Chasing the news
report from 10 years of digital legal deposit in Denmark

Tonny Skovgård Jensen, Sabine Schostag and Niels Bønding, The State and University Library, Aarhus, Denmark

Abstract:

This year the current legal deposit act in Denmark has a 10 years anniversary. Since 2005 content published in "electronic networks" has been part of the legal deposit program, presenting a whole new range of challenges for our library as a legal deposit institution. Our take on this task has not been focused on the type of content, like entertainment, information or news, but rather by the type of media channel: We have been harvesting the Danish part of the world wide web, together with The Royal Library in Copenhagen, and we have been collecting radio and tv-broadcasts, including also news. In the same period, we have been preparing digitization of 32 mio. pages of printed newspapers, now producing 1 mio. digital pages each month.

This means that we have a rapidly growing collection of digital news, some of it born digital, some of it digitized, but not in one specific news collection. Each type of content has its own set of formats, access rights and access platforms, and each type has its own technological challenges. This is not in any way an ideal situation, but still a lot of content has been collected, saved and made available, and a growing range of access opportunities are emerging, and use is increasing.

This paper/presentation will offer an overview of what we have done and learned so far, chasing the news everywhere and catching some of it for the benefit of present and future researchers.

Digital legal deposit - where we started

Journalists, editors and news media on various platforms have always been chasing the news – the breaking news story that would sell more papers or get more people watching the television news. Memory institutions like our library, The State and University Library in Aarhus Denmark, has been collecting news media as an important source of documentation of how people lived, how they debated to set the agenda for future actions. This was considered so important that legal deposit laws ensured that this collection would be done. For many years the printed newspapers were the only news media that was collected, and this has been going on for decades in almost the same way. Radio broadcasts have been part of the news media since 1925 in Denmark, and television since 1951. For many years these ephemeral broadcasts were not systematically collected, and they were not a part of the legal deposit program. During the 1980's things started changing – slowly. The State Media Archive was established in 1987 as a department of The State and University Library, and we started to collect television and radio broadcasts on tape, not based on legal deposit laws, but on voluntary agreements with the two national television broadcasters that existed at that time. This embracing of radio and television as cultural heritage was not just a consequence of technological development making it economically feasible to collect it, but also a slow recognition that radio and television were also a quite important part of the society life, and as such important sources for future research and documentation, parallel to printed newspapers, books and films.

It was in this setting that the internet started to expand and ultimately change everything, including what we define as cultural heritage and is included in the legal deposit legislation.

This short background gives a context to better understand the scope of the transformations that have shaped the news media since the appearance of the internet during the 1990's, as a fundamentally new and groundbreaking new way of interacting and communicating. While we all know that transformations are still going on right now, today, it can be hard to tell what is significant changes and what is less significant. In this paper we will take a retrospective look at a specific period of collection building in Denmark, a period defined by the reach of the current legal deposit legislation in Denmark, which has a 10 year anniversary in 2015. When launched it was quite modern, making both radio and television broadcasts as well as a complete collection of the Danish part of the internet subject to legal deposit collection. What we will look at in this paper is how we have implemented the law in Denmark, how well it has withstood the "transformations" we have seen, and where the major challenges are. We will focus on "news", and we will focus on the digital. But news is in our context mixed up with other sorts of content, as we will see, and the line between printed, broadcasted and online content is more and more difficult to discern, as it is all part of various digital workflows, and companies are using the internet to cross from newspaper to broadcaster and from broadcaster to internet news media.

To get an understanding of the "transformations", let us first take a brief look at where we started in 2005. What the world looked like at that time? How was content used, how was news consumed, which technology and infrastructure was common at that time?

- Facebook was just about a year old, and had about 5 million users. Worldwide.
- YouTube was founded this year. Few people could imagine at that time the enormous amount of video that was soon to be distributed online.
- Twitter was founded the year after, in 2006
- You would have to wait two more years to see the first iPhone, one more year to hear of the first android smartphones, and 5 years to see the first Ipad.
- State of the art devices were still PDA's and also the first smartphones, like the Sony Ericsson P910, but slow connections and slow hardware made them quite limited for practical purposes.

In other words, it was a completely different setting than the one we see today. Television news was something you watched on your tv-set, and it was broadcasted, not delivered on demand. Mobile phones were not so smart, and were still mostly used as phones, and internet services were something that you used when you were sitting at your desk, using your desktop PC.

