PRESERVATION OF DIGITISED BOOKS IN A LIBRARY CONTEXT

Eld Zierau **Claus Jensen**
The Royal Library of Denmark
Dep. of Digital Preservation
P.O.BOX 2149
1016 Copenhagen K, Denmark

ABSTRACT

The focus of this paper is on which digital objects to preserve when preserving digital library materials derived from original paper materials. It will investigate preservation strategies for digital objects from digitised paper material that must both be preserved and simultaneously retain a short route to dissemination. The investigation is based on a study of digitisation done a decade ago and digitisation done today.

In the last decade mass digitisation has become more commonly used since technological evolution has made it cheaper and quicker. The paper explores whether there are parts of digital material digitised a decade ago worth preserving, or whether a re-digitisation via mass digitisation today can create a relevant alternative.

The results presented show that the old digitised objects are worth preserving, although new digitisation can contribute additional information. A supplementary result is that investment in digitisation can mean lower costs in the long term. Manual adjustments for the image processing can result in considerably smaller images than images made in cheap mass digitisation. Although initial manual work is more expensive, the storage and bit preservation expenses are lower over a long period.

1 INTRODUCTION

This paper explores digital preservation of digitised material in a university and national library context, where there is a close relation between preservation and dissemination of digital material. The study is a result of a research project at the Royal Library of Denmark where the goal is updated preservation strategies for the library. It uses the Archive of Danish Literature (ADL) as a case study, which is a web based framework built at the start of the century. ADL is mostly limited to books, book metadata and book collections. In order to analyse the difference between 2000 and 2010, we have done experiments on re-digitising books from ADL.

The hypothesis investigated in this study is that we can reuse existing data from digitisations (10 years or older). If this hypothesis holds, it will also mean that it will be economically beneficial to preserve the old data in the sense of preserving the investment of the early digitisation. The results of exploring the hypothesis will influence preservation and dissemination strategies for the Royal Library of Denmark.

During the last decade, many mass digitisation projects have taken place all over the world. Examples of use of mass digitisation by Google can be found in [2]. Another example is Norway's National Library using Content Conversion Specialists (CCS)[1]. A decade ago, the available technology imposed limits on how automatic, fast and cheap a digitisation process could be. Today mass digitisation can be done much more cheaply and rapidly. However, there is no straight forward way to see if there is a difference in quality of the produced digital material. Quality according to requirements is important for whether digital material is worth preserving, therefore the differences will influence the strategies for preservation.

The study will explore preservation strategies mainly regarding functional (logical) preservation aspects of digitised objects, where a digital object must be preserved to be understandable and usable in the future. But functional preservation related to representation of complexities like consecutive pages is not part of this paper. The underlying bit preservation, which must ensure that the actual bits remain intact and accessible at all times, is only mentioned briefly.

Dissemination must be taken into account when evaluating preservation options in a library context. In comparison with e.g. traditional archives, libraries face additional challenges to preservation, since digital material in many cases must be disseminated to the public or researchers through fast access. Differences in purposes and goals for dissemination and preservation place different demands on the formats in which digital materials are preserved and presented, respectively. For example, many libraries have chosen TIFF or JPEG2000 as the preservation format for books and images [3], [8]. Dissemination, on the other hand, may use formats that consume less storage, e.g. JPEG or GIF, or formats with additional information for dissemination, e.g. pyramid-TIFF[2] (derived from TIFF) or JPEG2000 for images needing zoom functionality.

Besides the influence of dissemination requirements, the study will evaluate the choice of digital objects for preservation on different parameters. These are; quality of contents of the digitised object, the ability of the format to be used as a preservation format, the cost of producing objects, size of objects (related to ongoing storage costs), and the risks associated with the choices for production and storage of the digitised objects.

[1] See http://newsroom.ccs-digital.info/index.php?option=com_content&task=view&id=12&Itemid=26

[2] TIFF, Pyramid. The Library of Congress (National Digital Information Infrastructure & Preservation Program)

2 CASE STUDY: THE ADL SYSTEM

In order to explore the hypothesis, we will use the Danish ADL System as a case study. This system was developed by the Royal Library of Denmark together with "Det Danske Sprog- og Litteraturselskab" (DSL) which publishes and documents Danish language and literature. The Royal Library developed the framework, while DSL selected literary works to be included. The ADL system is a web based dissemination framework for digitised material from the archive for Danish literature. Today it contains literature from 78 authors represented by over 10,000 works of literature (e.g. novels, poems, plays). ADL additionally contains author portraits as well as 33 pieces of music (sheet music) and 118 manuscripts. The publication framework is still available on http://www.adl.dk/.

