DOI: 10.2352/issn.2168-3204.2016.1.0.122

Copies and Originals, Preservation and Access: The Art of Balance in a Digital World

Niels Bønding; Project Manager; Gry Vindelev Elstrøm; Digital Collections Manager; Karen Williams; Project Manager; State and University Library, Aarhus, Denmark

Abstract

Since the Legal Deposit Act of 1925, The Danish Newspaper Collection has kept two physical copies of each newspaper published in Denmark along with a microfilm copy as backup. Now, The State and University Library, Denmark is changing its preservation strategy for the newspaper collection. One reason for the change is our newspaper digitisation project, which will produce 32 million digital pages in three years. We pay for the digitisation by discarding one of the paper originals, thus saving on storage expenses, but at the same time, our goal is to acquire a digital copy that will take the place of one of the paper copies.

Another important change has to do with the tradition of microfilming the newspapers. We are moving away from microfilming and into the digital realm, where we want to receive the daily e-paper from the publishers and use this as a backup copy instead of the copy on microfilm. However, the question is if the e-paper can serve as backup for the physical paper when we cannot be 100 % certain the two newspapers are identical. We have seen examples where the two editions do not contain the exact same information. If that is a general occurrence, will the e-paper then suffice as a copy?

This paper is not telling the final story. We are in the middle of this change and the paper attempts to capture the state of affairs of newspaper preservation in Denmark post-"the good old days" when preservation and dissemination were both done on paper and celluloid.

The Newspaper Collection

When legal deposit of Danish newspapers originally was introduced in 1781, discarding newspapers of lesser value was considered a possibility. In 1915 when the law establishing the Danish Newspaper Collection was in the making, the Danish Home Secretary Ove Rode remarked that "it is not entirely unfortunate, that much of the paper whereon the newspapers are printed are of such a condition that it does not last forever. If it did we would surely see the time where it would be impossible to store the large quantities of newspapers." [1] This remark followed the mind-set of The Head Librarian of The State and University Library, Vilhelm Grundtvig, who in 1910 had written about "the newspaper nuisance". He thought that the sheer volume of the Danish newspapers could easily end up being a dead weight diverting time and space from more urgent tasks. [2] Until 1925, four copies of each newspaper were deposited. In 1925, a change in the legal deposit law reduced that number to the two copies we still receive today.

In spite of being not always highly valued or treated with affection, most of the newspapers that have been deposited throughout the years are still with us. Only few people dispute that the newspapers are an important source of knowledge and key to understanding life in Denmark through the centuries. The printed newspaper copies are kept on separate locations. One copy is kept at The Royal Library in Copenhagen and the other stays in Aarhus at The State and University Library where The Danish Newspaper Collection has been situated since it opened in 1918.

Preserving all these newspapers costs money and like all other tasks in the public domain, it is not immune to scrutiny with respect to costs and benefits. In the end, Ove Rode's question of space was the trigger for the Danish Newspaper Digitisation Project. Many of the newspapers kept in Copenhagen were (and most still are) stored in an old historic building named 'The West Indian Warehouse'. Originally used for storing produce imported to Denmark from the Virgin Islands, it is located at the harbour in Copenhagen [3]. The building is in dire need of restoration and the newspapers need a more stable environment to ensure their long-term preservation. This is why The Royal Library began investigating the range of options for changing the preservation set-up.

The traditional choice would have been to build a new storage facility for the newspapers, but since this is very expensive, the stakeholders began working on the idea that the funds could be used for digitising the newspapers instead. This entailed that the second printed copy would be discarded. In principle, discarding a newspaper copy would go against the original preservation strategy of keeping two printed copies of each newspaper.

In 2007, the State and University Library opened a newly built climate-controlled storage facility on a new location in Skejby near Aarhus. The building holds the Danish Newspaper Collection. Storing one copy of the printed newspapers in Skejby and the microfilms at the library's main address in Aarhus under optimal conditions facilitated the viewpoint that the need for storing yet another copy of the printed newspapers in Copenhagen was up for debate.

The decision to preserve one printed copy, one digital copy and one microfilm copy was made in 2009 by The Danish Ministry of Culture, The Royal Library and The State and University Library.