From the 2005 point of view it is suddenly very easy to see the transformations that have been going on. And all of these transformations have also affected the way that news is being distributed. So how does a legal deposit act from 2005 handle such a transformation, and how have we at The State and University Library implemented the act to try to adapt to the changes?

The scope of the 2005 legal deposit act - and what we have done to implement it

Radio and television

From a news perspective the significant changes in the legal deposit act were that radio and television now for the first time became part of the legal deposit collection. Radio and television broadcasts are now also being collected, and since the beginning in 2005 we have recorded it off-air and stored it digitally only. In the beginning we collected various content ourselves, using a mix of several antennas and satellite dishes, resulting in a heterogeneous mix of formats, and accompanying technical metadata. For the past 2 years we have bought almost all the content from a major cable television provider, resulting in a much more homogeneous quality.

- Close to 2 million broadcasts are now in the archive, and around 30.000 new broadcasts are added every month.
- The archive is around 2.000 TB in size, growing around 300 TB each year.
- 14 television channels and 6 radio channels are collected 100% (among these all the public service channels), and further channels are collected partly. The legal deposit act specifies that all “Danish” content must be collected. We do not evaluate each single program, but rather each channel. For those channels where a major part of the content is non-original content like American series, we only collect those time spans with completely or partially Danish content.
- Metadata: We do not have resources to manually register and annotate each single program. Instead we buy metadata from commercial companies.
- Digitization: We are working on a plan to save our collections of pre-2005 content on tape.

The Netarchive

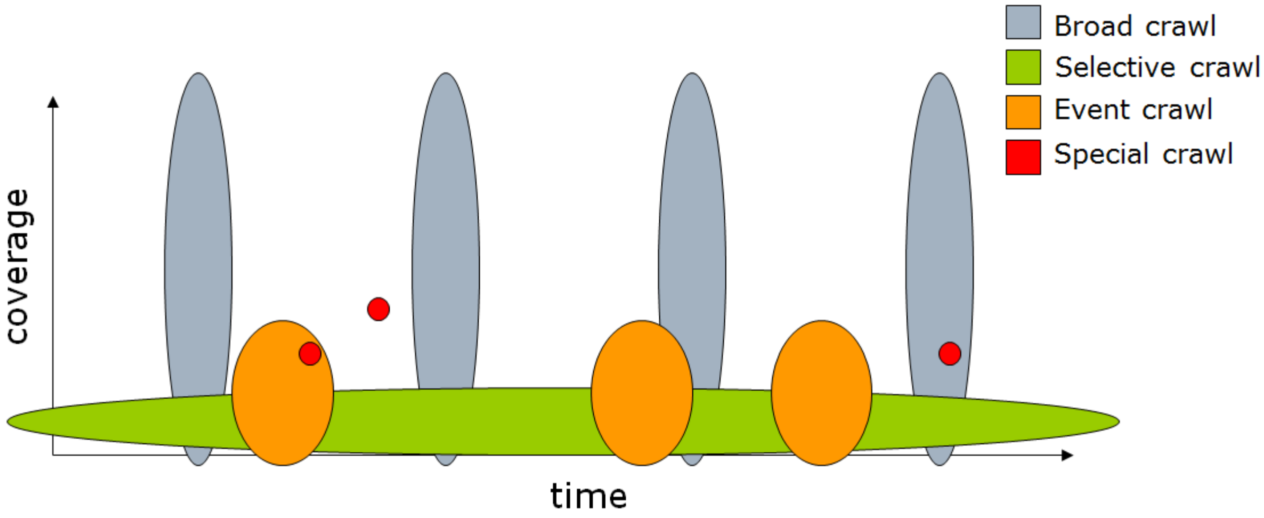
Even more significant collection and preservation of the web was now a legal deposit task. The Netarchive, the Danish national web archive, is a collaboration between the two National Libraries, the State and University Library and The Royal Library in Copenhagen.