The case study is interesting because it reflects a system built on the basis of technologies from the start of this century. Today, new demands have arisen for dissemination and preservation, and new technologies exist to produce the digital material.

2.1 Present Architecture and Contents

The ADL system does presently offer display of book pages based on the framework designed a decade ago. Each page can be viewed in three different ways derived from original scanned TIFF files with page images; as a 4-bit GIF image, as a pure text representation or as a download of a PDF containing the page image for print. This is a typical application from 2000 where 4-bit GIF was chosen to allow quick dissemination of ADL web-pages to users using relatively slow connections.

Digitised manuscripts and sheet music were added at a later stage. These are represented via JPEG images, because JPEG is a better dissemination format for e.g. handwriting on yellowed/coloured, deteriorated paper.

The structure of the web framework is based on authors, their literary works and the period when the authors were active. The website is based on dynamic HTML pages generated from information in a database.

2.2 Present versus Desired Preservation

The focus is on the preservation of the digital objects, thus we will not go into details of the functionality of the system. It should however be mentioned that scalability for fast response time and ease of maintenance of dissemination applications must be taken into account in the final decision on the preservation strategies.

At present, the ADL is only preserved as a part of the Danish web archive. This means that the only data preserved is the data visible on the internet, which does not include e.g. TIFF files and special encodings. Further actions for preservation await the results of this research project.

The preservation strategy considered for ADL data is a migration strategy. The focus here is to preserve and possibly reuse earlier digitisation as a basis for a migration into emerging dissemination and preservation formats. Emulation[1] does not support changes in presentation form and is therefore not considered.

The ADL information which is the target for preservation is: the digitised representation of the book *items* (as defined in [10]) including page images and encoded texts, manuscripts and sheet music as well as the related information such as period descriptions and author descriptions. Since the author and period descriptions were written especially for the ADL system, these are born digital.

ADL is presently disseminated from its own platform which is not aimed at preservation. This will change when a preservation strategy is implemented for ADL. Dissemination in a library context is strongly related to preservation. For example, if in dissemination we use high consumption storage formats similar to the preservation format, we may want the preservation and dissemination modules to share a copy of the data, or to be able to produce dissemination copies quickly for a cached storage. Sharing a copy under bit preservation should however be done with care (see [13]). Deriving a dissemination copy requires that it will be possible to identify and retrieve preserved data on request. Furthermore, the cost of transforming data for dissemination must be minimal. In the long term, a shift from e.g. TIFF to JPEG2000 in dissemination must be coordinated with the preservation formats, and vice versa, to support scalability and efficiency.

3 EXPERIMENT SETUP

The experiments focus on the issues related to the digitisation of book items. This excludes manuscripts and sheet music, author descriptions, period descriptions, relations between information such as citations etc. The book items included have good print quality. In ADL only the text and how the text is expressed through layout and text structure are worthy of preservation. This means we do not view look & feel and illustrations as important.

The goal of the experiments is to investigate how the book information should be preserved, when we consider the costs, available technology and the higher demands for dissemination. The experiments focus on questions related to our hypothesis that the quality of the original material is such that it provides a solid basis for future development.

3.1 Preservation Scenarios – Data to be Preserved

We will investigate our hypothesis in terms of the value of the original data from ADL compared to the value we can get from a re-digitisation. On this basis we can explore different preservation scenarios that can be used in a preservation strategy for the library.

The data we will investigate is:

[1] See e.g. "Keeping Emulation Environments Portable" (KEEP). http://www.keep-project.eu/

Book item, which must be preserved, if a later rescan can be expected to add value in form of extraction of additional information or substitution.

OCR and encoded text from the original digitisation must be preserved, if it contains information that is expensive or hard to recreate.

Low resolution page images from the original digitisation must be preserved in case we conclude that we do need page images, but not necessarily in high resolution.

High resolution page images from the original digitisation must be preserved in case we expect to do future OCR adding new information, or in case we expect to do manual inspection on letters that are hard to read. Look & feel and illustrations for dissemination can also be issues, but not in the ADL case.

The preservation scenarios considered are defined in terms of combinations of data we choose for preservation. The scenarios are listed in Table 1.