The Digitisation Project

An analysis of the possible ways of digitising the newspapers was initiated. The State and University Library drew on the experiences of digitisation projects in Australia, USA, Sweden and Holland as well as the current digitisation prices on the market. From the early stages of the analysis it was obvious that digitisation from paper was quite expensive. The price per page was many times higher than digitising from microfilm. Since the budget was low, digitisation had to be based on microfilm but with the amendment that digitising the paper copies was an option in cases where the microfilms were particularly bad and/or the newspapers were particularly valuable with respect to preservation and/or dissemination.

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. Since

DOI: 10.2352/issn.2168-3204.2016.1.0.122

the West Indian Warehouse holds 32 million pages this became the magical number in the project. Subsequently, the State and University Library issued an EU-tender for digitising 32 million newspaper pages from microfilm. Because time was an important factor, we chose to opt for a project where the 32 million pages would be digitised within three years.

Microfilm Choices

After deciding to digitise from microfilm, we still had to determine how we could ensure the highest possible quality: we had to choose between using the positive or the negative copies of our microfilms. This was a choice between the unused and potentially best negative copy, and the more expendable second generation positive copy. In the end, it was a choice between risking our negatives on one side, and getting lower quality from our (often used and scratched) positives on the other. We chose to use our negatives but we made sure to balance the risk in the tender so that the supplier would have to insure the microfilms for an amount that equaled our expenses of creating a new microfilm from the paper copies.

Our collection of negative microfilms is not at all uniform. Covering a timespan from the mid-1950s up to today, with the filming done by a private company for the first many years, there is a great variety in both film quality and filming procedures. First, some of the microfilms are not in mint condition due to poor storage conditions for a number of years. Secondly, throughout the years the microfilming was not always done in the best possible way. In some cases this has resulted in e.g. that the microfilm pictures are out of focus. Thirdly, sometimes the printed newspaper pages were in poor condition when microfilmed, for example torn, folded or in other ways damaged.

This left us with the question of whether to examine the microfilms before sending them off for digitisation, or leave the examination to the supplier. Based on our accumulated knowledge of the microfilm collection plus a series of spot checks, we chose to include the examination of the microfilms in the tender. We found it timesaving to include this in the vendor's workflow as opposed to creating a separate workflow ourselves. The reason we could make this decision was that we found the overall quality of our collection of negative microfilm to be quite good.

Automatic and Manual Quality Control

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

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 scrutiny – for example if the pages are bigger or smaller than expected or if the OCR percentage drops below a certain point. 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.

Another trade-off is found in the manual quality control of 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 delivered every day, it is impossible to manually check all of them. 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. Since each microfilm contains 500-1,500 pages this is not a huge number, but it is important to notice, that it is a minimum number of pages. In several cases, we have found reason to check more than five pages and we will continue to do so when necessary.

Project Status

Project status as of February 2016 is that approximately 19 million pages have been digitised by our supplier Ninestars, and out of these 11 million pages are available in our online portal, Mediestream [4]. We have yet to discard any newspapers, but the processes are in place and we have completed our analysis of a handful of carefully selected titles so in principle we are ready to begin discarding.

Therefore, for 18 million pages we now have:

- One printed copy preserved under optimal and climate-controlled conditions.
- One microfilm copy preserved in a secure and climate-controlled facility separate from the location of the printed copy.
- One digital copy preserved in alignment with the State and University Library's strategy for digital preservation [5], which actually means that two digital copies are kept at geographically separated locations.

Soon this will be the situation for all 32 million pages. But only for 32 million of the 74 million pages that our newspaper collection consists of. We still have 42 million pages and the continuous and daily accession of new newspapers to take care of.

Continuous Accession of the Danish Newspaper Collection

Alongside the digitisation project, the Danish Newspaper Collection still receives newspapers and microfilms them in order to have a back-up copy. This has been the practice for many years, but the microfilming has become increasingly expensive. The number of suppliers of equipment and microfilm is declining while at the same time the prices are going up. In addition, where the microfilms in the beginning had both a preservation purpose and a dissemination purpose, the latter is now surpassed by the potential of the digital copy. However, there might still be some good, sound preservation arguments for continuing to microfilm the newspapers. The question at hand is now whether born digital e-papers can be as sufficient a backup-copy as a microfilm copy of the paper original.

