



## Newsletter August 2010

Netarchive.dk celebrates its 5 <sup>th</sup> anniversary.....	1
Revision of the Danish Legal Deposit law.....	1
Statistics 2010.....	1
Bulk (cross-sectional/snapshot) harvests .....	2
Selective harvesting .....	2
Event harvests .....	2
Registration of the content of the archive .....	3
NetarchiveSuite – latest releases.....	3
First Ph.d. on the webarchive.....	3

### Netarchive.dk celebrates its 5<sup>th</sup> anniversary

On July 1<sup>st</sup>, 2005, Netarkivet.dk started the first harvest of the Danish internet domain (.dk) which means that we could celebrate our 5<sup>th</sup> anniversary this year. On June 28<sup>th</sup> about 30 people (past and present staff members, members of the Editorial Advisory Board and the directors of the Royal Library and State and University Library) gathered at the State and University Library for reminiscing, being merry – and trying to predict the development for the next five years.

### Revision of the Danish Legal Deposit law

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 Royal Library and the State and University Library together with the Ministry of Culture reviewed the law during the fall of 2009 and found that the law's formulation of what is covered by legal deposit is broad enough to include foreseeable technological developments in public communication and information dissemination. The main purpose for revision is, therefore, to be able to provide access to the archive for a broader public, as the current law only allows access to researchers. The challenge is how to provide general access without providing access to sensitive personal data.

Netarkivet.dk has spent most of the year finding ways to separate safe data from sensitive data. We decided on a strategy of differentiated access to the archive, so that general access should be given to selected websites, at first websites of public institutions (central, local and regional governmental bodies) then private companies and institutions with a vested interest in keeping their websites free of sensitive data and finally all the websites that can be defined as safe (i.e. without sensitive data). Access should be from the reading rooms of the two libraries (and later from the reading rooms of other Danish libraries) after the websites had been screened for unique identification numbers (every Danish citizen or resident has such a number). The access should be by URL only and it would not be possible to search across domains. Online access for researchers, currently allowed in the present law, would remain unchanged. A final version of our proposal for differentiated access was sent to Ministry of Culture in mid-July 2010. We are now keeping our fingers crossed.

### Statistics 2010

As of August 10<sup>th</sup>, 2010, the archive contained 155.387 Gigabytes (155 Terabytes) with about 4.5 billion objects. The most common file types are (still) HTML, JPEG and PDF, but videos are now in fourth place.

### **Bulk (cross-sectional/snapshot) harvests**

During the first five years we have completed nine bulk harvests and initiated the tenth on July 15<sup>th</sup> 2010. The fall 2009 harvest reaped 24.5 TB, while the spring 2010 harvest reached 28.4 TB (in comparison, the first bulk harvest in 2005 resulted in 5.2 TB). The QA process this past year has focused especially on improving the filters of the harvester, in order that we get meaningful data only. We are, therefore, more than eagerly awaiting the analyses of the harvest initiated in July.

The Danish internet domain (.dk) now has more than 1.1 mil. domains of which about 1 mil. are active. In addition, we harvest about 44.000 Danish sites on other domains (.com, .org, .nu, etc.).

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

One of the challenges of selective harvesting is to capture cross-media activities. For example, the Danish Broadcasting Corporation, DR, has several platforms for their news. We collect the broadcast news programmes separately (broadcast is also included in the legal deposit law), and the news website [dr.dk/nyheder](http://dr.dk/nyheder) is one of the key selective sites we harvest. Again, other ‘satellite-sites’ have emerged from [dr.dk/nyheder](http://dr.dk/nyheder), offering news feed and invitation to join the debate, eg. on <http://www.facebook.com/DRNyheder>. It is of great interest to media researchers that we get as much of these types of data as possible in the archive.

### **Event harvests**

2009 was an eventful year for the Netarchive in the sense that several events had to be harvested. During such harvests we try to capture pages that are new and presumably short-lived. As mentioned in last year’s newsletter we participated in a joint IIPC project, collecting websites concerning the June 7<sup>th</sup> election of members to the European Parliament. Other events focused on during the year were the World Outgames, held in Copenhagen July 25<sup>th</sup> – August 2<sup>nd</sup>, the IOC (International Olympic Committee) meeting and the 13<sup>th</sup> Olympic Congress in Copenhagen in October, municipal and regional elections in November, the outbreak of swine-flu and the possible pandemic threat was in our focus from May to December, as well as the ejection of Iraqi refugees from a church in Copenhagen in July, and of course COP15, the international climate conference in Copenhagen in December, which was in our focus the entire year. Not all events became events in our sense of the word, that is, they generated no or very few new and temporary pages that were not already caught in the bulk or selective harvests, but they all demanded attention of the collection staff.

2010 has so far proven less eventful (from a Netarchive point of view). We focused on the earth quake in Haiti from January to May, looking for temporary pages from Danish relief organisations, pages on major fund raising events and pages belonging to news media not covered by selective harvests. In February and March we also focused on the Danish participation in the Winter Olympics as part of a joint IIPC project.

### **Registration of the content of the archive**

On the Danish version of our website, we now document our collections by listing bulk harvests and event harvests (dates and size) and the selected websites, which are harvested frequently: <http://netarkivet.dk/indsamlingsDoku.html>.

In the Danish Bibliographic Council the debate about a national strategy for cataloguing net resources in connection with the future of the Danish National Bibliography continued. In a workshop held in September 2009 Netarkivet.dk presented various methods to give access to the content of the archive to members of the council as an alternative to creating individual catalogue records for selected online resources. The outcome was a recommendation by the council that supported a broader opening up of the archive and urged that the contents and the tools of the Netarchive.dk be taken into consideration when making plans for the future national bibliography. The council also pointed out that given the nature of the Internet, a traditional and primarily manual selection and cataloguing will cover only the most elementary subjects and types of materials and meet only some of the users' needs. However, the legal restrictions still forms a major barrier to utilizing the archive in finding new ways of helping users find their way in an online world.

### **NetarchiveSuite – latest releases**

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. The latest stable release, version 3.12.1, was released July 6<sup>th</sup> 2010 and the latest development release (3.13.0) was released June 15<sup>th</sup> 2010. For more information see <http://netarchive.dk/suite>.

### **Editorial Advisory Board**

On November 10<sup>th</sup> 2009, The Ministry of Culture announced the names of the new members of the Editorial Advisory Board of the Netarkivet.dk. The four members, who serve a four-year term, represent the university community, the publishing sector and the media sector. Two meetings have been held in January and June 2010 with the new board.

### **First Ph.d. on the webarchive**

On december 11, 2009, Vidar Falkenberg defended his Ph.D. thesis at the University of Aarhus, Department of Information and Media Studies. His thesis dealt with the development of online newspapers in Denmark (1993-2009) and is the first major research project based on the archived material. Falkenberg also included a discussion on the methodical issues in using archived websites as empirical sources. (Danish info: <http://www.imv.au.dk/nyheder/2009/1211>)

This newsletter was prepared by Grethe Jacobsen, The Royal Library, [gja@kb.dk](mailto:gja@kb.dk).