	Scenario						
Data	1	2	3	4	5	6	7
Book item	*X*				*X*		
High res. page image		*X*				*X*	
Low res. page image			*X*				*X*
OCR & encoded text				*X*	*X*	*X*	*X*

Table 1. Scenarios 1-7 for preservation of data.

For the column representing scenario 1, Table 1 has an 'X' in the row with book item. This means that in scenario 1, only the book items will be preserved, thus the digital preservation is skipped. Likewise, Table 1 shows that in scenario 6 both OCR & encoded text and high resolution page images are preserved.

Choice of scenario will be evaluated against what data we need to preserve, the associated risks, the expected costs, and consequences of choice of preservation format with respect to e.g. maturity and available tools.

3.2 Digitisation Process

To better understand the set-up of the experiments, we here give a brief sketch of the digitisation process as it was performed with the original ADL data. The process, illustrated in Figure 1, was defined on the basis of the then current experiences and observations made in e.g. the DIgitised European PERiodicals (DIEPER) project [9] and best practices within digital imaging [7].

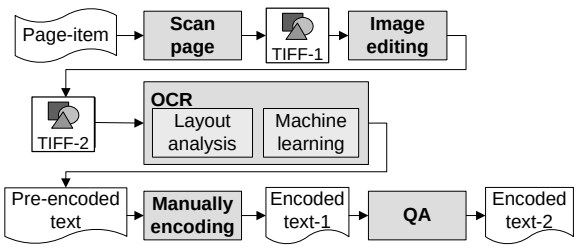

Figure 1. Simplified digitisation process.

Not all the sub-results illustrated in Figure 1 are available in the ADL case, since some data has been deleted or modified. The available information is; the page items, the image edited TIFFs (TIFF-2) and the encoded text after quality assurance (Encoded text-2).

Page item(s) from books. In ADL the books had their backs cut off, because the scanners used were document scanners with automatic page feeding.

Scan page(s) of page items. In ADL scanning was optimised and adjusted to get the best quality of the pages [9]. Grey-tone scale or adjusting the depth was used. Most pages in ADL were scanned with TIFF - 400 DPI Grey scale, 400 DPI black/white, a few with 600 DPI- Black/white and some with 200-300 DPI for pages with a high degree of background noise.

Image editing to deskew, image centering, and light & contrast adjustment. In ADL these edits were made to get a better presentation in dissemination. Furthermore, deskewing and adjustment of light and contrast enhanced the OCR results.

OCR, Optical Character Recognitions. In ADL the OCR was performed with FineReader version 5.0 or 6.0.

OCR - Layout analysis of pages. In ADL this analysis was used to identify and record objects types like text blocks, images, tables etc. and their appearance order. Common errors for the ADL scans were that the objects were not identified or they were identified with wrong type, e.g. as an image instead of a text. Manual analysis and predefined blocks for objects were used as help.

OCR - Machine learning of letter recognition. In ADL this covered assignment of Danish dictionaries that the OCR was mapped against.

Pre-encoded text results from the OCR. In ADL this included automatic formatting with page-breaks, line-breaks, new paragraph, etc.

Manually encoding as additional manual encoding result to the OCR result. In the ADL case, the pre-encoded text, represented in XML files, was sent to Russia, Sweden or India for manual encoding of various TEI-P4/TEI-lite[1] codes including e.g. speakers in plays.

QA, quality assurance of results. For ADL data this was mainly made automatically by the upload script to the ADL database which carries out a syntax check of the XML. On this point the best practices had not been followed.

3.3 Experiment Selection

Six books were chosen for the re-digitisation experiment based on wide representation of the following criteria:

- Aspects in recognition of characters. For example black-letter[2] typeface is hard to recognise for the OCR, and the results may differ for different fonts and print types
- The genres of the books. For example plays are harder to encode than novels and poems
- Illustrations included in the book. These can present challenges to the OCR block recognition

[1] TEI (Text Encoding Initiative)
[2] See http://en.wikipedia.org/wiki/Blackletter

- Notes in footer or margin in the book. These can present challenges for text encoding of the notes

The books chosen were:

{a} *Ludvig Holberg, Værker Bd. 1*. It contains essays and marginal notes. Marginal notes are not included in the current ADL encoded text.