According to the 2005 legal deposit act, The Netarchive is obliged to collect and preserve “content distributed in electronic communication networks”. Both national libraries were part of preparing and shaping the legislation, and so were two internet research pioneers, Niels Ole Finnemann and Niels Brügger, from the University of Aarhus. The official remarks to the act also specify the collection strategies. The Netarchive has 4 collection strategies:

1. Broad crawls: 4 snapshots of the whole Danish part of the Internet every year. We try to capture as much content as possible with the broad crawls.
2. Selective crawls of frequently updated web sites: we harvest content from 80-100 sites between 6 times a day and once a month. The selective crawled web sites come from three subject areas:
    a. News sites (Web sites owned by publishers of printed newspapers, broadcast stations, online news media).

   b. Web sites from public institutions (ministries, municipalities, public portals)
   c. Experimental web art

3. Event crawls: An occurrence generating new web sites, and increasing activities on existing sites, nowadays especially on social media, will be the reason for an event crawl. Events might be planned, such as parliamentary elections or cultural events, or they might happen unexpectedly, such as the terror attack in Copenhagen in February 2015
4. Special crawls: e.g. a crawl of a web site which is to be closed, or a crawl on demand (as a service for our researchers). This particular collection strategy was not in the official remarks, but was added later, because it turned out there was a need for it.



Web archiving is done automatically with a web crawler called Heritrix. Apart from the extra attention some of the news sites get by being part of the selective or event crawls, there is no distinction or special treatment for the news content. Anyway it would be hard or impossible to separate news content from the rest. It would be hard even to define news content in any way that would comply with the way people think about news today, from the traditional news media like broadcasters and newspaper publishers to the enormous amount of websites, blogs, twitter-accounts, discussion fora etc. that in many very different ways participate in communicating news. We catch a lot of it with our broad crawls, but certainly not nearly everything.

Many news sites are part of the selective crawls. This means that they are not just harvested 4 times per year, but is harvested in a more frequent pattern, individually designed to fit the specific site, and frequently checked manually by the curators, something that is impossible with the broad crawls. Typically these news sites are collected several times each day, making it possible to harvest quite a lot of the content on the top levels before it is replaced by something new, moved down to deeper levels of the site or removed. A growing challenge with news sites is the use of various kinds of paywalls which prevent access. According to the legal deposit act we are entitled to be granted access, free of charge, which means that they have to provide us with passwords that we can feed to the crawler software. This is all very well, but it is a time consuming task to set it up, and to check continuously that it still works. It also only works at all when our crawler software is able to enter the password correctly on the site. Frequent change of technology on the sites make difficult. Some major newspapers have helped us out by setting their sites up to allow free access from the IP-addresses that our servers use, which means that we get access without passwords. This works quite well.

Almost 600 TB of data has been collected by now, and the archive is growing more than 100 TB each year. Collecting "everything" of the Danish part of the Internet, as specified by the act, is not just a matter of pressing a button and then wait for the computers to do the job. The software used for the harvesting, the web crawler Heritrix, which is used by almost all national web archives, was developed by the Internet Archive in the 1990's. At that time web pages were built with html-texts – and maybe a few images. Web pages have evolved dramatically since then, and are now much more complex, with streamed videos and sound, Flash content, javascript etc. As new technologies constantly emerge on the web, Heritrix and the community around it, try to keep up with the changes by developing plug ins and by applying new tools. Software developers and curators work together, both on a national and international level to test and apply additional software to keep up with the ways that new trends and technologies change the structure and content types of web pages.

The continuous change of web pages is not only a technical challenge. Once you archive a book acquired by legal deposit, this book will always be the same book. Web pages on the contrary are modified, especially news sites are constantly updated, and the way they are updated, the technology they use, the layout and design is constantly changed. So once a news site is being harvested according to a specific schedule, the curators can't just leave it alone and continue with the next task. News page harvestings have to be monitored constantly to check if we actually captured the content we wanted to. Until now we have no automated way to monitor the crawls and alert us if e.g. the number of harvested bytes changes significantly.

Before we started to collect digital content online we worked in more or less the same way as we had been working for decades. With digital collection this has changed dramatically, forcing us to constantly being able to adapt to new circumstances, new technologies, new important actors, new business models.

Newspaper - digitization and collection of digital copies:

Printed newspapers are still being collected and microfilmed like before. This was not changed in the 2005 legal deposit act. There was no obligation for publishers to deposit the digital files that were made to produce the printed newspapers. However we are now preparing a pilot to collect pdf versions of the printed newspapers, to test to which degree the content is identical to the printed versions, and how well structured and uniform the file formats and metadata can be. To the extent that these files are actually published on the web, they are subject to the legal deposit act as "content distributed in electronic communication networks", provided that there is public access to it, for free or by payment. However we do not want to harvest them from the web, but need to get them in a more structured way, and this requires the cooperation of the newspapers and the distributors, and rely on agreements with them.

