netarchive.dk

Newsletter – March 2007

Welcome to the second international newsletter from the netarchive.dk project. It will present the status of the project on a collecting as well as a technical level and also mention more general news from the project.

Collecting policies

Recently the steering committee and the advisory board of netarchive.dk discussed and revised the following two policies on how to decide when an event is relevant for harvesting and how to define the necessary criteria when harvesting websites outside of the .dk TLD.

In connection with netarchive.dk an event can be characterized as follows:

- The event creates debate in the Danish population and can be said to influence Danish history or have a changing effect on its future development
- The event creates new websites
- The event gets serious coverage on already existing websites

Websites outside the .dk TLD should be harvested if they fulfil one or more of the following criteria:

- Language: at least 50% of the text is written in Danish
- The owner of the website lives in Denmark
- The owner is of Danish origin and the website is aimed at a Danish audience
- The website is aimed at the Danish population in general
- The website is aimed at inhabitants in Denmark whose native language is not Danish

The complete guidelines will be translated into English and published on our website: http://netarchive.dk

Snapshot harvesting

The netarchive.dk project has produced three snapshots of the .dk domain so far and a fourth harvest is currently running. Due to web server overload, harvesting big websites which pay their web hotel according to the amount of traffic their website gets, and crawler trap issues, we have decided on a default upper limit per domain of 500Mbytes. Domains which reach this limit will be manually processed in order to either increase the limit or keep it at 500Mbytes.

Approximately 7,000 domains have reached the limit so far. A lot of crawler traps were discovered in these and will be filtered off in future harvests. Due to an unexpected large growth in the overall data amount for these snap shots we have been forced to make guidelines to help us decide whether a domain should be granted a higher limit or not. These guidelines will be evaluated at a later stage.

So far websites that fall within one of the following categories will not be archived beyond the first 500Mbytes:

- The website is .dk, but the content is not Danish or aimed at a Danish audience
- The website is mainly of private content, typically sites with private photos, video clips, family, travels, hobbies etc.
- The website is a manufacturing business, but does not sell own products.
- The website consists mainly of pornographic content.

2,500 domains have been allowed to reach a limit of 1Gbyte and at the moment 250 domains have a 2Gbytes limit.

After each snapshot harvest all domains that have never before been processed but which reach the limit will be checked manually. We will of course at some point have to revise the list since websites tend to switch owners, contents and profiles.

Harvesting very large websites

Websites of a considerable size, such as http://dr.dk (Danish Broadcasting Corporation), are handled individually since they differ greatly from most of the other 800,000 Danish websites. They are harvested in separate jobs with different configurations (e.g. higher bandwidth per host) and much higher limits. The biggest website has a limit of 150Gbytes and when you reach that high a limit the need for filtering crawler traps gets obvious if you do not want to fill your archive with rubbish.

Deduplication

We are happy to announce that the Icelandic DeDuplicator module has been integrated with our system so that deduplication is now made automatically in all harvests.

The first results are very good. Performance is very high and harvesters still run at our default maximum bandwidth of 1,500kb/s per instance, meaning that one server with a dual CPU can still run two instances at full speed and the deduplication at the same time.

The initial results show that snapshots include around 30% duplicates, only deduplicating on non-“text/” mime types whereas the selective harvests include between 60 and 70% duplicates. This means that a lot of storage space and in the end a lot of money is saved when using deduplication.

Access-systems

The project has not yet implemented a user-friendly end user interface. This is mainly due to the very restricted legal framework. Only researchers at PhD level or higher can currently get access to the harvested material.

However, we granted and implemented remote access to the first real user just recently, Access was given on a very simple per harvest basis using our own proxy server and a per date harvest selector (basically a web page showing all harvests of a specific website with hypertext index-selection).

We plan on implementing advanced tools like nutchWAX, WayBack and WERA and we have done experiments as well as calculations on e.g. hardware requirements to build up an advanced end user access platform. The current amount of data (34.5 Tbytes) should be indexable and searchable for a small number of users on four machines holding 4Tbytes of index data.

The implementation of the access platform will hopefully begin in autumn 2007 and thus follow the open source release of our system

Danish domains outside the .dk TLD

During the last four months we have done quite a bit of work on discovering relevant Danish websites outside the .dk TLD. We have experimented with several different strategies:

1. Google search on Danish localities (like city names), limiting the search to non .dk sites
2. Extraction of websites redirected to directly from front page-URLs on .dk domains
3. Search for domain names in all previously harvested material.