{b} *Ludvig Holberg, Værker Bd. 3*. It contains plays and images. The book item was exchanged with another copy of the same edition. The only difference is a stamp on page 4 in the original.

{c} *Henrik Hertz, Dramatiske Værker Bd. 3*. It contains plays and is printed in a black-letter typeface.

{d} *Karl Larsen, Doctor Ix*. It contains a novel, and the print seems a bit thin.

{e} *J. P. Jacobsen, Lyrik og Prosa*. It contains poetry and diaries. Pages 96 and 97 are missing in ADL.

{f} *Anders Bording, Samlede Skrifter DDM*. It is printed in a black-letter typeface, and was included only because OCR had originally been given up.

The re-digitisation was carried out in two places and carried out with two different approaches. One represents cheap mass digitisation and the other represents re-digitisation as done for the original ADL data, and costing 8 times as much:

Subcontractor 1 (SC1) who carried out the mass digitisation. Here scanning was done in 400 DPI 24 bit colour set up in a standard configuration. OCR and text encoding was made in a semi-automated production via Germany-CCS docWORKS version 6.2-1.16 using FineReader version 7.1 and using ABBYY Morphology Engine 4.0 for an ABBYY dictionary. The results of the scanning were given in JPEG2000 and TIFF. The production of encoded text was based on TIFF and given in the ALTO[1] xml format. The automatic process did not involve image editing, manual encoding, special setup of OCR layout analysis and special setup of machine learning (see Figure 1).

Subcontractor 2 (SC2) who used an approach similar to the original ADL digitisation. OCR and text encoding based on the original ADL TIFF files, i.e. without scanning of page items and Image editing (see Figure 1). SC2 made OCR by FineReader version 9 and encoding in TEI-P4 with manual codes as specified for the original encoding. The process started on the basis of 'TIFF-2' from ADL (see Figure 1).

We decided only to do experiments on scans of the books via SC1. The interesting part is to see what the differences are between the original scans and scans made in an automatic process with standard scanning configuration for all books. We did not expect to see many differences because of the good quality of the ADL material, which is also the reason why scans were not part of the SC2 setup. The OCR set-up will in most cases still work best on scans with grey-tone 400 DPI [9], [11]. In the SC1 case the scans are in colours. Despite this, the cheap digitization price and the assumption of small differences made us settle for SC1.

Book {f} was only sent to SC2. The reason for this was that the encoding of this book had been given up earlier, thus it could not be expected to give a better result in an automated process. The most interesting investigation in this case is to see if improved technology enables OCR and text encoding of {f} today.

Encoding of marginal notes excluded in ADL from book {a} was included both in the SC1 and the SC2 results. This forms a basis for investigating if the new encoded files can replace the original encoded ADL files, because of added value of the margin encodings.

We did additional in-house experiments with JPEGs in order to investigate the question whether images can be preserved in a format requiring less storage space with less quality. For this experiment we made a sample selection of pages from the original ADL TIFFs and JPEGs derived from these TIFFs. Corresponding pages from the JPEG2000s and JPEGs received from SC1 were used. These experiments were conducted using FineReader version 10.

4 EXPERIMENT RESULTS

In this section we will start by presenting the experimental results compared to the original ADL data. Next we will relate these results to the different preservation scenarios given in Table 1.

4.1 Scanning Results

The new scans have an acceptable quality, but differ in adjustment, number of pages, colours and storage sizes. The quality is acceptable in the perspective that letters are readable from a screen presentation, i.e. it can be used for dissemination and proofreading. Furthermore, as described later, the scans can be used as basis for OCR. Illustrations are not important here, but it can be noticed that in dissemination used for ADL, they appear similarly to the old ADL scans.

The adjustment difference is due to lack of an image editing process in the automatic scans (see Figure 1). The same reason applies to the extra pages in the SC1 scans, since blank pages or pages with edition information have been removed in the original ADL image editing process.

The reason for the difference in colours is that the new scans are done in colour, while the originals were made in grey tone or black and white. This is also part of the explanation for difference in the storage sizes. However, the increased storage size is also caused by extra margins (thus larger image) which are removed in the editing process of the ADL scans, and the extra depth in some of the SC1 scans compared to ADL scans.

Table 2 gives an overview of the difference in sizes compared to the ADL scans. Note that there can be variations in these numbers depending on character density, original ADL scanning technique etc.

[1] ALTO (Analyzed Layout and Text Object). 2004. Technical Metadata for Optical Character Recognition, version 1.2.