Furthermore we have been working hard to digitize our historical legal deposit collection of printed newspapers. With newspaper archives being digitized around the world, expectations rise among both private and academic researchers to be able to access the newspaper content digitally. We have a digitization project running now, ingesting around 1 mio. pages from an external digitization company each month – around 4 mio. files, 30 TB of data each month. The current project is planned to run until the beginning of 2017, producing around 32 mio. pages and a total of around 800 TB data.

Preservation

Preservation of digital content is absolutely crucial for a legal deposit institution building major digital collections. It is a vast topic in itself, and we will only briefly describe the efforts we have been taking to secure our collections. We need of course hardware and software that is robust enough to handle the amount of data that we have. The bit preservation is secured by keeping multiple copies, on different locations, on different technology platforms and controlled by people in different parts of the organization. Part of this bit preservation is using checksums to constantly ensure that the different copies are still identical, and procedures to intervene when they are not. We have developed our software and procedures by taking part in various international projects like PLANETS and SCAPE, and we have been doing our first internal audit, following the guidelines of ISO 16363 (Audit and certification of trustworthy digital repositories), which sets out comprehensive metrics regarding technical and organizational aspects of what is required to be a "trustworthy digital repository". This kind of work is absolutely necessary to keep your digital collections safe. As collection owners we have the responsibility to keep the collections available and usable, and a benchmark based on international best practice is a great help to find out whether what you do is actually enough – before you find out the hard way.

Access and use

The ultimate purpose of building collections is that someone is going to use it - if not today, then at least at some later point in time. The legal deposit act does not specify access permissions, but the copyright act specifies important exceptions to the general copyright protection that makes it possible for us to give users access to legal deposit collections. However the access is mostly granted as on site access, for people visiting the library, and for some media types also some sort of additional online access rights for researchers. Actually on site access is granted for all three institutions handling legal deposit, which means that it also covers The Royal Library in Copenhagen and the Danish Film Institute. For digital legal collections this is important, since we can then use our online portal not only to provide access on our own institutions, but also at the other institutions as well. But if we only granted access on the basic level described in the law, our users would not be very happy. This means that negotiation of additional rights is also necessary.

This has been successful within the radio and television collection. We have built a portal called Mediestream.dk where everybody can search our digital collections online; however due to copyright restrictions users will have to come to the library to actually see and listen to the content. Through negotiation of extended collective licensing we have been able to buy additional rights for students and researchers at universities in Denmark. They can log on to the service and access and use the content online. The universities pay us for this additional online access, which makes it economically feasible for us to pay the licensing costs.

Unfortunately The Netarchive is much more complicated, both technically and legally. This means that all the news content collected online for the last 10 years is pretty hard to use. Technically the challenge is that webpages are not static but dynamic. We have to collect them in a way that makes it possible to reproduce the page as close to the original experience as possible, including the ability to link. Hence we use a format called warc. Until now finding content in The Netarchive is based on typing a specific URL. Keyword or full text search is not possible. We are in the process of indexing the whole archive, making it possible to do full text searches. This is an extremely heavy task requiring expensive hardware working for many months. Full text search requires indexing of the individual parts that together make up a webpage, like text, links, names of pictures, audio and video files, html tags etc. The process has not been completed yet, but we estimate that more than 10 billion objects will be indexed, requiring several servers working in a cluster to build and operate the index.

The Netarchive is a fantastic source, but very hard to get access to. Only researchers, including PhD students, can get online access, and only by application, stating their specific research purpose. Students can get access in connection with their theses, but only on site access, at the library, and they also have to apply in advance, for a specific purpose. Copyright is one part of the challenge, but protection of personal and private data is even more critical. There are scattered personal and private data on the Internet, published on purpose or by mistake. Since our web crawler cannot detect personal and private data, everything is harvested to one archive, and personal data are mixed up with for instance news content. Not only is news content in the same archive as sites with protected personal data, there is also sometimes private and personal data on the news sites, sometimes as part of the editorial content, sometimes as user comments that are also being harvested. On the live internet mistakes can be corrected and personal content may be removed, but in the archive it will remain the same, making it a problem for those affected. Screening of the archived material, both automatically and manually is necessary before we can open it to a wider group of users. We have been testing methods to do this on an automated basis, but it will probably never be possible to do it completely automated. But we are hopeful. The radio and television collection started out being a closed archive too. In the 1980's we just archived recordings on various types of tapes of radio and television programmes – and today you can access streamed radio and television in Mediestream. We all know that internet content has been playing a major role in society for the past 10 years, and we strongly believe that The Netarchive and the content it holds will once become a unique and invaluable source of understanding our time. In time we believe that researchers will be searching and browsing The Netarchive just as they today browse the collections of newspapers, on paper, microfilm and computer. One way forward may be splitting the archive into separate parts. If we had a part that only contained news media, we might be able to argue legally that this content is not covered by the personal data act, but rather by what is called the media responsibility act in Danish. However this is complicated both technically and legally, and with a new EU decree on personal data protection coming up, this is something that we cannot handle right now.

