Webarchiving: Legal Deposit of Internet in Denmark. A Curatorial Perspective

Sabine Schostag and Eva Fønss-Jørgensen

Sabine Schostag (sas@statsbiblioteket.dk) is librarian and web curator, and Eva Fønss-Jørgensen (efj@statsbiblioteket.dk) is Chief consultant, National Library Division, both at the State and University Library, Aarhus, Denmark

Summary

Since 2005 archiving the dynamic internet has been required by law in Denmark. This article tells the story of the recent seven years of experience with archiving the internet: What is covered by the law? How do we organize the work? How do we collect the web in practice? Who has access to the web archive? And finally, what are the challenges and future perspectives? The article focuses on the curational aspects and does not go into technical details.

Introduction

When the Danish parliament passed the Act on Legal Deposit of Published Material of 22 December 2004, Denmark got one of the most advanced and comprehensive legal deposit laws in the world. Legal deposit in Denmark has been regulated by law since 1697, and the following 300 years only printed materials were covered. A revision of the law in 1997 introduced legal deposit of audiovisual media in terms of commercial cd’s and videos, and also – in view of the upcoming internet – static web documents were included, such as reports and other text documents, e.g. published in Word format. The recent revision in 2004 further complied with the fast technological development: Radio and television became part of the law, and so did the dynamic internet. The new law went into force on 1st of July 2005.

The law

The intention of the law is – as far as possible – to get a total coverage of the Danish part of the internet. In view of the rapid growth and ephemeral nature of the web, three different and complementary collection strategies are prescribed in the law:

– Broad crawls: Four times per year a ‘snapshot’ is taken of the whole Danish internet. Encompassed are all domains from <.dk> and Danish websites published under other extensions, such as <.com>, <.org> etc.
– Selective crawls: In order to capture websites which are frequently updated and therefore cannot be covered completely by the snapshots, a selection of 80–100 sites is continuously harvested – from once a month up to six times a day. Selective harvesting focuses on three types of web sites: News sites (all national and selected regional media); Dynamic and frequently visited websites representing the civic society, the commercial sector, and public authorities; and finally, Experimental and/or unique sites, documenting new ways of using the web (e.g. net art).
– Event crawls: Occasionally, web sites dedicated to a specific event will appear and disappear as the event is over. An event is defined as something that creates a debate among the population and is expected to be of importance to Danish history or have an impact on the development of the Danish society. An additional criterion is that the event causes the emergence of new websites devoted to the event and is dealt with extensively on existing websites. The law prescribes event harvesting to take place two to three times per year. An event can be either predictable (such as national and local elections), or unpredictable (such as the ‘cartoon crisis’, political riots, or natural disasters).

Only the publicly available part of the internet is covered by the law. Private content such as family sites with restricted access via login are not covered, and
that is also the case for organizational intranets, directed towards e.g. the employees of a company.

**Organization**

For many years the administration of legal deposit in Denmark has been undertaken jointly by the State and University Library and the Royal Library with a specific division of labor. The task of collecting the internet was likewise given to the two institutions, which in 2005 formed a ‘virtual organization’ called Netarchive.dk, aiming at developing technical and curational procedures for running the archive according to the law. A project team was established with a steering committee, developers (crawl engineers) and web curators from both institutions. The team is assigned a total of 5 man-years. The efforts are distributed as follows: curation: 2 man-years; technical development: ½ man-year; technical operation: one man-year; project coordination: ½ man-year. The two institutions share the efforts evenly.

The curators agreed on sharing the task as follows: The Royal Library team would take care of the broad crawls, the State and University Library team would handle the selective crawls, and we would cooperate on the event crawls. Further, an advisory board with representatives from the research community, the media industry and other relevant organisations was established, primarily to support the work with selective harvesting.