Format	Storage factor of ADL TIFFs
SC1 TIFF	*10 times bigger*
SC1 JPEG2000	*2 times bigger*

Table 2. Storage space factor of page format.

Besides size difference for the individual pages, the SC1 will require extra storage space for the extra pages which were deleted in the ADL editing process, and the missing pages for book {e}.

4.2 OCR Results

The detailed OCR results are based on samples of pages selected on basis of variations in the page layouts.

4.2.1 Latin Typeface Pages in OCR from TIFFs

For books with Latin typefaces, the character recognition is fairly good as shown in Table 3. The numbers in Table 3 are number of differences (errors) in per mille where spaces and line breaks are excluded. For SC1 and SC2 the numbers are given for the errors in the OCR. For ADL the numbers are for the errors in the OCR with subsequent corrections.

Book \ Origin	ADL	SC1	SC2
{a}	0,1	1,4	1,2
{b}	0,0	2,5	1,3
{d}	0,0	3,3	2,0
{e}	0,3	7,5	3,7

Table 3. Number of errors (per mille).

The ADL OCR seems best, but has had subsequent corrections. Generally the difference between ADL, SC1 and SC2 OCR is small, therefore we cannot conclude whether one result is better than the other.

4.2.2 Latin Typeface Pages in OCR from JPEGs

The internal experiment with JPEG shows that the JPEG OCR was relatively good for Latin typefaces, as shown in Table 4 (calculated in the same way as for Table 3).

Origin	SC1		ADL	
Book	JPEG2000	JPEG	TIFF	JPEG
{a}	0,6	0,9	0,2	0,6
{b}	0,9	0,9	1,1	1,7
{d}	1,4	4,2	2,5	4,2
{e}	1,9	3,7	2,1	2,5

Table 4. Number of errors in the OCR (per mille).

Experiments with TIFFs from SC1 were also performed but the results were exactly the same as the results from the experiments with the JPEG2000s.

Most of the JPEGs have errors in letter recognition. Especially book {a} has many errors in the SC1 JPEG. In many cases the Danish 'ø' that is recognised as 'o'. We chose to study this problem further, an arbitrary 'ø' from the SC1 book {a} results, illustrated in Figure 2.

Figure 2. OCR of ø.

The images show that the ADL scans in TIFF are much sharper than the SC1 scans. One reason is due to optimisation in the ADL scanning and image editing (see Figure 1). In the conversion to JPEG the line in the 'ø' fades in the SC1 JPEG. This is not the same in the ADL JPEGs because they originate from TIFFs which are optimised in light and contrast.

The result gives an indication that a new encoding can be based on JPEGs for some books, although it will require extra quality assurance and manual corrections. The result also indicates that manual inspection can be based on the JPEGs for later corrections in the old OCR.

4.2.3 Black-letter Pages in OCR from TIFFs

As expected the OCR of black-letter typefaces was not without problems. Some black-letter letters can be quite hard to distinguish even for a trained human eye. Examples are "d" and "v" as well as "f" and "s".

The OCR of black-letter text requires a special additional OCR program which was not part of the SC1 package, therefore the OCR results for these books are not interesting. The SC2 faced challenges with both books which differ in black-letter fonts, and in print quality. It required addition of a special Danish dictionary and manual work to get a reasonable result.

OCR of book {f} had been given up earlier, and it did give SC2 additional challenges. Especially black-letter capitals were hard to recognise in this book. The result includes about 15-20% errors in character recognition, and many black-letter characters are interpreted as images. The earlier attempt resulted in approximately 40% errors in character recognition. Although the new OCR results are an improvement, there is still a need for a lot of manual work in order for the text to be acceptable for dissemination.

Book {c} had been through OCR and text encoding earlier with success, and SC2 has produced a much better result with less than 1% errors. Still the ADL is better with only few errors. This does not lead to conclusion that the ADL OCR was better, since it can be cause by subsequent corrective actions in the ADL XML. The comparison was made by manual inspection in order to determine whether the SC2 and/or ADL text is wrong. As described in "Improving OCR Accuracy for Classical Critical Editions" [1] there is no way to determine this automatically.