A Case Study on E-papers

When we started to discuss replacing microfilm copies with e-papers as a serious option, we quickly became interested in investigating whether there are differences between the e-paper and the newspaper. One of the interesting examples from our investigation arose from the 2014 FIFA World Cup in Brazil where the final game was played in the evening Danish time. The two pages below are page 1 and 23 of the e-paper of 'Morgenavisen Jyllands-Posten', 14th July 2014:

DOI: 10.2352/issn.2168-3204.2016.1.0.122

The pages above show no sign of the final game being played because the match had not ended by the time the e-paper was finalized for publishing.

Quite the contrary was the case with the printed newspaper that landed on the doormat of many Danish households the next morning, as is seen below:

This is page 1 and 23 of the printed newspaper from the same day. One would think that the e-paper would contain the latest news,

DOI: 10.2352/issn.2168-3204.2016.1.0.122

but the publisher did not find it necessary to issue a new e-paper edition containing the information on Germany's 1-0 victory over Argentina. The story seen here on page 23 appeared on the website instead – and with a different picture.

In this case, the e-paper cannot serve as an exact back-up copy of the original newspaper. It can even be argued that the original is the e-paper, since it was the first one published, but the publisher categorizes them both as first editions. This means that the conclusion could just as well be that this is the case of two different editions, each deserving a unique place in our collection.

Missing Contents

Another issue when it comes to e-papers is that we have seen many cases where ads seen in the printed newspaper are not present in the e-paper. This is especially the case for e-papers issued at the beginning of the 21st century where the ads were sent to the printer after the e-paper had been made. One could argue that the ads are of minor importance compared to the editorial material. While that might be true for some people, to us the ads are not without importance since they are also considered part of our cultural heritage, and we would rather not have to accept a back-up copy without them. Fortunately, this lack of ads has not been a problem in recent years. The printing workflow seems to have undergone changes in recent years so the ads are included in both the printed newspaper and the e-paper.

Where does this leave us with respect to using the e-paper as a back-up copy? Well, first we can conclude that the two copies are not always identical and because of this, we can hardly claim one is a copy of the other. The question is, if we benefit enough from the e-paper to want it anyway.

On the other hand and quite importantly, the e-paper provides us with pictures in colors. The microfilm copy is all black and white and because of this, it is not a 100% copy of the original newspaper either. However, in regards to the contents of the newspaper, we must say that the microfilm copy comes very close. Here the e-paper is more unstable. On the other, the text contents present in the e-paper are in return 100% digitally accessible from the beginning. This could be very important in assessing the value of shifting from microfilm to e-paper. However, in order to reap the benefits we would have to accept the downsides: lack of ads and differences in editorial content. Can we live with these?

The lack of ads is to the best of our knowledge a solved problem. Newer e-papers do not demonstrate this shortcoming, but that is not the same as saying the problem will never return. In all cases, we must check the material we receive from the publishers – random sampling as a minimum –, which we would do any way.

Tracking Changes

The issue of editorial content is a different matter. The differences here are so infrequent that random sampling would have to be quite extensive in order to give a qualified evaluation of the material. Moreover, what good would it do? We might acquire information of yet another event that took place round midnight and thus resulted in two different editions of the newspaper. But with the e-paper received we would have both of them preserved, and as long as the production of e-paper and printed newspaper are closely linked, the differences will be small.

It is important to keep a close eye on the newspaper publishers. For now, the e-paper is presenting itself as a flat printable pdf-file, but it might not stay that way for very long. As has been predicted many times, the paper editions are facing extinction in a few years' time and that could in effect help to "set the e-paper free" since it is currently made in the image of the traditional paper edition.

Investigating E-paper versus Printed Newspaper

Apart from these two known – and sizeable - problems, there might be all sorts of other issues when switching one of our preservation copies from microfilm to e-paper. Thus, in order to get a better idea of this we decided to launch a pilot project, which focused on receiving the daily e-papers from four different publishers for three months.

The pilot project was completed during autumn of 2015, and the goal was twofold:

1. To make it clear if the digital versions of the newspapers could replace microfilm as back-up copies of our legal deposit newspapers.
2. To gain experience on the characteristics of the files in order to estimate the task of ingesting them in our digital repository.

In order to reach the first goal, we compared the e-paper and the newspaper for each edition. Not down to the individual articles, but on the level of sections, inserts and number of pages. When the three months had passed, we concluded that the two copies were not completely identical, but when it came to the parts of the newspapers we usually preserve on microfilm [6], nothing was missing.

