netarchive.dk

# Newsletter August 2011



## Focus on Webcurating

Many webcurators will probably recognize the experience of the curators of the Netarchive.dk, who in the beginning of webarchiving (some 6 years ago) mostly sat on the sidelines trying to understand the technical issues. However, as technical issues were defined, solved if possible or being worked on, the focus began shifting to collection issues: what to collect, how often, how broadly, how to register, how to provide access and how to draw the attention of the intended users to our collections.

Other webcurators apparently had the same experience as may be seen from the initiative IIPC took when in 2008 it created a webcurator e-mail list. In September 2010, when the IIPC Working Groups met in conjunction with iPres 2010 and IWAW 2010 in Vienna, Austria, one of the offers to participants was a "Workshop/open forum on *Building Web Archive Collections*" organized by Gildas Ilien, Bibliothèque nationale de France, and Grethe Jacobsen, The Royal Library for the Netarchive.dk. Grethe Jacobsen gave a presentation on "Collection policy in theory and practice: the case of Denmark". The session drew a sizeable audience, who participated in the lively discussion, and obviously it fulfilled a need. Consequently, an all day session was organized for the IIPC General Assembly in Hague in May 2011, called "Out of the Box: Building and Using Web Archive collections" again a great attraction as seen from a blog report. Birgit Henriksen, The Royal Library for the Netarchive.dk, participated in a researcher-curator discussion panel.

## How to count activities and collections of a web archive

Netarchive.dk has also been active in the writing of a Technical Report for ISO on "Statistics and Quality Issues for Web Archiving". Grethe Jacobsen has been member of the working group, charged with the task of describing how one should count the bits and bytes that an increasing number of national libraries and other cultural heritage institutions are using resources to collect and present to their users. The target audience is not only the library and web archiving community but also library directors, politicians and officials who are responsible for organizing, prioritizing and funding this activity. Putting a dynamic activity as web archiving collecting on fixed formulas that everybody in the target audience will understand has not been an easy task, but certainly an interesting learning experience. The working group has met in Berlin, Paris and Munich in 2010 and a finished draft was circulated within the group in August 2011. The final technical report will form the foundation for incorporating web archiving statistics into the existing standards for library statistics, that the ISO TC 45/SC8 is in charge of.

**Revision of the Danish Legal Deposit law passed – no broader access**

The legal foundation for our web archiving activities is the Act on Legal Deposit of Published Material of 22 December 2004. The act was due for a revision during the 2010/11 session of the Danish parliament. The main purpose for revision was to test whether we could provide access to the archive for a broader public without providing general access to sensitive personal data (the current law only allows for access by researchers). The Royal Library and the State and University Library presented a report on possible technical solutions to solve this issue in early Fall of 2010. However, the Danish Data Protection Agency wanted a 100 % guarantee, that no sensitive personal data would be accessed by others than those with permission. This guarantee can only be given if all data is checked manually for sensitive data, a so expensive and time-consuming task that it would be impossible to implement. Consequently, the Ministry of Culture decided to drop the issue of broader access for the time being. The changes in the law were limited to dropping the section that demanded regular revision of the law.

**Statistics 2011**

As of July 1st, 2011, the archive contained 222 Terabytes with about 6 billion objects. The most common file types are (still) HTML, JPEG, GIF and PNG.

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

During the year 2010/11 we completed the eleventh bulk harvest, this time respecting robots.txt. Not surprisingly this harvest reaped less than the previous harvest (23.9 TB as compared to 26.5 TB). It was an experiment aimed at testing the assumption that we would miss significant published material if we respected robots.txt (which the law allows us to ignore). Closer analysis of the "missing" data shows that pictures and PDF files especially are not collected, so the assumption still holds true. We will dig in for a further analysis and comparison of this data. The next broad crawl, begun on August 17th, 2011, will again ignore robots.txt.

The Danish internet domain (.dk) now has more than 1.1 million domains of which about 1 million are active. In addition, we harvest about 44.000 Danish sites on other Top Level 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 101 such sites of which 46 are news sites, as news sites are the most frequently updated sites and they do not always cumulate all content. We define the term "news" rather broadly: news from politics, sports, economics, gossip etc.

Given our resources for analysis and quality assurance of the selective harvests the number of sites currently collected is a little bit high. Therefore, we are working on a strategy paper with collection criteria for selective harvests. The paper will specify the criteria and keep the number of selective harvested sites between 80 and 100.

**Event harvests**
2010 was rather uneventful year for the Netarchive.dk in the sense that no major events created new and presumably short-lived pages. Recently, we made a minor event harvest on the terror bombing in Oslo July 22nd, 2011, mainly in order to capture potential and short-lived activities on the websites of Danish right-wing fundamentalists.
We are now getting ready for a national election which has to be held in November 2011 at the latest. Rumours have been circulated for most of the year, and so the web curators have been more or less on alert during that time in case the Prime minister called an election before time.
We are also getting ready to harvest web pages with Danish content related to the Olympic Games in London 2012 and are concurrently participating in an IIPC project chaired by the British Library to harvest all webpages concerning OL 2012.

**Reorganized website under way**
Our bi-lingual website www.netarchive.dk is now 6 years old and in need of an update. The Steering Committee has approved a plan for an update which will take place during the fall of 2011. We expect to present a renovated site in January 2012. Meanwhile you can enjoy a new feature, an English FAQ with a different target group than the Danish FAQ. The English version is aimed at webcurators, researchers, information science students and other interested in how we do things at the Netarkivet.dk. When we announce a new site, please have a look at it and let us know, if you have any comments or questions. The FAQ will also be featured in the renovated website, of course.

**NetarchiveSuite – meetings in Vienna**
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. Two national libraries, Bibliothèque nationale de France and Österreichische Nationalbibliothek are now using the system and developing it together with the Netarchive.dk.
Prior to the IIPC meeting in Vienna in September 2010 webcurators and developers from the three institutions met and exchanged ideas and tips, specified desired features for future developments and created priorities for development of the system.
The latest stable release, version 3.16., was released June 28th, 2011. For more information see http://netarchive.dk/suite.

**Editorial Advisory Board**
The Ministry of Culture appointed a new Editorial Advisory Board in November 2009. The four members, who serve a four-year term, represent the university community, the publishing sector and the media sector and advise the Netarchive.dk in issues concerning collection and use of the archive. Between August 2010 and July 2011 two meetings have been held.

**Organisational and staff changes at the Netarchive.dk**
In May, the contract of the daily manager for the past three years expired and we had to bid goodbye to Claus Lomborg who has ably filled the position during that time. As funding has become precarious, due to the general financial crisis, it was decided to distribute tasks a bit differently, thereby reducing costs. The position of daily manager was reduced to a part-time job,

now held by IT consultant Tue Larsen, who also works with digital preservation at the Royal Library. Tue is in charge of the technical operations of the Netarchive.dk, while questions concerning legal deposit, collections policies and general PR for the archive are handled by the webcurators at the Royal Library and the State and University Library.

This newsletter was prepared by Grethe Jacobsen, The Royal Library, gja@kb.dk. She will retire on September 30th, 2011, and hereby bids a fond goodbye to all her friends in the webarchiving community. It has been really exciting to work with you.