**Collection in practice – from a web curator’s everyday life**

**Broad crawls**

As mentioned above, a broad crawl should capture as much content as possible from the Danish part of the internet. Before starting a broad crawl we have to ‘feed’ our production system with a list of domains to be harvested. Top Level Domains (TLD) do not pose a problem, as DK Hostmaster who is responsible for the administration of all <.dk> domains is obligated by law to provide us with this list. As to Danish websites on other Top Level Domains such as <.com>, <.info>, <.org> or nowadays anything you could imagine, it is more complicated and over the years we have done quite a bit of work on discovering relevant Danish websites outside <.dk>. We have experimented with several different strategies, e.g.:

1. Google search on Danish localities (like city names), limiting the search to non-.dk sites
2. Search for domain names in all previously harvested archival content
3. Domains nominated by the public via www.netarkivet.dk

We run them all through an IP geographical locator (GeoIP) to reveal hosts located in Denmark.

All broad crawls are done in two steps: First step with a limit of 10 MB per domain, because most websites contain less than 10 MB. Thus we are able to harvest almost 85% of all Danish websites. Second step is with a default limit of 1 GB per domain which captures most of the remaining 15%. Finally, we crawl the few really big Danish websites, such as dr.dk – the website of the Danish Broadcasting Corporation.

<table>
<thead>
<tr>
<th></th>
<th>2010</th>
<th>2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of websites harvested</td>
<td>898137</td>
<td>956824</td>
</tr>
<tr>
<td>Number of websites bigger than 10 MB</td>
<td>131736</td>
<td>140633</td>
</tr>
</tbody>
</table>

Our very first broad crawl took almost half a year, i.e. we did not comply with the prescription of the law, which says four broad crawls a year. However, due to technological improvements of both hardware and software we are now able to do four broad crawls a year. With our second broad crawl in 2012 we beat all records: it lasted only 53 days. On August 15 we pressed the start button for our third broad crawl 2012.

**Selective crawls**

The scope of selective crawls is – popularly spoken – to fill the gaps between the broad crawls. More precisely this means to gather content from websites that are frequently updated so this specific content would be missed between two broad crawls.

Sites with content to be considered in selective crawls typically cut into three subject areas:

- News sites (about 60% of all selected sites)
- Dynamic and very frequently visited sites representing civic society, the commercial sector and public authorities (about 30%)
Experimental and/or unique sites, documenting new ways of using the web, e.g. net art (less than 10 %)

According to the law, selective crawls are supposed to be made on 80–100 websites. So our first step in 2005 was to find 80-100 relevant websites among a total number of 6–700,000 Danish websites (today there are more than one million).

We used different methods to select the relevant sites. In the beginning, we studied chart lists, best-on-the-web lists and statistics on the Danes’ online behaviour. Apart from news sites, most of the sites we examined accumulated their content, in which case it is captured by the broad crawls.

Once we had found the 80 sites, our selection work was not finished. The web is extremely dynamic and constantly changing. The selected sites are not static, that is to say a selected site would not necessary be part of the selective harvests forever. Sites change, sites die, new sites emerge etc. To maintain a collection of 80–100 selected sites is an ongoing process of screening the Danish part of the web for relevant sites. The Advisory Board contributes with proposals for new sites to be crawled selectively, as well as anybody can propose sites on our website www.netarkivet.dk

Furthermore, we were confronted with the questions: How to organize our work? What exactly do we have to do, and who does what at which time? We elaborated a workflow (Fig.1) with seven steps:

1. Identification of a site
   A potential candidate for selective crawl is identified.

2. Initial examination
   The website is examined. First of all we determine whether the site cumulates all content or not. If so, it will be captured by the broad crawls, if not, we create a harvest profile.

3 Analysis and determination of the harvest profile
   Just as in the broad crawls, the harvester has to be fed with URLs. The harvest profile determines, which URLs we will harvest from the site into the production system. To maintain a collection of 80–100 selected sites is an ongoing process of screening the Danish part of the web for relevant sites. The Advisory Board contributes with proposals for new sites to be crawled selectively, as well as anybody can propose sites on our website www.netarkivet.dk
times a day to once a month. We compiled a form sheet for the documentation of the harvest profile.