4.3 Encoding Results

The encoding results are in different formats. The SC1 result is given per page in ALTO which is an XML representation of automatic derivable typographical information about the appearance of words in the layout,

e.g. paragraphs, line breaks, word positions etc. The SC2 is given in TEI-P4 which additionally contains text structural information such as interpretations of a chapter, a paragraph, a line group, a poem, a literary work, a speaker etc. However the SC2 TEI-P4 has no position information and it follows a different kind of XML tree structure than ALTO.

The different file formats and contents also influence how much storage space the files require. Due to the very detailed positions information in the SC1 files, these files require about 20 times more storage space than the ADL files. This will, however, vary according to text density on pages and inclusion of illustrations.

4.3.1 Typographical Encodings

It was not straightforward to compare the SC1 ALTO files with the SC2 TEI encoded files, because the representations are so different. However, we found that most of the information in the ALTO files is included in the TEI files (line breaks, paragraphs etc.).

Positions cannot be compared except for accuracy, even if there had been positions in the SC2 results. The reason is that positions from SC1 are related to the scans in the SC1 result and thus are very different from the original ADL scans. This means that the ALTO results only are valuable if SC1 scans are preserved as well.

The results from book {a} with marginal notes are noteworthy. These notes were left out of the original digitisation, and can therefore only be compared between the SC1 and SC2 results. In the results from SC1 the marginal text was encoded separately from the section text, and marked as marginal text with the specification of the position of text blocks and individual words. In the results from SC2, marginal text was placed above the text-section that the note belonged to. This is not very precise since the notes have a more specific placement in the layout.

4.3.2 Text Structural Encodings

Generally, the results from SC2 do not have as good a quality as the original ADL encodings. For instance the stage directions in drama, introducing and ending a scene, are encoded in the ADL text, but only given as italic encoding of each line in the SC2 text. Hyphenation is encoded in several of the ADL texts, but none in the SC2 text, the same applies for line groups.

There are big differences in the use of TEI-codes between the TEI-files in ADL and the corresponding TEI-files from SC2. This difference will complicate merging the two results. For example, TEI div-tags were used in both files, but not in the same way and with different naming of the div tags.

5 GENERAL SCENARIO RELATED RESULTS

We will here look at the feasibility of the different scenarios. This will be done by evaluating the different materials used in the scenarios based on the results given in the previous section.

5.1 Book Items

For ADL, the only case where preservation of the book items is a necessity is when pages are missing. Here re-scans can create the missing pages. However, this also points at the importance of more thorough quality assurance of the scanned pages, which could eliminate the need to preserve the book in the ADL case. However for other kinds of material there will be cases where a re-scan is needed for other purposes. For example, if in the future higher resolution of images or look and feel of the page is needed. Thus a decision not to preserve the book items will add a risk of losing such information.

Reproduction of lost material from a digitised book will have to be based on the book item. This can be an expensive and time consuming process, especially if large chunks of data are lost, thus preserving book items alone will not meet the requirement for fast dissemination.

In the specific ADL case, the recommendation will be to ensure better quality assurance and possibly skip preservation of books. This rules out scenario 1 and 5. Note that this will need to be accompanied with a higher level of bit preservation. Furthermore, there will be many cases for digitised books in general, where it should be supplemented with preservation of the books.

5.2 OCR & Encoded Text

We can conclude that OCR & encoded text needs to be preserved. From the experiments we found that there is valuable information in the encodings which will be hard and expensive to reproduce in a re-digitisation process (especially text structural information, e.g. stage directions and poetry line groups). Furthermore, loss of OCR & encoded text information will entail a long route to re-dissemination. Thus scenario 1, 2 and 3 are not to be recommended, since they do not include preservation of OCR & encoded text.

On the other hand, OCR and encoded text should not be the only information preserved. The reason is that there is too high a risk that information is wrong (spelling errors, fonts, italics, bold etc.) and cannot be detected or corrected by inspection of the book/page images. Another risk is that valuable information is lost (e.g. marginal notes) or because of inaccurate encoding (e.g. marginal note positions). Furthermore, it eliminates the possibility of asking the public for help to identify mark-up errors, as for example done in Australia [4]. Thus, scenario 4 cannot be recommended.

One should carefully consider how to preserve the OCR and encoded text. It needs to be preserved in a way that allows enrichment of the encoding and respecting how dissemination information can be derived. The final recommendation on how to preserve OCR and encoded text will therefore be closely related to consideration of modelling the book objects for logical preservation and preparing for future enhancements with annotations from the public and researchers. This work is described separately in [12].