The three strategies resulted in a total of more than 3,3 million unique domains. We ran them all through an IP geographically locator (GeoIP) to reveal hosts located in Denmark

This exercise reduced the number of domains to less than 47,000. The top 10 of the TLDs for that portion are as follows:

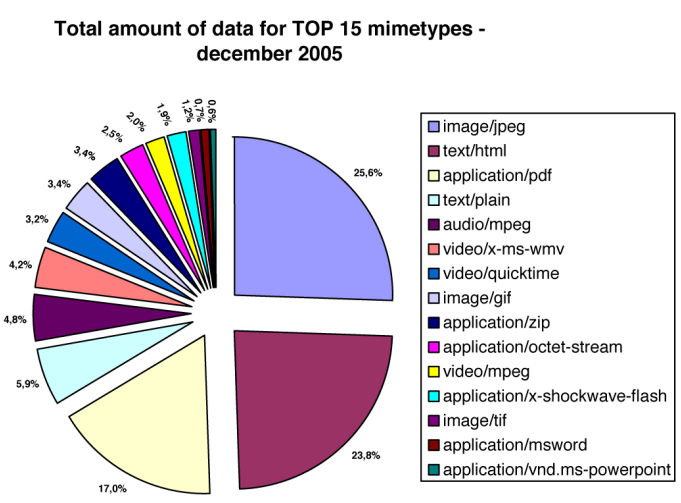
Top Level Domain	***Number of Domains***
.com	22,750
.se	8,058
.net	4,913
.org	2,381
.nu	1,473
.de	1,108
.no	1,106
.info	1,092
.eu	488
.biz	474

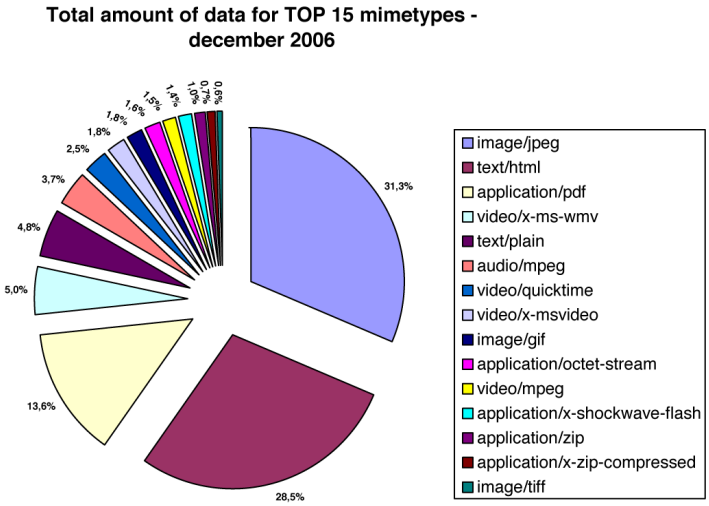
The GeoIP tool has proven to be quite accurate and it was decided to automatically include all domains placed by GeoIP in Denmark with the exception of national domains which are known to be harvested in snapshots harvesting jobs similar to our own (currently only .se and .no – Sweden and Norway)

So the currently running snapshot harvest of the Danish domain includes around 38,000 domains outside the .dk TLD. This corresponds to approx. 6% of the total number of active domains and we expect the amount of relevant websites to be between 8 and 10% (based on a simple spot test on links in some Danish printed magazines), so there are still some thousands of websites to discover.

Mime type development on the .dk domain

After the third snapshot harvest of the .dk TLD we have generated a lot of statistics. Among the more interesting information is the statistical figures which reveal how mime types are represented in the overall amount of data. TOP 15 mime types have evolved from 2005 to 2006 as follows:





The most significant change is that number of jpg-pictures has increased from covering approximately 25% of the total amount of data to more than 31%. Furthermore it is becoming clear that video formats are becoming more popular. Video is now represented with four mime types on the TOP 15 list compared to three in 2005 – so adding video material to websites is definitely a growing activity.

The netarchive.dk system in Open Source

The last piece of news is that we continue our work on the open source release of our entire system. Since the last newsletter we have done a lot of refactoring of the code to make it more modular and we are currently working hard on internationalization of the GUI as well as on writing documentation.

The system will be released under the LGLP open source license on July 1st 2007. Please check our website http://netarchive.dk for updates on the matter and feel free to write directly to the project's manager of the development part, Christen Hedegaard (chh@kb.dk), if you have any questions.

Best wishes
Bjarne Andersen
Daily Manager
bja@netarkivet.dk