Let’s have a look at our news sites. For some news sites we can make do with harvesting the front page and two more levels in order to capture all ephemeral content, others are more complicated. Take www.dr.dk, the website of the Danish Broadcasting Company, and one of the biggest Danish websites. The site cumulates bunches of content, but news from different areas are not kept for a long time on the site. There is a main news page (frontpage), and you find sub news pages for politics, sports, culture and so on. In addition there are the regional broadcasts’ news pages. It would be impossible to cover all these pages with just one harvest definition and one URL without capturing a lot of redundant material. You can find dr.dk in 8 harvest definitions and within these definitions we harvest content from about 60 URLs.
4. Data entry into our production system
Without implementation of the harvest profile into the production system nothing would be harvested. When we started up we had about 20 different harvest definitions (Fig. 2). A harvest definition tells the system which URLs from which domain are to be crawled in a given depth and frequency and at which time schedule. The determination of depth and frequency is called configuration. Many domains can be connected to one harvest definition (e.g. frontpage URLs from about 40 websites are crawled 6 times a day) and URLs from one domain can be crawled in many definitions (e.g. URLs from dr.dk are crawled in about 20 definitions). This depends on the size and structure of a website.

5. Quality assurance (QA)
Having captured the content to our web archive, we check whether we have got what we intended to get. First step of QA is having a look at sample pages from the harvested web sites. Do they look like they did on the "living web"? If so, QA is finished for the page, if not, we inspect the crawl logs in order to find the reason.

6. Monitoring
When we have ticked off the QA, we leave the page alone for at least half a year, that is to say we keep an eye on the bytes/objects harvested in each crawl. If the number of harvested bytes in each crawl within the same definition is roughly constant, we do not take further actions. But if we observe remarkable deviations, we make closer investigations.

7. Follow up
After at least six months of monitoring, we go back to the analysis again: Is the content of a given web page still of interest for the selective crawls? Does the harvesting profile still match the structure of the page? Thus, every six months either the page will take another tour through the workflow, or it will be excluded from the selective sites.

Currently, we collect selective content from 95 sites, of which 56 are news sites.

Even though web content is virtual, it takes a lot of space. A couple of years ago we introduced deduplication to our web archive, i.e. the system automatically ‘erases’ content already harvested in the previous crawl. Deduplication has really had an impact on this type of harvests, showing a saving of 50-70% on bits to be archived. This also means that we can harvest deeper than absolutely necessary, just to be on the safe side. Deduplication does not work on content in html format, but html content only takes very little space.

Event crawls
Event crawls enable us to capture ephemeral web content emerging in conjunction with a specific event. Event crawls are done in addition to broad crawls and selective crawls.

In practice this means that we identify URLs linking to web content related to an event and feed our harvesters with these URLs. The URLs will be crawled until we decide that the event has faded out or finished. Thus each event crawl results in a special thematic web collection.

The average number of events is supposed to be two to three a year. There are planned events (e.g. parliamentary elections, the Olympic Games) and events that just happen. The planned events are easy to harvest. The first event, we harvested, was the local elections in November of 2005.

But how to predict occasional or unplanned events? Does a tsunami on the other side of the earth have an impact on the Danish part of the internet, which would qualify an event harvest? Who can predict the impact of a given event on the Danish part of the internet? The first event of that kind was the so called Cartoon crisis. The catalyst of this event was the cartoon of Muhammad with a bomb in his turban published by one of the biggest Danish newspapers. Nobody in the Netarchive team would have guessed the dimension of this event. It became the issue of our second event harvest.

In 2011, we wanted to be at the leading edge concerning an occasional event: an announced international neo-nazi demonstration in the city of Aarhus. We prepared and started an event harvest, but it turned out that people payed scant attention to this event, which went on without any further ado.

So far, we have harvested 24 events, and we have just started a new one on a tax cause of the Danish Prime Minister's husband. This event is of potential interest because a commission has been settled to investigate a supposed political leak, and also the role
of the press is part of the case. We do not yet know whether this event will create new web pages and debates on the internet. But we are prepared.

Content of the Danish web archive (2012-08-26)

<table>
<thead>
<tr>
<th>Category</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>283 TB</td>
</tr>
<tr>
<td>of which:</td>
<td></td>
</tr>
<tr>
<td>Selective crawls</td>
<td>33762 GB</td>
</tr>
<tr>
<td>Event crawls</td>
<td>14603 GB</td>
</tr>
<tr>
<td>Digital objects</td>
<td>9.0 billion</td>
</tr>
</tbody>
</table>

Tools for harvesting the web

How do we capture and archive all these Danish web pages? First of all we needed a web crawler, also called web harvester. Just as you can “collect” all the corn from your cornfield with a harvester, you can use a web harvester to collect all the URLs and objects you want.