5.3 Page Images

As long as the page images are needed as part of the dissemination, page images should be preserved, since loss of pages will mean a lengthy re-dissemination process or complete loss of pages. Furthermore, as described under book items and OCR & encoded text, page images are the best choice as a complement to preservation of OCR & encoded text. Thus the choice is between scenarios 6 and 7. The questions that remain is what image formats we can accept as a preservation format, at what cost, at what risk, and how to retain the possibility of a short route to dissemination.

A *preservation format* will need to be a well documented format, preferably loss-less, and supported by tools. If we look at a format like JPEG, this is a lossy format which loses data when edited. However, it may still be considered as a preservation format of page images, or perhaps loss-less format as e.g. GIF or loss-less JPEG[1] might be considered. As for TIFF, this is a simple, mature, high resolution format supported by many tools, but it consumes much storage space. JPEG2000 is a high resolution format, which requires less storage space, but is more complex, and not as mature and well supported as TIFF. For supplementary considerations see [3,8].

Costs can be related to the creation of the digital objects or to ongoing storage and maintenance in connection with bit preservation (e.g. cost of hardware migration and integrity checks between data copies [13]). Looking at the ongoing costs, one way to reduce the costs is to choose a format which requires less storage space; another way is to store the format in a compressed form. Table 5 gives approximate percentages of storage reduction compared to the ADL TIFF, including numbers for LZW[2] compressed TIFFs. Note that it does not make sense to LZW compress JPEGs, and that LZW works best on black & white.

Format	% Storage of ADL TIFF
ADL TIFF LZW	*8%*
ADL JPEG	*50%*
ADL XML	*0,1%*
SC1 TIFF	*(10 times bigger) 1000%*
SC1 TIFF LZW	*(2 times bigger) 200%*
SC1 JPEG2000	*(2 times bigger) 200%*
SC1 JPEG	*15%*
SC1 XML	*2%*

Table 5. Storage space factor of page format.

The percentages are based on the experiments, but will vary for each book because of differences in letter density and inclusion of images. Table 5 includes factors for LZW compression of the different formats.

The least storage consuming format is the LZW compressed TIFFs, which were created by an optimised scanning process. Since the SC1 JPEG2000 is twice as big as the ADL TIFFs, we can conclude that have gained considerable storage reductions through optimisation of the scanning process. However, compressed JPEGs may give a better result. To draw a general conclusion of whether an optimised scanning process will be cheaper, it must be investigated further if the extra costs for manual work in the optimised scanning process can compete with the cost savings in storage.

The choice of format must be evaluated against the risk that the format may add. For instance, if a format is a lossy format, there will be a risk related to whether transformation or edits have a negative effect. If page images are saved in low resolution, the risk is that it is too low for future needs, and for later automatic extraction of additional information. Furthermore, if all books are treated identically there is a risk that e.g. books printed with black-letter may have a higher risk of losing information, even using manual inspection. On the other hand, differentiated preservation strategies for different books will influence the complexity of preservation strategies. If we use compression of the formats, this will add a risk to the bit preservation. Furthermore, compression may also add processing time in dissemination.

For ADL, we end up with a recommendation of preserving LZW compressed TIFFs, since investment, in an optimised scanning process, has already been made, it is a stable format, and book items still exist, if compression corrupts the TIFF.

6 DISCUSSION

For this study we are privileged to have unique material from an application built a decade ago. However, the state of the books and lack of information about the original digitisation process has influenced how much we can conclude. It could be argued that the condition of the material added too many uncertainties to the results. However, no matter how the ADL data has achieved its good quality, we have been able to conclude that these data is worth preserving, and we have been able to analyse which preservation strategies to choose, thus the case study has fulfilled its purpose in the investigation of our hypothesis.

In this study we have only investigated digitised books with high print quality and only focused on the text, layout, and structure of the text. Other books with other characteristics may need other digitisation and preservation strategies. For example the page images of sheet music and manuscripts in colours may need higher resolution in preservation, and might give better results in mass digitisation. Another example is, if text structural information is unimportant then all information can be extracted automatically. In any case compromise and choices must be made at the start of scanning.

In the ADL case we saw that it seems beneficial, regarding ongoing storage costs, to digitise books using a more manual approach than mass digitisation traditionally takes. The actual difference in cost can be

[1] See http://en.wikipedia.org/wiki/Lossless_JPEG
[2] See http://en.wikipedia.org/wiki/Lempel-Ziv-Welch

hard to evaluate. A model for calculation of migration costs is given in [5], and results will differ according to the period it covers and how it is used.

