netarchive.dk

Newsletter August 2008

ECDL September 14-19, 2008 in Århus – special workshop on web archiving September 18-19

We would like to draw your attention to the European Conference on Research and Advanced Technology for Digital Libraries **http://www.ecdl2008.org/**. The conference will include a workshop on web archiving, arranged by the IWAW (International Web Archiving Workshop) http://iwaw.net/08/index.html

The law

The legal foundation for our web archiving activities is the Act on Legal Deposit of Published Material of 22. December 2004, which was supposed to have been revised during the 2007/08-session of the Danish parliament (*Folketinget*). However, as we have not made much progress on the issue of access (see below), the revision was postponed to 2010.

Statistics

As of July 1st, 2008, the archive contained 71 Tbytes, of which 56 Tbytes have been collected through six cross-sectional harvests, 9 Tbytes through selective harvesting and 6 Tbytes through event harvesting.

Release of NetarchiveSuite in open source July 2007

The two institutions behind the Netarchive.dk, the Royal Library and the State and University Library, have developed a system for web archiving called NetarchiveSuite, and in July 2007 it was released in open source under the LGPL license. NetarchiveSuite is a complete web archiving software package with a primary function of planning, scheduling and running web harvests of parts of the Internet. It scales to a wide range of tasks, from small, thematic harvests (e.g. related to special events or to special domains) to harvesting and archiving the content of an entire national domain.The NetarchiveSuite is built around the Heritrix web crawler, and it has a built-in bit preservation functionality. The system architecture allows for the software to be distributed among several machines, also on several geographical locations as is the case of the Netarchive.dk which operates both from Copenhagen (The Royal Library) and Århus (the State and University Library).

Workshop on NetarchiveSuite September 2007
On September 6-7, 2007, the Netarchive.dk held a workshop for web archivists in order to present the NetarchiveSuite and give potential users a chance for a hands-on experience. The workshop drew 17 participants from 10 national libraries or institutions. Currently two libraries (The National Library of Scotland and Österreichische Nationalbibliothek (Austrian National Library)) are using the system, while three other libraries are testing it. In September 2008 there will be another opportunity for getting to know the system at the ECDL (http://www.ecdl2008.org/).

Staff changes
Claus Lomborg has taken over the task of being the daily manager of the Netarchive.dk as of July 1st, 2008, from Bjarne Andersen, who will be devoting all his time to the ICT development of the State and University Library. Bjarne, who most ably has guided the archive safely through its first three years, will fortunately still be affiliated with the Netarchive.dk, as he will become a member of the Steering Committee from October 1st 2008.

Cross-sectional (snapshot) harvests
We had planned on doing four harvests a year and should thus have completed 12 harvests by now. We have only reached half that number for several reasons: the rapid growth of the Top Level Domain DK, technical problems and the limited number of machines to do the harvest. We are adding more machines and plan to do two more harvests in 2008 and hope that in 2009 we can complete the desired four harvests.

The snapshot harvest is done in stages. First, we harvest all websites up to 10 MB (almost 90 % of Danish websites are smaller that that), then we harvest those websites that are larger than 10 MB. The third step is to check manually the websites that are still not fully harvested to see if they contain more material or if the site has a crawlertrap that we haven't detected before. Once the sites have been "approved" for further harvesting we collect the sites up to 2 GB, unless they have a higher limit and check those who hit that level, and determine whether the harvest should stop or a new limit be set for the site. During the first five harvests we set the limit for manual check at 500 MB, but the number of sites, that hit this limit, increased in 2007 to more than 6,000. We know that 60 % of them were less than 1 GB, leaving some 2,400 to be checked manually, so we raised the limit to 1 GB, beginning with the sixth harvest in spring 2008.

One problem to be solved is log-in sites. It is technically possible to circumvent a log-in, but the harvester cannot distinguish between websites with published materials from sites with private materials, and if we ignore the log-in, we would be collecting materials that are not public and not covered by the Legal Deposit Act. At the moment the only recourse for handling this is to manually check all log-in sites, which is becoming unrealistic for cross-sectional harvesting given our resources. Instead, our plan is to identify technical features that would distinguish access to public sites from access to private sites in order to capture more of the public websites with log-in.