Almost the only existing web harvester is Heritrix. Heritrix is the Internet Archive’s web crawler, which was specially designed for web archiving. It is open source and written in Java. The main interface is accessible using a web browser (Fig. 3).

Heritrix was developed jointly by the Internet Archive and the Nordic national libraries on specifications written in early 2003. The first official release was in January 2004, and it has been continually improved by employees of the Internet Archive and the community.

Furthermore, we needed a curation tool – a software to manage our different crawls, so the developer team at the State and University Library created NetarchiveSuite. NetarchiveSuite had been developed for curation of both broad crawls from predefined domains and crawls of selected domains with regular intervals. That is to say NetarchiveSuite was designed for the web harvesting prescribed in the legal deposit law. The NetarchiveSuite is built around the Heritrix web crawler.

The main functionalities of NetarchiveSuite are planning and performing archiving of parts of the internet (more or less big parts). NetarchiveSuite differs from other web archiving tools by its usability for both capturing content from a few websites and
for building huge national collections on top level domain as for example <.dk>. With NetarchiveSuite, we can create all the harvest definitions we want to manage our captures of the web. A bit preservation function is part of NetarchiveSuite, so we are able to continuously replicate collected data to different geographical locations. On 4th of July 2007, NetarchiveSuite was released as a complete software package under the LGPL license. With all its functionalities NetarchiveSuite is designed for big institutions.

In conjunction with ECDL/IWAW/IPRES conferences\(^5\) we have held workshops for web archivists all over the world in order to present the NetarchiveSuite and give potential users a chance for a hands-on experience. After a couple of years our lobby activities resulted in a NetarchiveSuite community with Austrian and the French national libraries. Österreichische Nationalbibliothek was the first to join the NetarchiveSuite community in 2008, followed by Bibliothèque Nationale de France in the same year. We hope that more web archives will join the NetarchiveSuite community. The more we are the better we can improve our tool.

Developers and curators from the three institutions meet once a year, and in between the annual meetings we hold teleconferences and use a wiki for common documentation. The curators primarily use these meetings for coordinating a wish list for new features to NetarchiveSuite. Lately, it has become clear that the advantages of NetarchiveSuite can actually complicate our daily curation work. Over the years, we have produced such long lists of harvest definitions and schedules etc. that it gets difficult to keep track

---

### Configurations

<table>
<thead>
<tr>
<th>Domain name</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;.dk&gt;</td>
<td></td>
</tr>
</tbody>
</table>

The following domains are aliases of this domain: bornholmsradio.dk, danmarkskanalen.dk, digital-tverrgangen.dk, drysen.dk, nordmag.dk, radiodanmark.dk, radionorden.dk, radiosyd.dk, senuer.dk

---

Fig. 4: List of configurations
of our actions. Therefore, we are currently working on developing features for a more clear presentation in our NetarchiveSuite GUI (graphic user interface). This might be the ability of sorting long lists of harvest definitions and configurations (Fig. 4) according to different criteria, such as listing inactive definitions and configurations at the bottom of the list.

Documentation

Documentation is an important part of curation. The NetarchiveSuite GUI does not offer enough facilities for documentation of such huge data collections as the Netarchive. So what would be a good tool for documentation? First, we decided to create a form sheet for every domain we crawled selectively. In the beginning we wrote the form sheets in Word and filled them into a system of folders and subfolders, where we moved them around according to their stage in the workflow. However, this system of folders and several levels of subfolders was growing and became rather complicated and confusing, so we decided to change over to something more flexible. We made our choice – and moved our documentation to a wiki: we converted each “domain sheet” to a wiki page. A table representing the workflow connects these wiki pages.

First and foremost this wiki documentation was made for ourselves. It was and is a must for our daily work as curators. The central document for the selective crawls is a table reflecting the main actions in our workflow (Fig. 5): Identification, analysis, data entry, QA and monitoring. There are many advantages in wiki technology and structure, especially the feature of creating links criss-cross throughout the whole wiki. The workflow table links to a document for each website we crawl. This document is reused in a table reflecting every website we ever have included in our selective harvests and for which period. You can find a reproduction on the documentation table on our website, while the workflow table is an internal tool.

We also document our event harvests on the wiki, and last but not least we document technical problems, i.e. everything from electrical power interruption, which stops our harvester, to advanced website features, which we are unable to capture with Heritrix. This is important for the documentation of our collection and also useful information for researchers who access our web archive.

Access – who and how?

Access to the Netarchive is very restricted, and it is difficult to explain to people, why they cannot get access, when the content has been and often still is available on the open internet. However, once the web is archived and stored in the State and University Library and the Royal Library, it is subject to the administrative law.

Only researchers at PhD level or higher can currently get online access to the archive. Others with a scientific purpose are allowed to access the archive on the premises of the legal deposit institutions. The access conditions are regulated partly by the Copyright Law, partly by the Act on Processing of Personal Data. The reason for this limited access is that some of the content of the archive may contain sensitive personal data, and the Danish Data Protection Agency therefore deems the whole archive sensitive. Also, access for a wider audience would require an agreement with the copyright holders.
We have investigated several ways to open the access for a broader public. We do not consider the copyright restriction as the biggest problem, because we habitually make agreements for wider access to the cultural heritage, primarily for study and teaching purposes at higher education. One of the special features of Danish copyright is the use of extended collective licenses, which make it possible to license the use of a large volume of works protected by copyright.

The problem with the sensitive personal data is more complicated, and we have tried to solve it during the recent years. Among other things, we have experimented with automatic detection of personal identification numbers (in Danish: CPR-numre) in the archive. If we could detect them, we might be able to make a de-classified archive more broadly available. The experiment was quite successful, but according to the Danish Data Protection Agency the method was not safe enough. We also tried to examine whether the archive could be subject to another legislation, e.g. the Danish Archives Act\textsuperscript{7}, but neither this was successful.

Consequently, our users so far are researchers at PhD level or higher. Each user has to fill in an application, verify that he/she fulfills the access requirements and declare that he/she does not intend to use personal during and after the research period. Until now, about 20 researchers have got access to the archive. Examples of research projects based on the Netarchive are

- The history of online newspapers - a historically based analysis of online newspapers as a distinct medium, with the purpose of identifying the characteristics of online newspapers.
- The interaction between television news and news on the web (case: DR and TV 2)
- The media coverage of the cartoon crisis.
- How did Danish politicians use the new information and communication technology during the 2007 parliamentary election?

Access tools

Sometimes we might have been tempted to think: fortunately access to the archive may only be given to a few persons. The practical part of giving access was rather complicated. We did not really have search facilities to offer to our researchers. Until recently, we had to guide/coach each individual user entering our web archive.

But this is history now. A couple of years ago we started to develop a wayback machine based on the Internet Archive Wayback Machine\textsuperscript{8}. With an internet wayback machine you can travel back in time on the World Wide Web. Therefore, we were able to make a de-classified archive more broadly available. The experiment was quite successful, but according to the Danish Data Protection Agency the method was not safe enough. We also tried to examine whether the archive could be subject to another legislation, e.g. the Danish Archives Act\textsuperscript{7}, but neither this was successful.

Consequently, our users so far are researchers at PhD level or higher. Each user has to fill in an application, verify that he/she fulfills the access requirements and declare that he/she does not intend to use personal during and after the research period. Until now, about 20 researchers have got access to the archive. Examples of research projects based on the Netarchive are

- The history of online newspapers - a historically based analysis of online newspapers as a distinct medium, with the purpose of identifying the characteristics of online newspapers.
- The interaction between television news and news on the web (case: DR and TV 2)
- The media coverage of the cartoon crisis.
- How did Danish politicians use the new information and communication technology during the 2007 parliamentary election?

Access tools

Sometimes we might have been tempted to think: fortunately access to the archive may only be given to a few persons. The practical part of giving access was rather complicated. We did not really have search facilities to offer to our researchers. Until recently, we had to guide/coach each individual user entering our web archive.

But this is history now. A couple of years ago we started to develop a wayback machine based on the Internet Archive Wayback Machine\textsuperscript{8}. With an internet wayback machine you can travel back in time on the
internet, i.e. the wayback machine is a tool to look at archived web content. Now our users get access to the archive via a password protected proxy connection to the wayback machine (Fig.6). They can browse all archived URLs.

Next step will be the implementation of free text search, which will make it much easier to carry out thematic studies in our archive.

But more advanced and specialized tools are demanded by the users. The very nature of such a huge amount of electronic sources calls upon the possibility of data mining and triggers new research methods. The Netarchive is a central element in the Danish Digital Humanities project (DigHumLab), which has a subproject called Netlab9. One of the aims of Netlab is to develop a workspace (e.g. for searching and visualization) and build the relevant skills for using software supported methods in the study of internet materials. The project will monitor the broad array of existing analytical software with a view to identifying, testing, adjusting or possibly developing a limited set of software tools which can be used for a broad range of different analytical purposes.

**Challenges and future perspectives**

You might compare our harvest challenges to virus and antivirus programs: once a virus is created and spread, the antivirus developers set out to chase it. Once a fancy technology has entered the internet, we have to teach our harvesters how to catch these new technologies. You are always one step behind.

In 2005, the central content of web sites was (html) text, maybe with some illustrations or photos. Today, you nearly can’t imagine a web page without either flash presentations, embedded videostreams or mashups such as a stream from a facebook page.

Just take a look at the Danish social democrats’ web page from 2005 (Fig. 7)

And then compare it with their page now: http://socialdemokraterne.dk/
One of our big challenges is capturing streamed content, both images and sound. We have experimented with capturing selected videos from YouTube — this was done more or less manually. In the long run we need a feature in Heritrix that automates the process.

As this is not a specific Danish challenge, it is worked on world wide. For example, the EU project LiWA has made an approach to capturing rich media\(^\text{10}\), but this approach is not large-scale implementable. Other new web features such as flash presentations, advanced java script, mashups and interactive pages cannot be captured by Heritrix either.

An even bigger challenge is the social media. Facebook usage represents a big part of every day life in these years, and e.g. open Facebook profiles of politicians or electoral candidates are important to capture as part of the cultural heritage. But not only the technology used by Facebook with constantly updating pages causes us headaches: Facebook is blocking our web crawlers, and as Facebook is not subject to Danish legislation, we cannot refer to the legal deposit law, but only hope to obtain their cooperation. So far, we have not been successful.

New challenges have emerged outside the World Wide Web: Apps for tablets and smartphones are part of the electronic communication network, and they are in principle subject to legal deposit. But Heritrix, our harvesting tool, is far away from being able to capture apps and what ever might appear in the future.

Even though many of these challenges are not related specifically to NetarchiveSuite, we exchange experiences in the NetarchiveSuite community: the more cooperation, the better for finding solutions.

**International cooperation — IIPC**

Collecting the internet as part of the cultural heritage is a challenging task, both from a technical and a curational point of view, and many resources are needed to keep up with the development and constant growth of the web. Therefore, international cooperation is a must. The international NetarchiveSuite community has already been mentioned above.

As Denmark was one of the first countries in the world to include the web in the legal deposit law, Netarchive.dk has been an active member of IIPC (International Internet Preservation Consortium) from the start. The mission of the IIPC is “to acquire, preserve and make accessible knowledge and information from the Internet for future generations everywhere, promoting global exchange and international relations” \(^\text{11}\). IIPC has working groups for harvesting, access and preservation and develops a set of high quality, easy-to-use and open source tools for setting up a web archiving chain. IIPC also initiates global campaigns to collect specific events. One example is the 2010 Winter Olympics, where several national web archives, including Netarchive.dk, contributed by delivering relevant national URLs. The result is a common global archive which provides public access to the event\(^\text{12}\).

### Notes

2. https://www.dk-hostmaster.dk/presse/statistik/antal-registrerede-domaener/
3. Heritrix: https://webarchive.jira.com/wiki/display/Heritrix/Heritrix
7. An (unofficial) English translation is available on http://www.sa.dk/media(3000,1033)/Danish_Archives_Act.pdf
11. see also newsletter 1: http://liwa-project.eu/index.php/newsletters/