netarchive.dk

Newsletter August 2009

ECDL in Århus, September 2008.

In September 2008 the annual ECDL (European Conference on Research and Advanced Technology for Digital Libraries) was held in Denmark at the University of Aarhus. The papers and published proceedings are available here (http://www.ecdl2008.org/papers/). The conference included a workshop on web archiving, sponsored by the IWAW (International Web Archiving Workshop) http://iwaw.net/08/index.html. The State and University Library was organizer of the ECDL and several staff members from the netarchive.dk participated in the conference and the workshop, sharing experiences, problems and solutions with other web archivists.

The law – issues in the upcoming revision

The legal foundation for our web archiving activities is the Act on Legal Deposit of Published Material of 22 December 2004 (http://www.kb.dk/en/kb/service/pligtaflevering-ISSN/lov.html). The act is due for a revision during the 2010/11-session of the Danish parliament, the *Folketing*. The two libraries behind Netarchive.dk, the Royal Library and the State and University Library, have begun work on a draft for a revised legal deposit law, which will be sent to the Ministry of Culture in early 2010 in order to be ready for presentation to the *Folketing*. The major issue as far as web archiving is concerned is access (see below).

Statistics

As of August 10th, 2009, the archive contained 112.163 Gigabytes (112 Terabytes) with about 3.5 billion objects. The most common file types are (still) HTLM, JPEG and PDF, but videos are now in fourth place.

Cross-sectional (snapshot/bulk) harvests

The original plan called for four harvests a year, and we should, accordingly, have completed 16 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. In 2008 technical problems prevented us from doing more than one harvest. In early 2009 an upgrade of the hardware enabled us to complete a cross-sectional harvest in a little more than 3 months. During the first stage, all websites were harvested up to 1 GB and this took only 8 days, whereas harvesting the big sites in depth (and in the process uncovering new crawlertraps) took 87 days. During the first stage we collected 2.2 TB and in the second stage 20.4 TB for a total of 22.6 TB. We are currently (August 2009) well into the second harvest.

The TLD DK now has more than 1.3 mill. domain names of which about 1 mill. are active. In addition we harvest about 44.000 Danish sites on other domains (.com, .org, .nu, etc.)

The Danish Broadcasting Company (DR) has long wanted to broadcast the start of a cross-sectional harvest, and on April 16th 2009 a journalist and a camera (wo)man arrived and

filmed the initiating of the harvest. In addition, the journalist had interviewed a woman who had created a website with stories and photos of her children and was pleased to hear that the Netarchive.dk was preserving the site. The story was shown on the late evening news and gave a favorable impression of our activities.

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 87 such sites of which 43 are news sites. The Editorial Advisory Board has been most helpful in finding websites representing the two other categories.

Events harvests

The major events during the past year was the attack on the Danish embassy in June 2008, The Olympic Games in Beijing in August 2008 (only pages concerning Danish participation were harvested), and the financial crisis during the fall of 2008 (as far as it concerned Danish affairs), for which we harvested pages from relevant websites. Interestingly, we discovered that – according to the websites of the banks – there was no crisis, quite in contrast to what the news media and government sites were proclaiming. In April, May and June 2009 we participated in an IIPC project on the election of members to the European Parliament on June 7th, collecting sites dealing with the Danish candidates to the EP.

Registration of harvests

The system provides technical information on harvests (urls and dates for collection, problems encountered etc.) but we also need to document the decisions made on what to collect and not collect in order that future researchers may know the content of the archive. So far this documentation is found in the minutes of meetings, as lists etc. in a wiki, but we are now working on systematizing this information for the benefit of future users. On the Danish version of our website, we now document our collections by listing cross-sectional harvests and event harvests (dates and size) and the selected websites, which are harvested frequently: http://netarkivet.dk/indsamlingsDoku.html

Within the Danish library community we are debating a national strategy for cataloguing net resources in connection with the future Danish National Bibliography. The question is: What should the national bibliography cover? The National Bibliographic Agency (a private organisation which for many years has been responsible for making the national bibliography of books) is currently creating MARC records for the national union catalogue for about 2-3,000 net publications a year, mostly materials, which are analogous to printed materials. 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, partly because a national bibliography needs to reflect the whole national production of published materials, partly to help our users obtain more comprehensive and 'google-like' search results. A central issue here is also to decide how the resources for creating the national bibliography should be used. In a workshop to be held in September 2009 various methods to present the content of the archive will be discussed.

Access

Access is still limited to researchers, as The Danish Data Protection has determined that the EU directive on personal data also covers the archived data from the web, even if these data have been

publicised. We would like to give a broader access at least to some of the material in the archive. For this purpose, we are working on two tracks, one is legal, the other technical. A Danish legal scholar delivered in January 2009 a legal opinion on all the issues involved but also making clear, that the solutions to the problems of giving general access to web archives must be provided by the politicians. In addition, we are following closely what other web archives are doing in providing access.

On the technical track our strategy is to make part of the web, first and foremost governmental sites, available either online or from reading rooms at the Royal Library and the State and University Library. This fall we will, therefore, concentrate on testing various programs in order to present possible solutions to the issue of access to the Ministry of Culture as part of preparing for a revision of the law.

NetarchiveSuite – latest release July 2009

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 August 2007 it was released in open source under the LGPL license. In July 2009, NetarchiveSuite version 3.8.1 was released. At present, The National Library of Scotland is using NetarchiveSuite for harvesting, while two other national libraries, Bibliothèque nationale de France (BnF) and the Österreichische Nationalbibliothek (ONB), have joined the NetarchiveSuite development community and developers from the two institutions are successfully integrated in the NetarchiveSuite development team. The ONB estimates that they will begin using the NetarchiveSuite for harvesting in August, 2009, while the BnF plans to begin harvesting using the NetarchiveSuite at the end of 2009.
For more information see http://netarchive.dk/suite.

Editorial Advisory Board

An Editorial Advisory Board of five people, representing researchers and media professionals, assists the Netarchive.dk in formulating policies and selecting sites for selective and event harvests. The members are elected for a four-year term. The terms of the first board members expired as of July 1, 2009, and at present three new members have been appointed by the Ministry of Culture and the last two members should be appointed by late September.

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