. This constitutes an effective bandwidth usage of 2.9MByte/s, or roughly 25Mbit/s.
6. Downloaded 138,796,750 objects, which occupy ca. 5.3 TByte of storage space (5,300 GByte).

Limitations

As this was the first complete harvesting of the Danish domains, and as the Net Archive wished to proceed cautiously, it was decided that we would set an upper limit on the number of objects collected per domain. The limit was set at 5000 objects per domain, based on a hypothesis that this would secure the majority of Danish net sites in their entirety, and yet give a fleshed-out picture of the very large sites.

[1] Developed primarily by The Internet Archive (http://crawler.archive.org), but in cooperation with the Nordic National Libraries, among others, through the IIPC cooperation (http://www.netpreserve.org).

The limit was set for two reasons. Primarily, we didn't want to overload the Danish web servers any more than absolutely necessary. It was clear that collecting so many websites would generate much more traffic than normal, and naturally we wished to maintain a good relationship with the owners of the affected domains. The domain owners are, after all, our most important collaborators, even though they do not, in principle, carry out any of the work.

Harvesting with webcrawler software proceeds in following way: the software is seeded with a number of starting pages (in this case, the complete list of the .dk domain's front pages). These sites are downloaded, the software finds links and the process continues in this manner until no more pages remain on the domains that it started with, or until the limit of 5000 objects per domain is reached.

Secondly, the limit also acts as a practical guard against the so-called "crawler traps". Crawler traps are, as the name implies, places on the net where the webcrawler gets caught in the virtual world, that is to say, it ends up downloading conceivably infinitely many pages, if it isn't stopped one way or another. A typical and frequently found crawler trap is a calendar application, wherein one can surf around in a calendar with links to, for example, the next day, next month and next year. In such sites, the web crawler continues to find new links and download new pages completely without relevant content, as very few calendars contain entries centuries in the future (or past). To utilize an upper limit naturally causes much uncertainty in the derived statistics. This uncertainty particularly affects statistics about how large Danish websites are; that is to say, those websites that reached the 5000-object limit are actually larger than 5000 objects – we just don't know how much larger. However, as the statistics later in the article will show, the amount of sites which actually reached the limit was not so large, which is why this uncertainty realistically only encompasses a smaller number of sites.

The occurrence of crawler traps also inflates the statistics regarding the Danish websites somewhat, as crawler traps cause websites to look bigger than they actually are. We have no useful statistics about the frequency of crawler traps, primarily because they are nearly impossible to find automatically.

Robots.txt

The collection of the Danish part of the Internet does not respect the so-called robots.txt directives. Studies from 2003-2004 showed that many of the truly important net sites (e.g. news media, political parties) had very stringent robots.txt directives. Had these directives been followed, nothing at all would be archived from those sites. Therefore robots.txt is explicitly mentioned in the commentary to the law, not because robots.txt is a judicial standard (it is rather a sort of gentlemen's agreement), but precisely because the parties who were invited to the initial hearings (professional agents within the internet business) had the opportunity to comment critically on this approach.

The statistics for the first cross-section harvest showed that fully 35,000 websites had robots.txt directives, and thus had more or less strict rules for what web crawlers are allowed to download. The Net Archive does not have the resources to manually inspect the contents so many occurrences, so it seems natural, given the archive's purpose, to ignore the robots.txt standard.

The Net Archive has always been aware of the fact that robots.txt was invented to prevent web crawlers from making enquiries to URLs that could potentially create problems for the visited websites (e.g. to send a submission to a debate forum, to send an email to the webmaster, and so forth). We are always willing and able to find a quick and effective solution if it appears that the Net Archive's web crawlers have behaved inconsiderately, and we have in fact had a small number of cases of that nature. An obvious solution is to obey the robots.txt directives on those sites where it would create problems to ignore them, and this model has in fact been used in a few instances. A less confining solution is that the Net Archive can add rules to its system that certain URLs (or URL-syntaxes) may not be downloaded.

A few enquirers have also suggested using the robots.txt directives to omit materials which the producers find irrelevant (e.g. private photos, etc). It is for neither the Net Archive nor other people to define what is relevant or not in 2005. Researchers of the future may well uncover interesting things on the basis of material which, at a glance, may be appraised as uninteresting in our time, but which may become important sources of information in ten or fifty years.

