netarchive.dk

Harvesting the Danish internet – the first two years

On July 1st 2005 the most recent Danish legal deposit law went into force. Chapter 3 of the law concerns "Materials published in electronic communications network" and allows for the harvesting of websites within the domain <.dk> and of websites on other domains aimed at a Danish audience. Since that date the Royal Library (The National Library of Denmark) and The State and University Library have harvested more that 34 terabytes from the Internet. This article will discuss the experiences, which the libraries have gained during the first two years.

Strategies for harvesting

Before the law was passed and went into effect, experiments had shown that in order to capture the Internet three types of strategies had to be employed: a general snapshot of the entire <.dk> domain done four times a year, a selective but more frequent harvests of dynamic sites, and irregular harvests of selected websites in connections with events.

Snapshot harvest

The idea behind the snapshots is to get a complete picture of the Danish domain by harvesting all web pages four times a year. So far we have been able to complete three harvests, one in 2005, two in 2006, while a fourth harvest was initiated in March 2007. We are now able to take a complete snapshot within three months with the hardware capacity available. Thus from 2008, we should be able to do four snapshots a year. It is still an open question if and how much we can reduce this time span for a complete snapshot harvest, as we also have to take into consideration the traffic on the websites we harvest. We don't want to close any site down while harvesting, so for smaller websites especially we try to collect in a speed that will not obstruct normal Internet traffic.

After the first two snapshot harvests (Fall 2005 and Spring 2006), we concentrated on getting a de-duplication module integrated into our system. That task was successfully completed over the summer of 2006. It has resulted in a significant reduction (31%) on bytes to be archived from snapshot harvests. This percentage would be larger when we reach the goal of doing four snapshots a year, thereby increasing the frequency of duplicates.

Before a harvest can begin, it is necessary for the system to have a list of domains to be harvested. This list is supplied by the host master of domain <.dk>, who according to the law is obliged to provide the netarchive with a list of all domains registered within the <.dk> top level domain. This has worked very well, and we have been able to feed an updated list into the system before each harvest. The first harvest was done in stages in order to uncover crawler traps (such as calendars). During the first stage all websites were collected up to 10 objects and we found that about 235.000 websites fell within this limit. During the next stage the limit was set at 50 objects pr. site, then at 500 objects and finally at 5000 objects. During these harvests more than 99% of the Danish domains had been collected in their entirety. The websites containing more than 5000 objects,

which amounted to 3.500 sites, were manually examined by two staff members for crawler traps, alias's and other problems, which were then eliminated (URLs removed from the harvester) and the websites collected in entirety.

After the first snapshot harvest we decided to change the measure of limits from objects to Mbytes, so that the first round of harvest had a limit of 10 Mbytes. More that 600.000 sites fell below this limit and about 61.000 hit the limit. The second round was set at 500 Mbytes, which left about 6.300 sites to be examined manually. This was done to eliminate the problems mentioned above but also to ascertain if certain sites, typically private sites with pictures and videos of family gatherings, weddings etc. could be sampled rather than collected in entirety. We decided for the time being to limit those sites to 499 MB. This reduced the number of sites to be harvested up to 1 Gbytes to 263 websites. The largest website by far is the national broadcasting corporation with more than 150 GB (www.dr.dk).

The fourth snapshot has been started without running in steps, but this has in practice proved to be a bad idea (thus proving that harvesting in steps - first up to 10MB, then with only domains hitting that 10MB limit - is a good idea) because small and big websites get too mixed up in the same harvester-jobs making harvesters ineffective. So from the next snapshot we will return to the original strategy with covering all the small sites in a first run.

Selective harvest

The idea behind the selective harvest is to gather web pages that are frequently updated and which would be missed by the snapshot harvest.

Such types has been defined as

- News sites (national and regional media)
- "Typical" dynamic and heavily used sites representing civic society, the commercial sector and public authorities.
- Experimental and/or unique sites, documenting new ways of using the web (e.g. net art)

The current policy is to collect about 80 such sites. To select only 80 among all Danish sites is not an easy task, but this number has been determined by the resources available. News sites are the easiest category to select, amounting at the moment to some 30-35 sites (Denmark is a small country). But the media market is constantly changing, and the development must be continuously monitored. The other two categories are more difficult to select. It is well known that the web has fostered new ways of human communication (virtual communities of all kinds), and this is important to document. An Editorial Advisory Board representing researchers and media professionals assists the netarchive in the selection process (see below).

Once a site has been selected, it will be examined thoroughly to find out, which parts of site should be harvested and how often it should be harvested. The frequency may vary from several times a day to once a week. Not surprisingly, this type of harvesting requires close monitoring from the

staff to find the sites and to determine frequency as well as depth of harvest (news sites, e.g., may have databases with older news and articles, that do not change and will be collected through snapshot harvesting, therefore only the "front pages" of the website need to be harvested frequently). Duplication reduction has really had an impact on this type of harvests, showing a saving of 50-70% on bits to be archived, which also means that we can harvest deeper than absolutely necessary just to be on the safe side.

