Long-term archiving of digital data on microfilm

Steffen W. Schilke*

Gemeinsame IT-Stelle der hessischen Justiz

Friedrich-Ebert-Straße 28
D-61118 Bad Vilbel, Germany
Postal Address:
Oberlandesgericht Frankfurt am Main,
D-60256 Frankfurt am Main, Germany
E-mail: steffen.schilke@gmail.com
*Corresponding author

Andreas Rauber

Department of Software Technology and Interactive Systems,
Vienna University of Technology Favoritenstr.,
9-11/188, A-1040 Vienna, Austria
E-mail: rauber@ifs.tuwien.ac.at

Abstract: E-government applications have to archive data or documents for long retention periods of 100 years or more. This requires to store digital data on stable media, and to ensure that the file formats can be read by available software. Both applications as well as media technology have only short life spans. Thus, data has to be migrated at frequent intervals onto new data carriers and to new file formats. However, original file versions usually need to be retained permanently. In terms of cost, stability and technology independence, microfilm storage offers a promising solution for off-line storage. This paper reports on a feasibility study analysing encoding techniques that allow digital data to be saved onto microfilm, testing data recovery as well as cost issues.

Keywords: digital preservation; bit-stream preservation; long-term storage; microfilming.

Reference to this paper should be made as follows: Schilke, S.W. and Rauber, A. (2010) 'Long-term archiving of digital data on microfilm', *Int. J. Electronic Governance*, Vol. 3, No. 3, pp.237–253.

Biographical notes: Steffen Walter Schilke is working in the field of archiving and document management for more than a decade. Currently shaping the archiving architecture in the state of Hessen, Germany. He holds a Computer Science Degree and a MBA Degree. In addition he is Certified Document Imaging Architec (CompTIA). He is a member of the British Computer Society (MBCS), Bund deutscher Volks- und Betriebswirte (BDVB) and the German Gesellschaft für Informatik (GI e.V.).

Andreas Rauber is an Associate Professor at the Department of Software Technology and Interactive Systems (ifs) at the Vienna University of Technology (TU-Wien). He furthermore is President of AARIT, the Austrian Association for Research in IT and a Honorary Research Fellow in the Department of Humanities Advanced Technology and Information Institute (HATII), University of Glasgow. His research interests cover the broad scope of digital libraries and information spaces, including specifically text and music information retrieval and organisation, information visualisation, as well as data analysis, neural computation and digital preservation.

1 Introduction

The introduction of e-government services has radically transformed workflows and services being offered. With a delay of several years, a range of new challenges have been appearing stemming from this transformation. Among them is the need to provide suitable long-term preservation of digital data in archival institutions. This, in turn, has led to the implementation of specific digital preservation solutions as part of e-government initiatives in most national or state archives. Digital data is prone to decay on several levels, one being the storage of data (bit-level preservation), ensuring that the data is securely stored on data carriers that can be read with current technology.

A second layer is the logical preservation: digital objects require specific software to be opened and read, which in turn require specific operating systems, device drivers, and, ultimately, hardware to run. Finally, semantic preservation is essential to facilitate correct interpretation of objects, similar to conventional analogue pieces of information. These challenges are now being addressed in numerous research and development projects around the globe to ensure that the electronic objects created during e-government processes can be authentically preserved.

Several solution are being implemented, usually relying on regularly migrating data both from old storage technology to current one, as well as format migration to current versions of file formats. However, specifically the latter usually incurs changes to the objects, some of which may seem undesirable with respect to their future usage, especially since these changes accumulate over a series of migration steps. While careful planning procedures try to limit this effect (Strodl et al., 2007), it cannot be avoided completely. Thus, most initiatives recommend to always maintain the original format version of an object to allow reconstruction and access e.g., via emulation if required. This, in turn, calls for a cost-effective strategy for bit-level preservation of digital data on a durable storage technology not requiring regular maintenance.

Additionally, most systems require a stable back-up copy to be maintained for all data (including migrated versions) in addition to online versions for continuous use. Unfortunately, most digital storage techniques do not offer themselves for these purposes: hard disc RAID arrays require regular operation and need to be replaced every few years. Tape drives as long-term storage of massive amounts of data require regular re-winding of tapes to maintain them readable. Further, with current

development cycles the durability of the tapes has surpassed the support life-time for tape readers, rendering the respective tapes unreadable unless migrated to new types of tapes.