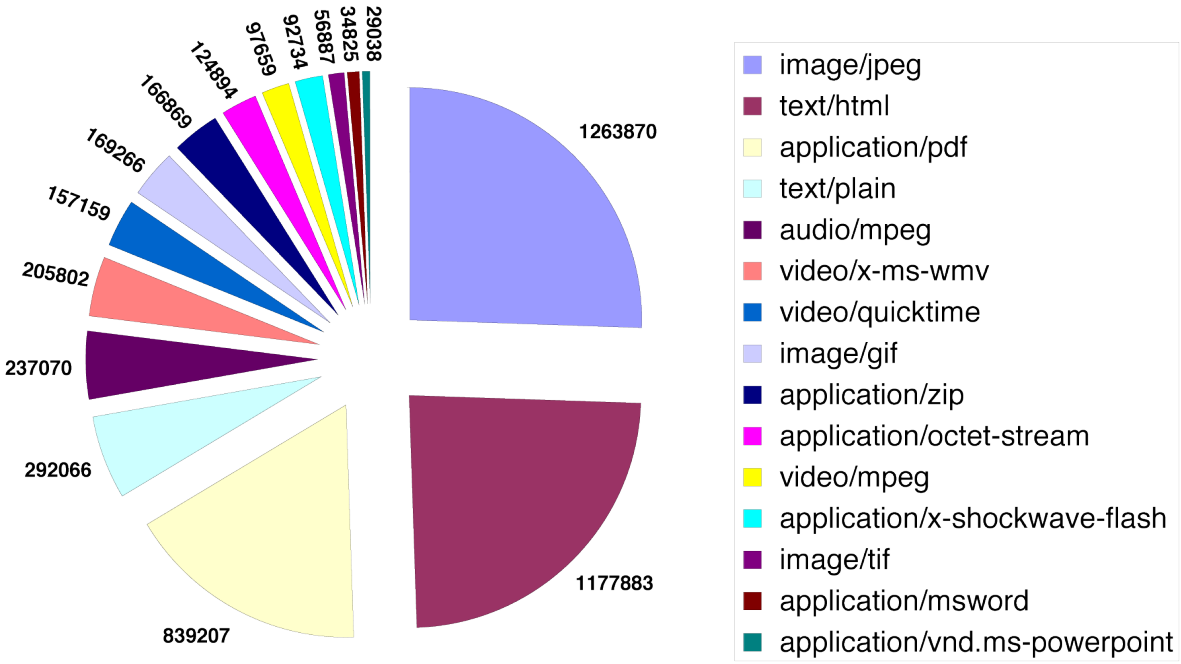
Since the Net Archive's purpose is to preserve the Danish part of the Internet, in principle perpetually, we can't be limited by what people today think is relevant or irrelevant, which is also the primary reason that we do cross-section harvesting at all.

In some other countries (Australia, to name one example), they have until just recently only collected websites using a selective strategy. However, they have now realized (and received the corresponding legal directive) that we cannot say today what will be interesting in the future, which is why the safest course is to try to preserve everything. As the price of storage space is falling steadily, this practice is now a feasible assignment, as compared to just 5-10 years ago, where quantities of data of this size would have to be stored on DAT tapes in order to keep Internet archiving projects on a reasonable budget.

File types

The collection of the entire Danish domain yields many interesting statistics. The following statistic shows the breakdown of file types for the collection. One of the things it shows is which formats are most popular – for example, whether Danes prefer Word .doc files or Acrobat .pdf files for publishing documents on the Internet.

The following chart shows the amount of data in megabytes for the 15 file types that occupy the most space in the collection.



Most surprising is the fact that the JPEG graphic format is the top scorer. We interpret this as a sign that digital cameras have become quite common, and that Danes are more and more frequently placing their private photo albums on the public portion of the Internet. Test samples have shown that we have downloaded private photo albums with thousands of photos from a number of websites, many of them with pictures in very high resolution.

Another smaller surprise has been that PDF files occupy third place in the number of megabytes. This is a clear sign that, in 2005, PDF files are the vastly preferred format for publishing documents online (aside from documents in HTML-format). It will be interesting to follow these statistics in coming years.

Unsurprisingly, AV materials (audio/* and video/*) take up a great deal of space in the chart, as files in these formats tend to be quite large.

All in all, these 15 file types cover more than 95% of the collected amount of data. Given that the collected list contained 613 unique MIME types (of which a goodly number are not officially registered MIME types), this is quite a large fraction. Long-term storage of files in the Net Archive probably won't be able to keep all file types readable in perpetuity, but this statistic shows that we can preserve a great portion of the collected archive if we can "just" preserve a smaller number of different file types. Many of the file types on the top-15 list present a major challenge when it comes to preservation, but that problem is beyond the scope of this article.