Unfortunately, the examination of the files also revealed that our IT-department would have to do more work than we had hoped for in order to set up a permanent workflow. The span of different pdf-profiles in this small section of titles alone was quite large, and there is no reason to expect a lower degree of heterogeneity when introducing the full set of 40-50 titles currently in print. This means that we will have to wait a while to start collecting the e-papers, but when we begin, we will have taken another step away from the old preservation strategy.

For the newspapers of the future, we will then store:

- One printed copy preserved under optimal and climate-controlled conditions.
- One digital copy preserved in alignment with the State and University Library's strategy for digital preservation, which actually means that two digital copies are kept at geographically separated locations.

For these newspapers, we will not produce a microfilm copy. We are still discussing this in relation to the safety of the collection, and while it is difficult to imagine us creating new microfilms, this fact plays into the discussion of keeping a third digital copy far away from the other two, which are both situated in the Aarhus area not more than a few kilometers apart. Related to this discussion is also the fact that when the 32 million pages are discarded, the remaining physical copy and the microfilm negatives will also be at the same two locations in Aarhus.

The Patchwork of Completion

This is where we are today. Closing in on having 32 million pages digitised and taking care of the shift from microfilm to e-paper when it comes to back-up copies of the daily legal deposit of titles. However, the Danish Newspaper Collection consists of 74 million pages. What about the rest?

DOI: 10.2352/issn.2168-3204.2016.1.0.122

Of the remaining 42 million pages, 16 million have been microfilmed but 22 million are only preserved in paper format. In order to digitise these two chunks we have to find the money for two new largescale digitisation projects. This is of course something we would like to do, but at present, our focus also lies with a smaller portion of newspapers.

When the funding related to the West Indian Warehouse became a reality, we had already explored the possibilities of creating a digital collection of newspapers quite extensively. The business idea that came to supplement the digitisation was to cooperate with the Danish newspaper companies [7]. We sell a digital archive copy of their old newspapers to the newspaper companies, and part of our contract is that we receive a copy of their digital archive in return. We will then have a clean cut between our digitised newspapers and their digital born newspapers. Our challenge is the gap at the other end. Because the exchange of archives has already been done for a number of titles, and we have not yet started to receive the e-papers, we will – in spite of the contracts – still have a gap to cover.

A second challenge lies hidden in the historical newspapers of the digitisation project. We have from the beginning chosen to focus on the titles we have microfilmed in total. However, as we are working our way through the collection we find ourselves digitising more and more titles where not all editions have been microfilmed. In some cases, we have actually chosen to create these missing microfilms and then digitising them afterwards, but since we now have suspended microfilming all together – 2015 will be the last year the library microfilms newspapers – this is no longer feasible. For the purpose of retro-digitisation – as opposed to retro-filming like we used to – we have bought our first newspaper scanner, but the workflow from scanner to digital repository has yet to be established.

Nature abhors a vacuum, Aristotle said, and likewise a true archivist abhors an incomplete collection. Sometimes we can do nothing about this, but sometimes we can, even if the efforts result in a patchwork of different formats. In the case of backup copies, this completeness is very much wanted for preservation reasons, but the users of the collection are also stakeholders in the quest for adding the last pages to the searchable body of newspapers in our repository. The primary challenge is finding the resources to make this come true.

References

[1] Svend Larsen, "Pligtaflevering af aviser," in Den trykte kulturarv. Pligtaflevering gennem 300 år, København 1998, p. 195-222

[2] Svend Larsen, "Statens Avissamling i Aarhus," in Bibliotekshistorie 5, København 1999, p. 99-124

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

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

[5] 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 (both in English)

[6] We do not microfilm e.g. advertisement inserts.

[7] 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

Author Biography

Niels Bønding, Project Manager at the State and University Library, Denmark has managed the Danish Newspaper Digitization Project for the past two years. He holds a Master's degree in Philosophy, has an IPMA C certificate in project management and is a certified PRINCE2 Practitioner.

Gry Vindelev Elstrøm is Digital Collections Manager at the State and University Library, managing the long-term preservation of digital cultural heritage collections at the library. She holds a Master's degree in Nordic Studies and History.

Karen Williams works as Project Manager at the State and University Library, Denmark. She holds a Master's Degree in English and Nordic Studies and works with digitization and dissemination of the library's analogue and digital collections.