The JPEG case study only represents a special setup with specific parameters for OCR and conversion. The result may have differed with different tools and setups. However, similar results may be found in experiments with decrease of image quality.

Mass digitisation does have different advantages. For example, there may also be a time factor for how long a digitisation process must take, e.g. because of a political deadline or deteriorating material (an example can be found in [6]).

The preservation strategy must take into account how to model the complex structures so that the necessity of the short route to dissemination is respected. The reason is that it can influence how the metadata and the OCR and encodings are represented in the preserved data. In other words, the research presented here only provides input on the reuse of earlier digitisation, and which factors should be considered in new digitisation.

7 CONCLUSION

We can conclude that digital material from the ADL case study is worth preserving. In general it points at possible reuse of digitised books where look & feel as well as images are unimportant. The digital objects, that are the target for digital preservations are the page images and the OCR & encoded text. Whether the pages are preserved in high or low resolution, with or without compression, must depend on a risk analysis, and analysis of relation to dissemination.

An additional result of the study was that the chosen digitisation process can influence the ongoing storage costs in the future. This leads to a conclusion that the digitisation process must be chosen with care both regarding the immediate requirements, but also regarding the long term consequences.

Digitisation has evolved in the last decade, enabling cheaper and faster mass digitisation, as illustrated in the different digitisation approaches. The most evident evolution in the ADL case was enabling OCR of books printed in black-letter which were given up a decade ago. However, the manual work in scanning and in OCR & encoding corrections have added value to the material. For book material of as good quality as in ADL, the technological enhancements cannot compete with these manually added values.

A final preservation strategy in a library context can now be made on basis of this study as well as the modelling aspects related to functional preservation.

8 REFERENCES

[1] Boschetti, F., Romanello, M., Babeu, A., Bamman, D., Crane, G. "Improving OCR Accuracy for Classical Critical Editions" *Proceedings of the European Conference on Digital Libraries*, Corfu, Greece, 2009.

[2] Coyle, K. "Mass Digitization of Books". *The Journal of Academic Librarianship, Vol. 32, Issue 6*, 2006.

[3] Gillesse, R., Rog, J., Verheusen, A. "Life Beyond Uncompressed TIFF: Alternative File Formats for Storage of Master Image Files" *Proceedings of the IS&T Archiving Conference*, Bern, Switzerland, 2008.

[4] Holley, R. "Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers". *Technical Report from National Library of Australia*, Australia, 2009.

[5] Kejser, U.B., Nielsen, A.B., Thirifays, A. "Cost Model for Digital Curation: Cost of Digital Migration" *Proceedings of the International Conference on Preservation of Digital Objects*, San Francisco, USA, 2009.

[6] Kejser, U.B. "Preservation copying of endangered historic negative collections" *Proceedings of the IS&T Archiving Conference*, Bern, Switzerland, 2008.

[7] Kenney, A. R., Rieger, O. Y. *Moving Theory into Practice: Digital Imaging for Libraries and Archives*. Research Libraries Group, California, USA, 2000.

[8] Kulovits H., Rauber A., Kugler A., Brantl M., Beinert T., Schoger A. "From TIFF to JPEG 2000? Preservation Planning at the Bavarian State Library Using a Collection of Digitized 16th Century Printings", *D-Lib Magazine Vol. 15 No. 11/12*, 2009.

[9] Mehrabi, H., Laursen H. "Standards for images and full text" *Proceedings of Conference on future strategies for European libraries*, Copenhagen, Denmark, 2000.

[10] Riva, P. "Functional requirements for bibliographic records: Introducing the Functional Requirements for Bibliographic Records and related IFLA developments" *Bulletin of the American Society for Information Science and Technology vol. 33 issue 6*, 2008.

[11] Holley, R. "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs", *D-Lib Magazine Vol. 15 No. 3/4*, 2009.

[12] Zierau, E., "Representation of Digital Material Preserved in a Library Context". *Proceedings of the International Conference on Preservation of Digital Objects*, Vienna, Austria, 2010.

[13] Zierau, E., Kejser, U.B. "Cross Institutional Cooperation on a Shared Bit Repository". *Proceedings of the International Conference on Digital Libraries*, New Delhi, India, 2010.