The statistic over the fifteen file types that occupy the most space in megabytes is shown in figures below. The table also contains information about the number of objects collected of the relevant types, along with the average file size for the file types.

MIME type	Mbytes	% af all	amount	% af all	file size (Kb)
image/jpeg	1263870	25,56%	36914322	28,33%	35
text/html	1177883	23,82%	68634852	52,68%	18
application/pdf	839207	16,97%	1559645	1,20%	551
text/plain	292066	5,91%	841932	0,65%	355
audio/mpeg	237070	4,79%	85287	0,07%	2846
video/x-ms-wmv	205802	4,16%	41164	0,03%	5120
video/quicktime	157159	3,18%	26927	0,02%	5977
image/gif	169266	3,42%	21013245	16,13%	8
application/zip	166869	3,37%	79093	0,06%	2160
application/octet-stream	124894	2,53%	229051	0,18%	558
video/mpeg	97659	1,97%	20005	0,02%	4999
application/x-shockwave-flash	92734	1,88%	632402	0,49%	150
image/tif	56887	1,15%	11552	0,01%	5043
application/msword	34825	0,70%	190267	0,15%	187
application/vnd.ms-powerpoint	29038	0,59%	16800	0,01%	1770

It is not surprising that HTML files are by far the most widespread (over 50%) when measured in number of files. After HTML files come the two most widespread picture formats, JPEG and GIF. Altogether these three file types constitute more than 97% of the number of objects, though only 52% when measured in megabytes.

It is also clear that PDF is vastly more widespread than MS-Word documents, inasmuch as that there are more than eight times as many PDF files as Word files.

The publication of video clips on the Internet also appears to be an increasingly common practice, as shown by the fact that the first cross-section harvesting collected more than 88,000 video clips.

The average file size for all collected files is approximately 40KB. This is a marked increase compared to an investigation the two libraries made in 2001. At that point, the average file size was ca. 34 KB. This 17% increase is testament to the steadily increasing amount of bandwidth in Danish homes and greater space available from web hosting providers around the country.

The size of Danish websites

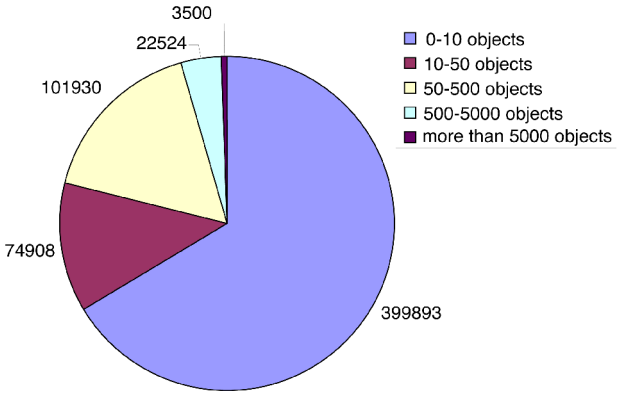
The first cross-section harvesting was only initiated on about 579,000 domains, due to problems with domain names that contained æ, ø and å. The statistics could therefore be somewhat distorted, in case the domains that could not be harvested didn't have a size breakdown comparable to that of other sites.

The number of registered domains is steadily increasing. There were 649,510 domain names registered in October, as compared to the 607,000 that were on the list with which the harvesting started in June 2005[2]. The number of domain names containing Danish characters was largely unchanged in the same period, from 28,250 to 28,426[3]. The Danes can clearly still find word and letter combinations that are not yet registered with DK-Hostmaster.

[2] http://www.dk-hostmaster.dk/index.php?id=235

[3] http://www.dk-hostmaster.dk/index.php?id=116

This is how the collection looks, measured by number of objects per domain.



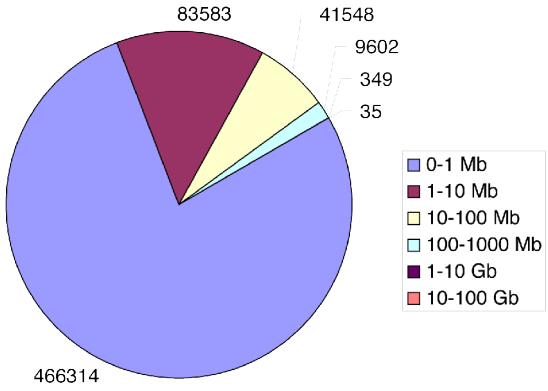
The chart in figure 3 shows that nearly 70% of all domain names are for websites which contain between zero and ten objects – in practice just a "front page" with perhaps a couple of pictures.

Only 4.4% of domain names (26,024) contain more than 500 objects, and only 0.6% (3,500) contain more than 5000 objects. This last number is somewhat uncertain due to the previously described limitation. Thus an average Danish website contains approximately 275 objects.

It appears that the number of registered domains is in fact significantly higher than the actual number of web sites. A large number of the registered domain names either do not answer on DNS-lookup, or the web server that has been given the domain name does not answer. Such was the case with a little more than 100,000 of the domain names, or fully 17% of all registered names. In actuality, we only archived material from 479,000 "living" Danish domains.

This number does not correspond to the number of "unique" websites, in that many content providers (primarily commercial ones) have registered multiple domain names for the same site – so-called "aliases". The first cross-section harvesting did not take this problem into account, but simply downloaded all the "living" sites from the collected list. At the moment we are working on features that will assist with the automatic identification of these aliases. We can thereby avoid archiving entirely identical websites under different names, partly to save space in the archive, but also to dramatically reduce the load on the affected web servers.

The Danish web sites break down in this way according to size:



There is again some uncertainty on the largest web sites, but the chart shows quite clearly that

the vast majority of sites contain less than 100 MByte of data (98,2 %), which means that just under 10,000 Danish sites contain more than 100 MByte. An average Danish website thus contains almost 12 MByte.

The statistics can also show something about how widespread the usage of different hostnames is in the Danish domain (for example sporten.tv2.dk, nyhederne.tv2.dk, and politik.tv2.dk). The vast majority of the domains use only www.domainname.dk. Only 34,036 (7.3%) of the visited domains have more than one host-name defined (and visited by the Net Archive's harvester). On average, these domains have 5.14 host-names. The majority have exactly two host-names (most use both www.domainname.dk and domainname.dk).

Of these domains, 5,829 have more than two host-names. In this group, the domains have on average more than 21 different host-names defined, which indicates that, once a domain has begun using more than two host-names, the likelihood of its using even more increases dramatically. The average number is inflated by a small number of domains with many host-names. First place goes to a site with more than 20,000 host-names for the same domain. In this instance, words from a dictionary-like application were used, which is why the number became unusually large.

External websites

The web crawler software is set up in such a way that it only finds links and goes further on those domains which were included in the starting list. However, "external" files necessary for the recreation of sites (e.g., pictures, etc.) are also included. The result of this has been that the Net Archive's harvesters have downloaded materials from a total of 155,208 unique domains outside the .dk-domain. We can quickly conclude that it is very common to link to, for example, pictures outside one's own website.

The number of objects that were downloaded from servers outside the .dk-domain are however "just" 7,934,537, which corresponds to just about 6% of the collected objects. We downloaded on average 51 objects per domain from external servers.

Conclusion

The first cross section harvesting of the collected .dk-domain has shown a number of interesting things about the Danish domains' size and content.

1. In the period of July to October 2005, there were about 479,000 "live" domains in Denmark.
2. An average website contained 275 objects and occupied almost 12 Mbyte.
3. 98.2% of all domains contained fewer than 100 MByte of data.
4. JPEG graphics are the file type which altogether occupies the most space.
5. HTML files are the file type of which there are the most.
6. PDF files are by far and away the most utilized for publishing documents which are not in HTML format.
7. The average file size increases in concert with bandwidth/server space.
8. The use of multiple host names is relatively widespread, and once more than two

host names are being used, the probability that more will follow increases greatly.

The experiences from this harvest have shown that both aliases and crawler traps are a real problem. The Net Archive is therefore working on finding methods to automate their identification. Because file sizes can vary greatly from one website to another, the upper limit on the number of objects has proven to be inexpedient (i.e., a website with 5000 video clips can occupy a great deal of space). Therefore, we now operate with an upper limit on the number of bytes instead.

At the same time, the Net Archive is examining possibilities for reducing duplication, as tests have shown that at least 50% of the material is static. For this reason, subsequent cross-section harvestings will download and archive large amounts of redundant data. Duplicate reduction can be executed either after download, which will not reduce the load on the content providers and the network, or during download, by not re-downloading static data, e.g., JPEG, GIF, and PDF files which haven't changed since the previous time.

The plan is to generate statistics after every cross-section harvesting, so that there will eventually be a good basis for comparison. Many of the statistics are interesting to follow, for example increases or decreases in the utilization and distribution of certain file formats, or the average size of both single objects and websites as a whole.

For further information, visit http://netarchive.dk.