Another kind of use is very strongly on the rise: Digital humanities and related ways of working with digital cultural heritage. Projects around the world proliferate and also in Denmark we see more and more projects being planned, and some that are already in process. One example of this is the Cosound research project. This is research that explores how to search sound archives in new ways, not using manually created metadata, but using algorithms that can help you for instance to find sound clips with a particular voice, feeding an example to the computer, and asking to find radio programs with the same voice. We have a PhD student working with us applying similar techniques on music archives. We get more and more requests to participate in that kind of projects, with people who would like to use our digital collections for various kinds of data mining. This is exciting, but again leaves us with a whole range of unanswered questions about technology, legal issues and economical issues.

We are planning several new activities to try to deal with this kind of challenges. One of them is being planned in cooperation with DeIC, Danish e-Infrastructure Cooperation, which is a national organization working to provide an IT-infrastructure for universities and other research institutions in Denmark. They also work with high performance computing, and we are right now working with them to explore the possibility of establishing what we call a cultural heritage cluster at the State and University Library. This is supposed to become a hardware, software and service offering, making it possible also for researchers in the humanities and social sciences to work with high end computers on the digital cultural heritage that we have in our collections, but also with other content that they bring in from other sources.

Looking forward

The amount of digital content in "electronic networks" is exploding, our collections are expanding rapidly, more and more users want to use it, and our resources are under heavy pressure. Considering the starting point of the 2005 legal deposit act - a world without YouTube, Facebook, Twitter, smartphones and tablets - it is no wonder that with the level of digital activity and innovation we see today, the way we manage digital legal deposit collections is under pressure.

Certainly there is no looking back here - we have to look forward, and find ways to document the online lives of our society. This is as important as ever before. However the political and economical climate is not pointing towards applying more resources to cultural heritage collection. So what is the way forward? We certainly do not have the answer to that question, unfortunately. One direction to look might be to reassess the basic paradigms that we more or less consciously stand on as cultural heritage professionals. Do these paradigms still fit in with the digital age, or do we need to redefine them? We believe in completeness, order, structure, control. This is the traditional way that we keep our collections searchable and usable. But this paradigm may not be equivalent with the amounts of digital content that we are surrounded by today. Two recent discussions in our small department may illustrate the point: We had worried discussions regarding the potential use of pdf-versions of printed newspapers instead of microfilming the paper copies. The worries was that the pdf-files would not be 100% but only perhaps 99.9% identical to the printed newspapers, since sometimes last minute changes are made to one of them. In the same week, in The Netarchive steering group, it was reported that we had now persuaded one of the web hotels that were blocking our crawlers to let us harvest the sites that they are hosting again. This opened up 40.000 domains that had not been collected at all for a long time. It might be time to reconsider both the way we approach the waves of digital content, and adjust the paradigms by which we measure our success.

Selected links and references

State and University Library, English website: http://en.statsbiblioteket.dk/

The Netarchive website: http://netarkivet.dk/in-english/

The Danish legal deposit act in English translation: http://www.kb.dk/en/kb/service/pligtaflevering-ISSN/lov.html

Article in English with more details of how we work with webarchiving: http://netarkivet.dk/wp-content/uploads/Artikel_webarkivering1.pdf

The online portal to our digital collections of radio, television and newspapers (not internet content): http://www.mediestream.dk

About the Cosound research project: http://www.cosound.dk/about-cosound-2/

About the SCAPE project: http://www.scape-project.eu/

About the Open Planets Foundation: http://openpreservation.org/

International Internet Preservation Consortium: http://netpreserve.org/

The warc format: http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml