Submitted on: 11/08/2014

Digitise to Discard: 32 Million Newspaper Pages in Three Years

Niels Bønding
State Media Archive, State and University Library, Aarhus, Denmark.
nieb@statsbiblioteket.dk

Karen Williams
State Media Archive, State and University Library, Aarhus, Denmark.
kw@statsbiblioteket.dk

Gry Vindelev Elstrøm
State Media Archive, State and University Library, Aarhus, Denmark.
gve@statsbiblioteket.dk

Tonny Skovgård Jensen
State Media Archive, State and University Library, Aarhus, Denmark.
tsj@statsbiblioteket.dk

Abstract:

This paper presents a detailed case study of a newspaper digitisation project at the State and University Library, Denmark. The State and University Library is digitising 32 million newspaper pages in three years on a limited budget. The purpose of the project is twofold – digitise to discard one of the two printed newspaper copies preserved at the State and University Library and the Royal Library in Denmark, and enabling users to search and read the digital newspapers online. The digitisation of newspaper pages is performed on the basis of microfilm.

The large scale digitisation project calls for innovative workflows, especially when it comes to quality control. Automatic and manual quality control processes and tools that can deal with 50,000 pages a day, one million pages a month, have been developed. The purpose of the quality control processes is to ensure a sufficient quality to enable the library to discard 32 million newspaper pages from an outdated storage facility. This supports a preservation strategy for newspapers entailing one printed copy, one microfilm copy and one JPEG 2000 file for each page instead of two printed copies and one microfilm copy.

The digital newspaper pages that will be the result of the project will be accessible through the State and University Library's own online portal, Mediestream or through partnerships with newspaper companies.

Keywords: Newspapers, digitisation, access, discarding, library.

---

1. Introduction

This paper is about digitising 32 million newspaper pages in three years. And it is about digitising newspapers with a motive that is not very common in the world of museums, libraries and archives – digitising to discard. To our knowledge there has never been an effort to process this many pages in so little time. It means that this project has conquered – and is still conquering – new land in the digitising area.

The goal of the State and University Library, Denmark[1] is to reach a cruising speed of digitising and post-processing one million newspaper pages a month. All pages are scanned from microfilm negatives and digitised as JPEG 2000 files with accompanying xml-files containing metadata. We need a production process where every element and activity is highly optimised. Obtaining this calls for all the ingenuity that we and our partners at Ninestars Information Technologies Ltd[2] can muster. Otherwise we will not be able to ensure that we uphold the highest possible quality level from start to finish in the chain of production.

This is where we are now: Hardware and software have been trimmed to perfection so they can stand the pressure of handling millions of files on a daily basis, and both the State and University Library and Ninestars have increased the number of resources needed for executing the manual jobs in the process. Will we succeed in reaching and keeping the desired pace of digitising and post-processing 50,000 pages a day? We do not know yet, but we have decided to share our findings so far in this paper.

2. The Basic Facts

The Danish Legal Deposit Law[3] prescribes that two copies of every newspaper printed in Denmark must be delivered to the State and University Library. This collection policy has been in practise since 1916 and by now the collection has accumulated to more than 100 million unique newspaper pages. One of the copies is preserved at the State and University Library in Aarhus[4] and the other copy is preserved at The Royal Library in Copenhagen[5].

---

[1] The State and University Library, Denmark is a public institution under the Danish Ministry of Culture. The library solves a huge range of tasks on behalf of students, researchers, lecturers and other staff members at Aarhus University, as well as public libraries, private companies and individuals - visit en.statsbiblioteket.dk

[2] Ninestars Information Technologies Ltd has years of expertise in digitising significant newspapers such as the Times of India, the South China Morning Post and The New York Times – visit www.ninestar.co.in

[3] The Danish Legal Deposit Law (Lov om pligtaflevering af offentliggjort materiale - www.pligtaflevering.dk/loven/index.htm - has been in effect since 1697. The law currently supports the deposit of all printed matter, radio, TV and the Internet. Read more about legal deposit in Denmark - pligtaflevering.dk/ (in Danish).

[4] The copies become part of The Danish Newspaper Collection, which the State and University Library has a national responsibility to preserve as part of Denmark's national cultural heritage collections - en.statsbiblioteket.dk/national-library-division/the-danish-newspaper-collection

[5] Information about the newspaper collection at the Royal Library in Copenhagen - www.kb.dk/en/nb/samling/ds/aviser

The State and University Library has microfilmed the Danish newspapers continuously from 1976. The microfilm collection covers all in all approximately two thirds of the printed collection. Before 1976 the microfilming was in the hands of the privately held company Minerva. The first microfilms from Minerva date back to 1955, and although microfilms are said to be stable as preservation media for up to 500 years, this requires that they are stored correctly. The microfilms from Minerva were not all stored under optimal conditions before the State and University Library acquired them in 2008. The fact that the condition of the microfilms vary presents the digitisation project with a number of challenges which will also be described in this paper.

The digitisation project has its starting point in the fact that the second printed copy of each newspaper, the copy kept at the Royal Library in Copenhagen, is stored in an old historic and protected building – the West Indian Warehouse which was originally used for storing produce imported to Denmark from the Virgin Islands - on the harbour in Copenhagen[6]. The warehouse is neither fit for preserving newspapers nor for handling loans from the collection but it has nevertheless been used for both activities for many years. The building is in dire need of being renovated and the newspapers need a more stable environment to ensure their preservation, which is why The Royal Library began investigating the range of options for both a couple of years ago.

The obvious choice would be to build a new storage facility for the newspapers, but since this would be very expensive, the stakeholders began working on the idea that the funds could be used for digitising the newspapers instead. This did of course entail that the second printed copy would have to be discarded – something that would go against the original preservation strategy of keeping two printed copies of each newspaper.

Fortunately the discarding of one of the copies was not against the Danish Legal Deposit Law. Changing the law would have required political action, whereas changing the preservation strategy under the law was an administrative decision. The decision of preserving one printed copy, one digital copy and one microfilm copy was made in 2009 by the State and University Library.

What played into the decision of changing the strategy was the fact that in 2007 the State and University Library had opened a newly built climate controlled storage facility on a new location in Skejby near Aarhus. Storing the printed copies of the newspapers in Skejby and the microfilms at the library's main address in Aarhus under optimal conditions would facilitate the viewpoint that the need for storing yet another collection of printed copies in Copenhagen was up for debate.

In 2012 the Danish government approved a special appropriation on the state budget for the digitisation and the discarding of the newspapers in the West Indian Warehouse. Subsequently the State and University Library issued an EU-tender for digitising 32 million newspaper pages from microfilm in three years.

Since the West Indian Warehouse currently holds 32 million pages this has become the magical number in the project. Digitising 32 million newspaper pages from microfilm and funding the project by discarding the same number of physical pages will both be a step towards modernising the preservation strategy for newspapers as well as greatly improving the users' access to the Danish newspaper heritage.

[6] Description of The West Indian Warehouse - www.kulturarv.dk/1001fortaellinger/en_GB/the-west-indian-warehouse-toldbodgade-40-copenhagen (in English)

3. Difficult Decisions

In order to make the number of pages meet the numbers on the budget we had to make some serious decisions. The first one was to let go of the dream of digitising the printed copies of the newspapers. Using the original and printed copies would potentially yield the highest picture quality but neither the timeframe nor the budget would let us anywhere near this option. Therefore we decided to use microfilm as the basis for digitisation, but since we were still pursuing the highest possible quality we opted for using our microfilm negatives instead of the microfilm positives.

The decision to digitise from microfilm poses a series of challenges concerning the actual microfilms: First of all, some of the microfilms are not in mint condition due to having been stored under poor conditions for a number of years. Secondly, throughout the years the microfilming was not always done in the most optimal way causing the microfilm pictures to be out of focus, for example. Thirdly, sometimes the printed newspaper pages were in poor condition when microfilmed, for example torn, folded or in other ways damaged.

Based on our accumulated knowledge of the microfilm collection plus a series of spot checks done on the collection, we chose to meet these challenges by not meeting them – that is, we will digitise all our microfilms of a chosen newspaper title even though we know that the quality of the microfilms will vary. Instead we will deal with the poor quality after the digitisation process has been completed. The actual processes involved in discarding the newspapers have yet to be fully decided upon, but we have already decided that digital images which reveal themselves to be of very poor quality will not result in the printed copies being discarded.

Even though we chose this somewhat easier way to reach our goal it still means that we have to process more than 30,000 microfilm reels in three years. As already stated, not all the microfilms are in mint condition, but instead of going through all the microfilms ourselves, we decided to include the examination of the microfilm in the requirements to the supplier. Since the supplier has to handle all the microfilms anyway this means that they will only have to be handled once instead of twice. This again is a time-saving decision.

So we are sending more than 30,000 unexamined microfilm reels to Ninestars for evaluation and scanning. Moreover, we have not included an option in the contract which allows us to leave out any of the microfilms no matter how poor their quality may be. We decided to digitise every microfilm, since it was our judgement that the overall quality was quite good. However, we do have to make sure that the quality of each individual microfilm is recorded, thus allowing us to locate the problematic microfilms later on when the time comes to discard the second printed copy of the newspapers. Information registered by Ninestars includes the minimum and maximum levels of emulsion density on every single scanned microfilm.

The final area where we had to make a difficult decision concerns the manual quality control of the result of the digitisation process - the digital images. Although we have automated as much as possible of the quality control, there are still things that can only be detected by having a skilled and trained person looking at the images. With 50,000 images/pages being delivered every day it is obviously impossible to manually look at all of them, and with the resources we have at our disposal even looking at 5% of the images is an overwhelming number of pages. So instead we have settled for a sampling rate that gives us no less than five random pages/images per microfilm. With microfilms containing 500-1,500 pages this is not a huge number, but it is important to notice, that it is a minimum number of pages and it can be increased, if our knowledge of the collection requires it.

4. What Quality for what Purpose?

A distinguishing feature of the project is the constraints put on the production process which originates in the way the project is financed. In order to be able to discard the newspapers in the future, we have had to adjust our overall strategy for preserving newspapers. The strategy now states the following:

- One printed copy is preserved under optimal and climate controlled conditions at our facility in Skejby.
- One microfilm copy is preserved in a secure and climate controlled facility in the basement under the State and University Library.
- One digital copy is preserved in alignment with the State and University Library's strategy for digital preservation[7], which actually means that two digital copies are kept at geographically separated locations.

An inspection of the collection of newspapers in the West Indian Warehouse revealed that not all 32 million newspaper pages can be disposed of right away. Some newspapers are simply too rare or too unique to discard. Other newspapers are missing in the collection of printed copies in Skejby. If possible, identified holes in the Skejby collection will be filled with the printed copies from the warehouse, so the collection in Skejby will be as complete as possible. If a newspaper has not been microfilmed, it will not be discarded either. Newspapers fulfilling the criteria of already being present in the collection in Skejby, having been microfilmed and not being very rare or unique will be discarded, if the quality of the digital copies turns out to be acceptable.

A fundamental challenge, when having digitised the newspaper pages that meet the criteria of already being present in Skejby and having been microfilmed, is to determine if the quality of the digital copy is adequate grounds for discarding the printed copy.

The first step to meet this challenge is to ensure that the digital copy is produced in the best possible quality. In order to do so we have established a control process consisting of three steps:

1. Automatic control at the supplier site, that is: Ninestars' facilities
2. Automatic control at the State and University Library
3. Manual control at the State and University Library

4.1. Automatic Quality Control

The automatic quality control which takes place at Ninestars' facilities is almost identical to the automatic quality control we perform in Aarhus when the files arrive in our systems. It was part of the requirements in the tender that the supplier should run the first automatic control on location in order to save valuable time. The automatic quality control carried out by Ninestars and the automatic quality control done by the State and University Library in Aarhus are performed using the same quality control software developed by the IT department at the library.

[7] *Digital Preservation Strategy for the State and University Library, Denmark* - en.statsbiblioteket.dk/about-the-library/dpstrategi and *Digital Preservation Policy for the State and University Library, Denmark* - en.statsbiblioteket.dk/about-the-library/ddpolicy

The automatic quality control process carried out by Ninestars is supposed to ensure that when the files arrive in Aarhus ideally there are no errors and inconsistencies left. But in reality this is of course not the case. For example: Ninestars imports data about the microfilms from our database in order to connect various metadata to the digital files, and we use the same database when we check that the delivery is correct. What we did not expect was that sometimes the date interval written on the cardboard box, which holds the microfilm reels, does not reflect what is actually on the microfilms, and therefore the information does not reflect correctly the publication dates which are actually included in the delivery. We have had to introduce a special process to handle this scenario and right now it is a completely manual process since we do not know to what extent this inconsistency will occur. Had we done a very thorough in-house check of the microfilm reels before shipment, it would have given us the chance to totally avoid this scenario. However, when we decided not to check the microfilms before shipment we did not even know that this discrepancy existed in our microfilm collection, even though we did know the quality of our microfilms varied greatly.

A number of different aspects of the files are checked automatically. First we concentrate on the file structure: Is the shipment of digital files intact on arrival in the systems at the State and University Library? If it is, then we continue with a series of checks on metadata:

1. We compare the received data with our own data – e.g. is the title of the newspaper correct.
2. We perform logical checks – e.g. does the start date lie before the end date.
3. We perform checks on the data formats – e.g. is the date format valid.
4. We check that the files adhere to the standards of the JPEG 2000 format.
5. We check that the pictures have the characteristics which we and Ninestars have agreed upon – e.g. no lossy compression.

Apart from what we have included in the automatic quality control process, we also have a feature which we have labelled "flag for manual control". The feature points to "unusual incidents" which require closer scrutiny – for example if the histograms of a series of pages are anomalous, or if the OCR percentage of correctness suddenly drops. How unusual these incidents are allowed to be is a trade-off between a chosen level of certainty and the number of human resources available to investigate the incidents. We have yet to conclude on the degree of unusualness, and the adjustment of this feature will be an ongoing task for some time to come.

4.2. Manual Quality Control

When the automatic quality control process has ended successfully, the files are transferred to the tool used for manual quality check. The tool has been developed by Ninestars specifically for our project. We use the tool to perform the manual quality checks on a sample of pages from each microfilm. We have consulted ISO 2859[8], so we have a fairly good idea of how to reach the desired level of certainty, but in the end it is the budget (that is the number of available human resources) that decides how many pages we are able to check. As previously stated, we have set the limit to five pages per microfilm (which contains 500-1,500 pages). The tool automatically picks five pages, but the flags from the automatic quality control indicating “unusual incidents” must be added to this number. If we are not careful manual quality control will increase and exceed our limits and capabilities. We have to adjust the process as we go along in order to use our resources in the most rational way.

[8] "DS/ISO 2859 - Sampling procedures for inspection by attributes, Parts 1, 2, 3, 4 and 10", Copenhagen 1992-2011.

4.3. Quality and Discarding

After performing both automatic and manual quality control as described we are convinced that the digital output is the best possible result given the quality of the microfilm. As a minimum we will have a digital representation of our microfilm negatives. In the end it is the quality of the microfilm negatives which determine the quality of the digital images we can get from the digitisation process. The quality control processes are designed to ensure that we get the best possible result while at the same time overseeing the enrichment of the digital pictures with metadata and OCR.

If this project was not financed with the intention of discarding one of the printed copies, we would be quite happy to produce dissemination copies, put them online and invite our users to help us identify the problematic time spans and unreadable pages on the microfilms, so we could adjust the information or replace the pages with new and improved copies based on a new microfilm or maybe even a copy created by digitising the printed newspaper. But we are digitising to discard and because of that we need to identify the problematic time spans and the pages of poor quality ourselves before the discarding process begins.

This activity is done partly by recording the values of emulsion density on the microfilms, which inform us about the physical quality of the film, and partly by recording the findings of both the manual and the automatic quality control processes. In regards to the manual check it is the source of the poor quality of the digital image that is interesting, and in particular to identify what caused the poor quality. We will record the problematic microfilms, and if possible determine if the problem stems from the microfilm or the original printed newspaper.

In the alto-files accompanying every JPEG 2000 the ABBYY Finereader, which is the tool Ninestars uses for OCR recognition, states its own recognition success in percentages for every page. When we accumulate and process the statistics on these scores we expect to be able to identify patterns that can guide us towards microfilms and newspapers of poor quality.

Finally, and most importantly, we have employees at the library who have accumulated in-depth knowledge about the collections of printed newspapers and microfilms over the years. They already know where a lot of the problems are, so when we add the meticulous registrations of our manual control and the statistics of the automatic control to their knowledge, we will have the best possible source material for fulfilling the purpose of the project – being able to discard 32 million newspaper pages.

5. Adjusting for Mass Production

Signing the contract with Ninestars meant that three new locations became central to the project:

1. Ninestars picks up the microfilms in Aarhus and transports them to **Hamburg, Germany** where the scanning is done.
2. The scanned pages receive post-processing (for example splitting, cropping and de-skewing) in **Chennai, India.**
3. Ninestars' IT-team in **Bangalore** is responsible for the processes of OCR and segmentation, and for developing the tool used for manual quality control in Aarhus.

The fact that Ninestars has an office in Hamburg which can perform the actual scanning of the microfilms was very fortunate for the State and University Library since we would have been

hesitant to ship our microfilm negatives to India due to the long transport distance from Denmark.

Now the microfilms are registered in our shipping system, barcoded and packaged safely, before they are shipped in aluminium boxes to Hamburg where they are scanned. The digital files are then accessed by Ninestars' staff in Chennai and Bangalore to undergo post-processing, OCR and segmentation. After that the digital files are transferred to the library and when the quality controls have approved the files, the microfilms are returned to Aarhus. On their return the microfilms are registered in the shipping system once again.

In order to reach one million newspaper pages a month it is crucial that we can continuously keep up the speed in all parts of the processes. We rely on a lot of hardware and software and we will have these processes monitored more or less 24/7, but the manual processes must also be top-tuned. As Mohan Doshi from Ninestars stated when a delegation from the library visited Ninestars in Chennai in February this year: "It is important to eliminate subjectivity. If the people who do the manual routines have no objective rules to go by, we will never reach one million pages a month."

A good example here is the B7-option of our contract which states that, "The Supplier shall include the section label in the page MODS file." The question is then: "What is a section label, and how do we express this unambiguously so that non-Danes can detect the section label in any Danish newspaper?"

The short answer is: We do not – at least not if we want the definition to coincide with the use and history of the concept in relation to Danish newspapers. Instead we had to settle for this narrow but operational definition of what a section is:

1. A section always starts with page number 1.
2. When a page number 1 is present, it indicates the beginning of a new section.

On the one hand, sections have historically been used in newspapers without separation, allowing the sports section to start on page 16 and finance on page 20, for example. On the other hand, we would not call all the physically separated parts of a newspaper sections – some of the parts are actually inserts with advertisements and are thus not actual sections in the newspaper.

The case for production is that all the processes have been trimmed repeatedly and we have started to increase the weekly production keeping a firm eye on the robustness of the production chain. Reaching a stable production at the desired speed will come in handy since it will leave us some time for all the rest of the tasks connected to this project. Tasks such as providing access, clearing rights and establishing partnerships with the Danish newspaper companies.

6. Access

An essential part of the digitisation project is to digitise to discard. The other essential part of the project is to be able to give access to a broader range of users than is possible now with lending printed newspaper copies or microfilm to our users.

We plan to give access to the digitised newspapers through an online portal. The library has already developed its own portal holding the library's collections of digital broadcasted radio

and TV and commercials. This online portal, Mediestream[9], will be expanded with a newspaper section. Adding the newspaper section to an already existing solution means that the users can focus their search for information on the collection of newspapers, but they can also search across all the collections in Mediestream. In the latter scenario the users will benefit from the synergy generated from having access to different media types.

Over the next few months different features favouring the characteristics of digital newspapers will be developed for Mediestream. Since the newspapers will have undergone OCR and segmentation processes, we are going to develop a full search index so the users can search through the entire corpus by using keywords. It will also be possible to search for a specific newspaper title and limit the search by date or a date range.

When the user has found the newspaper he or she was looking for the user will be able to flick through the pages of the rest of that particular newspaper. The user will be able to zoom, make a pdf-file, have a persistent identifier on page level etc. Metadata will be displayed next to each page. In time an entry point consisting of information about the specific newspaper titles and a calendar showing the dates for which newspapers have been digitised will be added. We expect to open our online newspaper portal in the beginning of 2015.

7. The Danish Copyright Law

We would really love to give free public online access to the newspaper collection but as in most countries around the world, Denmark too has strict legal restrictions on what the users can get free access to. In principle the Danish copyright law[10] prohibits us from giving free access to newspapers newer than 1875.

Negotiations with the Danish Copyright Holder Organisation (Copydan)[11] are on-going, and we hope that by paying a fee we will be able to give free access to newspapers dated as recently as 1920. If a user shows up at the State and University Library, Aarhus, The Royal Library in Copenhagen and the Danish Film Institute in Copenhagen[12], he or she will have free access to the complete collection.

People living far away from the three institutions will not necessarily find this option convenient. However, our collection of Danish newspapers is not the only collection of newspapers in Denmark. A significant number of local history archives and public libraries in Denmark also have collections of newspapers – either in printed form or on microfilm. This means that the users can use Mediestream to search for relevant newspapers, identify them and then turn to their local history archive or public library for actual access, if it's not possible to read the newspaper in Mediestream. The State and Library's collection of newspapers may be the most complete collection but is by far not the only one. The sheer fact that the newspapers become fully searchable is an enormous leap in increasing access to the contents – whether or not the users end up actually reading our exact copies of the newspapers or somebody else's is of minor importance. Providing the search tool to the newspapers is a great accomplishment of the project.

[9] Mediestream – www.statsbiblioteket.dk/mediestream

[10] The Danish Copyright Law (Bekendtgørelse af lov om ophavsret) (in Danish) www.retsinformation.dk/Forms/r0710.aspx?id=129901

[11] Copydan – Kunst, Viden og Underholdning www.copydan.dk/?lang=UK

[12] The Danish Film Institute - www.dfi.dk/Service/English.aspx

8. Partnerships with Danish Newspapers

Right from the beginning of the project we decided that we wanted to cooperate with the Danish newspaper companies[13]. We establish partnerships with the newspaper companies with the intention of selling them a digital archive copy of their newspapers currently in print or on microfilm and in return we receive a copy of their digital archive of newspapers, if they have one. The money generated from the sales will be used to digitise more newspapers and will hopefully also allow us to digitise newspapers which are to be considered orphan works because they no longer have a copyright holder who can be identified. The money is also spent on OCR and segmentation which increase the value of the search options in Mediestream.

Acquiring the digital archives from the newspapers is very valuable to us since the newspapers currently in print are not deposited with us in digital form. On top of this, we hope that we will be able to save time and money by not digitising newspapers which are already available in digital form.

The gain for the newspaper companies is that they can give their readers access to a complete historical archive as part of a subscription for the newspaper, for example.

9. Picking Newspapers for Digitisation

Since one of the goals of the project is to establish partnerships with the Danish newspaper companies, it quickly became clear that part of the policy of choosing titles for digitisation would have to favour the newspapers which had signed a contract by placing them first in line for digitisation. However, since the State and University Library is an institution owned by the Danish state and with obligations to the Danish public, we cannot allow money to make all the decisions when it comes to choosing what titles should be digitised during the project. Of the 991 newspaper titles in the Danish Newspaper Collection, the library has identified approximately 130 newspaper titles which are candidates for undergoing digitisation[14]. On the list are the newspapers currently in print but also newspapers which have been chosen in order to guarantee the geographical and political diversity of the digital collection. We have also listed newspaper titles which cover important historical periods of the history of Denmark.

10. The Future

The first version of Mediestream Newspapers will not have facilities for correcting the OCR-generated text from the newspapers, but in the future we would like to include this crowdsourcing feature. Other crowdsourcing options would be to invite the users to correct certain parts of the metadata or help us tag the photos in the newspapers, just to name a few. We know with certainty that not all 32 million pages will be perfect. We will receive images of poor quality. However, images that are unreadable by a machine are not necessarily unreadable by a human being, and we hope to be able to enthuse some of the many talented people in Denmark to help us increase the value of the collection even more by refining it. We have not yet had a chance to investigate the area and options of crowdsourcing thoroughly,

[13] Partnership with Danish newspapers – including the public list of newspapers which have signed a contract with the library - en.statsbiblioteket.dk/national-library-division/newspaper-digitisation/invitation-to-partnership-with-danish-newspapers

[14] Newspaper titles and description of the criteria used for identifying them as candidates likely to be digitised - blog.avisdigitalisering.dk/avistitler/ (in Danish).

but we hope to be able to do so in near future, since inviting our users to help us improve our collection is an obvious road to take with a project like ours[15].

The vision of the State and University Library is that our users should be able to discover everything in our collections online in 2020. This is an ambitious goal, and it means that we will keep trying to find funding for digitising all our collections, including the remaining two thirds of our newspaper collection. Since one of the goals of the project is to build and run a digitisation factory, it would be expedient to keep the flow going once the three years are over and the first 32 million pages have been digitised. By continuing to digitise we will make the most of our investment in software and workflow development as well as the accumulated and fine-tuned skills. The road to funding this vision is paved with the publicity we get when we give the Danish people online access to the newspapers in Mediestream and invite them to dive into the history of Denmark and the many stories the reported by the Danish newspapers since 1668.

[15] Visit en.statsbiblioteket.dk/newspaper-digitisation for more information about the project and follow the project on the blog www.avisdigitalisering.dk and on Twitter @Avisdigital and @Statsbib.