Event harvest

The idea behind this strategy is to be able to collect web pages from new sites, dedicated to one event and which will disappear as the event is over.

We have defined an event as something that

- Creates a debate among the population and is expected to be of importance to Danish history or have an impact on the development of Danish society
- Causes the appearance of new websites devoted to the event
- Is dealt with extensively on existing websites

The first event, we harvested, was the local election in November of 2005. Apart from being an election, which is always of interests to scholars, this one was of special interests because of the government's decision to reorganise the country's local administration, reducing the number of municipalities (*kommuner)* from 271 to 98, and replacing 14 counties (*amter)* by 5 regions. This also means fewer mayoral and councillor seats for local politicians as well as new political constellations of voters in a municipality. Compared with a previous election harvested as a pilot project in 2001, the 2005 election showed a great increase in websites for individual candidates. More than 1,000 candidates had their own website, most of which disappeared after the election and this pattern is expected to become the norm in future elections.

The second event, we chose to harvest, was the crisis during the winter 2005/6 arising from the publication of caricatures of Muhammad in a Danish newspaper.

The third event was a follow-up on the elections of 2005. The new local structure was inaugurated on January 1st, 2007, and during the last months of 2006 we did a special harvest of all municipal and county websites as we anticipated – rightly – that websites for old municipalities would disappear as would websites set up to inform citizens of the changes in local and regional administration.

Concurrently we are collecting smaller events that for a brief period cause new websites to appear (and disappear) between snapshots. One example: the clearing on March 1st 2007 by the police of a house in Copenhagen occupied by a group of youngsters gave rise to a series of web pages dealing with this event.

It should be stressed that all harvested material will merge into one archive, regardless of how it was harvested. The above strategies are strategies for collecting the materials, not for building collections. We, do, however, keep a log on our harvesting activities as documentation for future users of how the content of the archive was collected.

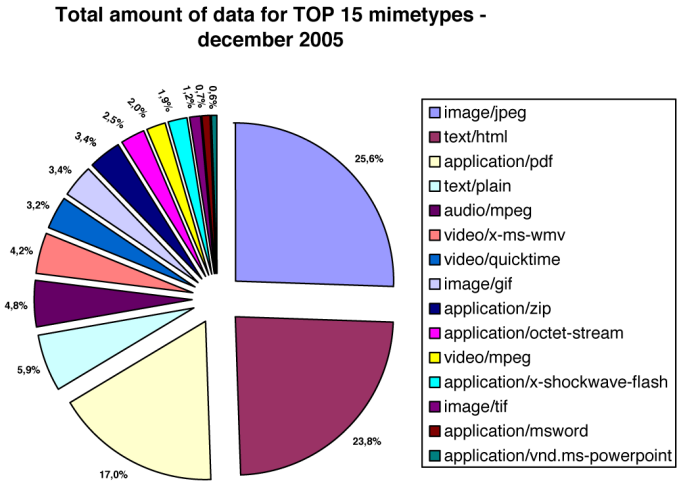
The materials harvested – statistics and types

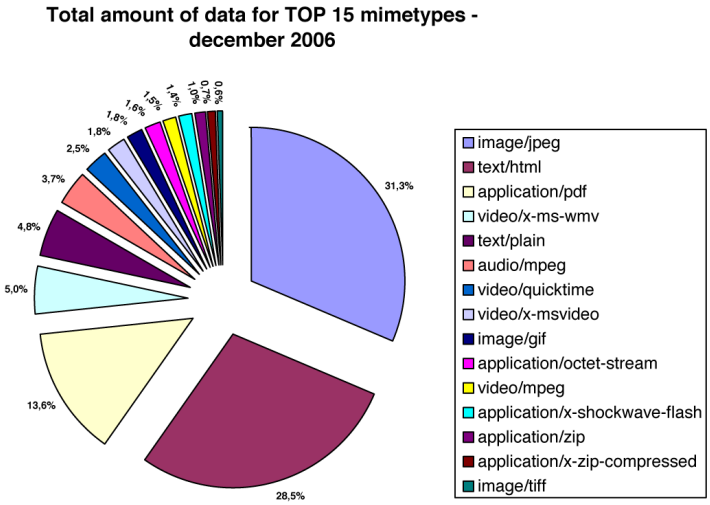
Domains

The first list of Danish domains from June 2005 contained about 607.000 domain names of which about 480.000 were active within the <.dk> domain, while the most recent list from March 2007 contained about 803.000 names of which about 640.000 are active. Most domains (more than 85% of the active domains) remain small in size, that is, less than 10 MB. The largest websites belong to the two national broadcasting corporations, the public DR and the semi-private TV2, which together holds more than 200 GB – and if including streamed material several TB. As of March 1, 2007, the archive holds 28,350 GB collected through the snapshot harvests; 3,593 GB from selective harvesting and 3,316 GB gathered through event harvests.

Types of documents

The latest snapshot harvest showed that pictures (jpeg-files) are still the most common type of file, now making up more than 31% of the total data collected (as opposed to 25% in 2005). The second largest is - less surprising – text/html files (28% in 2007, 23% in 2005). The third type is PDF, which share of the total has diminished (now 13,6% vs. 17 % in 2005). Video/x-ms-wmv has replaced text/plain as number four and video in general is now represented with four types of files in the 15 most common files types as opposed to three types in 2005. What this means in less technical terms is that there is a widespread use among the Danish population of the Internet as storage for private photos and home videos.





Problems encountered

Questions/complaints

We have received close to 100 enquiries and complaints. Considering the number of active websites we have visited (more than 650.000) this must be considered a very low number. It should be noted that we always leave our "card" identifying our harvester and containing an e-mail address that is checked daily. In the beginning most of the questions concerned our not respecting robot.txt convention, which we do not, as we want to harvest all public pages. Practically all complainers have been satisfied when being told about the purpose of legal deposit. At present we have encountered only three website owners who continue to be dissatisfied with our harvesting.

Crawler traps

As our harvester crawls the web through links it invariably falls into crawler traps such as calendars, which have an unending number of links. A lot of these are discovered and marked either at the domain on which they were found or at a global level avoiding them on other web sites holding the same trap.

Passwords

Sites, which require password, still present a problem at least as far as automatic gathering of login-information is concerned (e.g. automatic e-mail notification upon login-discovery would spam a lot of non-public websites that are not at all covered by the law and thus should not be archived). The harvester can technically do login while harvesting but collecting login-information is still a manual process.

The administrative set-up

Administering the Danish legal deposit law is shared jointly between the two legal deposit libraries, the Royal Library and the State and University Library. In order to secure an effective administration of the section dealing with "Materials published in electronic communications network," a virtual institution "netarkivet.dk" (netarchive.dk) was established. It is governed by a Steering Committee of 6 members (3 from each library) representing expertise in web, IT, legal deposit and collection building. The committee meets 2-3 times a year as needed to discuss and decide on economic, technical, policy and legal issues. Netarkivet.dk has a daily manager who reports to the Steering Committee and supervises the daily work of the netarchive.

To assist the netarchive in formulating policies for collection building, the Ministry of Culture has appointed an Editorial Advisory Board with five members from the university, IT and publishing sector, all present and future users and cooperative partners in establishing a national netarchive. This board convened for the first time in March 2006 and has since met twice with representatives of the netarkivet.dk. The Board has proven very helpful in discussing and formulating policies on collecting, defining "events" and assisting in identifying sites to be harvested selectively.

The policies

Together with the Editorial Advisory Board we have formulated policies on event harvesting (se above) and on domains outside <.dk> to be collected.

Domains outside <.dk>

As the law also covers Internet material that "is published from other Internet domains etc. and is directed at a public in Denmark", we also collect sites from outside domain <.dk> which are considered "directed at a public in Denmark" primarily sites in Danish but also sites of Danish companies, artists and institutions that are domiciled in Denmark. We have found about 38.000 sites, covered by the law, most of them at <.com>. The methods of locating these have been partially automatic, partially manual. The automatic method involved running a list of links on <.dk> websites pointing outside the domain (and therefore not harvested) against a GEO-IP presumably revealing websites with Danish owners. The manual method meant searching via Google for web pages which include Danish place names but which are not on domain <.dk>, which uncovered many sites for local associations and institutions.

Access

Access is, unfortunately, extremely limited for the time being. This is due, not to copyright legislation as might be expected, but to the Act on Processing of Personal Data. The Danish Data Protection Agency has decided that the data collected through the harvest may contain sensitive personal data and consequently access to the netarchive should be very restrictive. This means in practical terms that only researchers engaged in a research project on a post-doc level is allowed access. This, of course, goes against the philosophy behind the legal deposit act, namely that materials once published are kept available to the general public at the National Library. The Ministry of Culture is, therefore, about to negotiate with the Data Protection Agency to allow for wider access.

Staff and Equipment

The netarchive has at its disposal 4.5 full-time staff years, distributed among some 20 employees, IT engineers, computer scientists, librarians and library assistants.

The complete technical setup consists currently of 9 servers running linux (administration, harvesting and access) and a PC-farm of 25 machines running windows XP (for indexing and bit preservation). On our website: http://netarchive.dk you can find more several technical articles describing the system on both an overall level and in some very technical and specific details.

More questions?

See www.netarchive.dk or contact Grethe Jacobsen, gja@kb.dk or Eva Fønss-Jørgensen, efj@statsbiblioteket.dk (law and legal details) or Bjarne Andersen bja@netarkivet.dk (technical details)

Author: Grethe Jacobsen with assistance from Eva-Fønss-Jørgensen and Bjarne Andersen. Completed: May 2nd, 2007