



netarchive.dk

## Web archiving in Denmark August 2009

### A fact sheet

#### The Act

“Act no. 1439 of December 22, 2004 on Legal Deposit of Published Material” went into force on July 1<sup>st</sup>, 2005. Part 3 of the act concerns “Materials published in electronic communications networks” and allows legal deposit institutions to harvest websites within Top Level Domain .dk and websites on other domains aimed at a Danish audience. An English version of the law is found on <http://www.kb.dk/en/kb/service/pligtaflevering-ISSN/lov.html>. An amendment, postponing a revision of the act to 2011, was passed on February 20<sup>th</sup>, 2008.

#### The institution

The task of administering the act is shared jointly by the Royal Library and the State and University Library. The two libraries have established a virtual institution, “Netarkivet.dk” (Netarchive.dk) to perform the task of web archiving. Netarkivet.dk is governed by a Steering Committee of 6 members (3 from each library) representing expertise in digital preservation, IT, legal deposit and national collection building. Netarkivet.dk has a daily manager who reports to the Steering Committee.

#### Resources

The staff comprises 4.5 full-time staff years, shared by 20 employees, who are IT engineers, computer scientists, librarians and library assistants.

#### Hardware

The complete technical setup consists of 12 servers running Linux and a PC-farm of 45 machines running Windows XP. During the coming months the pc-farm will be replaced by 10 servers.

#### Software (open source)

Netarkivet.dk has developed a complete web archiving software package, *The NetarchiveSuite*. The primary function of the NetarchiveSuite is to plan, schedule and run web harvests of parts of the Internet. It scales to a wide range of tasks, from small, thematic harvests (see about harvesting strategies below) to harvesting and archiving the content of an entire national domain. The software has built-in bit preservation functionality. The system architecture allows for the software to be distributed among several machines, possibly on more than one geographical location. The NetarchiveSuite is built around the Heritrix web crawler, which is fundamental for the NetarchiveSuite behavior in harvesting the web. For more information see <http://netarchive.dk/suite>. Currently (August 2009) at least one country (Scotland) is using NetarchiveSuite for harvesting and three other web archives are testing it. Two national libraries, Bibliothèque nationale de France (BnF) and the Österreichische Nationalbibliothek (ONB), have joined the NetarchiveSuite development community. 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.

#### Strategies for harvesting

1. *Snapshot harvesting* (or cross-sectional harvesting) is done to get a complete picture of the Top Level Domain .dk. 7 harvests have been done since July 2005. The eighth harvest was started July 7<sup>th</sup>, 2009. A harvest is begun by loading a list of domains to be harvested, supplied by the Administrator of Top Level Domain .dk. To this list is added a list of URLs on other domains aimed at a Danish audience.

2. *Selective harvesting* is done to gather web pages that are frequently updated and which would be missed by the snapshot harvests such as (1) news sites (national and regional media), (2) “typical” dynamic and heavily used sites representing the civic society, the commercial sector and public authorities, and (3) experimental and/or unique sites, documenting new ways of using the web (e.g. net art). Currently 82 sites are collected daily, weekly or monthly.
3. *Event harvesting* is done to collect web pages from new sites, dedicated to one event and which is expected to disappear when the event is over. An event is defined as something that (1) 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, (2) causes the appearance of new websites devoted to the event, and (3) is dealt with extensively on existing websites.

The above strategies are strategies for collecting the materials, not for building special collections; therefore, all harvested material will merge into one archive, regardless of how it was harvested. The Netarchive.dk takes care to document the harvesting activities in order to provide future users with details on the web archive’s content.

### **Statistics**

The number of domains within Top Level Domain .dk has more than doubled in four years from 607.000 domain names (of which 480.000 were active) in June 2005 to 1.3 mill. domain names of which about 1 mill. are active in July 2009. In addition, more than 45.000 domains aimed at a Danish audience but outside Top Level Domain .dk (on domains .com, .org, .nu, etc.) are harvested. As of August 10<sup>th</sup>, 2009, the archive contains 112.163 Gigabytes (112 Terabytes) with about 3.5 billion objects.

### **Access**

Access is limited to researchers with at least a Master’s Degree and for scholarly and statistical purposes only, according to the Act on Processing of Personal Data. The Danish Data Protection Agency has decided that the data collected through the harvests may contain sensitive personal data and that consequently the entire archive is covered by the regulations of this act. The Netarchive.dk is working on finding legal as well as technical solutions that will enable us to provide general access to the web archive, analogous to access to the physical cultural heritage, for which the libraries are responsible (printed works, pictures, movies, sound recordings, radio and television).

### **More questions?**

See bilingual website: <http://netarchive.dk> (Danish and English) or contact us at [info@netarchive.dk](mailto:info@netarchive.dk).