Although the number of sites harvested now has exceeded 800.000 the number of complaints is neglible, less than 130, and they have been solved to everybody's satisfaction. It is usually a question of adjusting the harvest routine or modifying the architecture of the site to reduce to a minimum any inconvenience incurred during harvest. We only rarely get complaints about ignoring the convention of "robots.txt", so that has been generally accepted by the community.

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

Currently, we collect 82 such sites of which 30-35 are news sites. The Editorial Advisory Board (five people, representing researchers and media professionals) has been most helpful in helping us find websites representing the two other categories. We also check newspapers and lists of "most visited sites" as well as keep up with internet research to find relevant material. The websites that are collected are continuously monitored and changes in the list are frequent with the newssites being the most stable. About one-fourth of the sites have relevant content which is accessed via log-in only and that is also harvested. It should be added that in order to find the initial 82 sites many more sites were examined but rejected for this kind of harvest, either because they had an archival function, which means that a cross-sectional harvest would get the entire content of the site, or because they were not covered by the above criteria.

Events harvests

The major event during the past year was the parliamentary election in November 2007. Rumours about an upcoming election had been floating around for some months, and the staff at the Netarchive.dk had prepared a list of urls of sites (political parties, debate sites) to be harvested. So, three hours after the election was announced, the harvester had been fed the list of urls and was crawling the web. Then we looked for websites of individual candidates to feed the harvester. We also faced new challenges in getting the videos, now common on websites, and tracking down campaigns on community sites such as Facebook. We received most welcome help from Internet Archive in collecting videos and we were able to collect some material from Facebook. The harvest of the entire election campaign (which lasted three weeks) took up 2.2 Tbytes of space. Video and embedded digital objects account for part of the increase. In comparison: a local election in 2005 used up 293 GB of space.

Concurrently we have several small events. One of the lessons we have learned is that events may take many forms, and that we should not try to categorize them but keep alert and collect what looks like temporary web pages worthy of saving. It is hard to put on a formula, and it requires staff, who knows about web behaviour and who is tuned to this kind of collecting. At the IWAW Workshop our staff members will share experiences with other web archivists.

Registration of harvests

The system provides technical information on harvests (urls and dates collected, problems etc.) but we also need to document the decisions made concerning what to collect and not collect in order that future researchers may know the content of the archive. We have this documentation in minutes from meetings, as lists etc. in a wiki and are working on systematizing this information for the benefit of future users.

In the Danish library community we are debating a national strategy for cataloguing net resources. The National Bibliographic Agency (a private organisation) is currently cataloguing (i.e. making MARC records for the Danish union catalog) about 2-3,000 net publications a year, mostly materials, which are analogous to printed materials, and they would like to continue this strategy of selecting online ressources for manual cataloguing, while the institutions behind the Netarchive.dk would like to see automatic cataloguing tools developed which would enrich

metadata provided by the producers of internet materials to help our users obtain more accurate search results.

Access

Access is still limited to researchers. The Danish Data Protection Act has interpreted the EU directive as also covering archived data from the web, even if these data have been publicised. We are at the moment working along two tracks, one technical track, the other legal in order to make the archive generally accessible. On the technical track, we have been trying to create a program that could identify sensitive data among the harvested material. However, the studies we have conducted so far have revealed that even the most neutral personal data can become sensitive when the context is considered or when searches make it possible to combine data from many sources. We may still make more attempts along this track but without great hopes for finding a solution.

The other track is legal. We have asked two legal scholars to look at the legal issues involved. The crux of the matter is that the web archive is caught in the dilemma between protecting the individual and supporting the public's right to be informed and have access to data, that are made public. One of the scholars will write a legal opinion on issues involved and comparing practices across Europe. The opinion is due at the end of the year and we will have it translated into English.

This newsletter was prepared by Grethe Jacobsen, The Royal Library, gja@kb.dk. More info on www.netarkivet.dk. An article on the first two years can be found at http://netarkivet.dk/publikationer/CollectingTheDanishInternet_2007.pdf