This has led to the revival of a rather unexpected storage technique for digital data, namely microfilming. Microfilm, especially black/white film, has proven a very durable medium, requiring no maintenance apart from appropriate storage conditions. Microfilm as storage medium has a life span of more then 100+ years (e.g., Voges et al., 2008) and has the advantage that a media migration has to be done less frequently. It is already used for long-term storage of scanned images of paper documents. It is also in widespread use in archival institution, ensuring that expertise in its handling as well as appropriate technology is in place. A recent AIIM Market Intelligence Quarterly mentioned that 43% of the organisations that participated in the poll still use microfilm/fiche (Frappaolo, 2008).

Provided correct encoding schemas are used, microfilm can store both analogue representations (i.e., images) of objects as well as the digital data stream, offering the additional benefit of easy inspection and redundancy of representation forms, as it basically already includes a kind of 'migrated' analogue representation in addition to the digital object. Furthermore, the technology required to read data from a microfilm is rather simple and stable, promising readability and decodability for very long-time periods. Given the correct encoding scheme, even manual reconstruction of digital data is in principle possible.

In a nutshell, binary data is encoded in a printable form, either using a textual encoding such as UUencode, or a barcode image representation. This data is then transformed into an image and stored (exposed) onto microfilm. To recover the binary data, the images one the microfilm are scanned and passed through a de-coding process, e.g., applying Optical Character Recognition (OCR) software and subsequent UUdecode, or by decoding the scanned barcode image. This results in the original binary data stream that can be used for further processing.

This paper reviews issues involved in utilising microfilm as storage medium. It focusses specifically on different encoding schemas that allow the storage of digital data on an analogue carrier. Due to space limitation we exclude discussions of migration and emulation strategies, which are addressed in specific preservation planning reports (Strodl et al., 2007). It also provides estimates for storage capacity, and subsequently costs, of storing digital data on microfilm.

We particularly emphasise and analyse the use of standard technologies, i.e., conventional microfilm read/write equipment rather than purpose-built research prototypes. This trades higher performance potentials for easier deployability now. Results show the potential of this approach, specifically as off-line back-up storage of digital data to complement the working copy, which obviously needs to be stored in easily accessible and continuously maintained online storage media. While microfilm does not offer itself in any way as a copy for continuous use, it offers a very stable medium not requiring any continuous maintenance of the storage system. However, several problems surface when it comes to the evaluation of the encoding schemes available. While text-based encodings such as UUencode offer the possibility for visual error correction based on human inspection, the data recovery based on OCR does not provide sufficient quality requiring tremendous manual intervention to obtain the correct digital encoding. Specifically, the UUencoded data has to be written in rather large font sizes, leading to a very low storage density.

The bar code based encoding offers significantly higher storage density combined with efficient error correction codes, allowing perfect data recovery. Yet, as the encoding is not human-readable, any errors on the microfilm that might occur cannot be corrected by human intervention. Still, this seems to offer the best option for off-line maintenance-free back-up data storage of digital objects in their original form as accepted into a long-term digital archive for potential disaster recovery after a failure of the online system with storage costs around 0.22EUR per MB.

The remainder of this paper is structured as follows: Section 2 provides an overview of current research initiatives on utilising microfilm and paper for storage of digital information. Section 3 describes the set-up of the study presented in this paper, with results being summarised in Section 4. Finally, Section 5 summarises the findings, providing recommendations on how to implement microfilm back-up storage for the archiving of e-government data, and lists open issues in improving efficiency of this storage technique.

2 Related work

Different approaches can be taken to store digital data on microfilm as analysed e.g., by the Arche project (Wendel and Schwitin, 2007). The basic principle is to encode binary data in the form of either a sequence of textual characters or barcode like images. These are 'printed' onto microfilm. By scanning and de-coding the microfilm images the original binary data can be recovered.

The proposed methods are studied in detail in the Millennium project 'Bits on Film' VDI VDE IT – InnoNet (2006) and shall be implemented by using specialised hardware. The plan is to use a colour microfilm laser writer. This requires a special microfilm scanner in order to re-digitise the data. A similar project is Persistent Visual Archive (PEVIAR) (Amir et al., 2008; Müller et al., 2008). PEVIAR uses hybrid storage, i.e., simultaneously storing an analogue image and the digital representation as a colour 2D-barcode.

The study presented in this paper does not include hybrid storage, as the focus is entirely on the encoding and subsequent de-coding of the digital information. Obviously, hybrid storage having an analogue representation of data next to the encoded digital form, is possible as well. Only analogue storage as performed in many archival institutions is not sufficient even for static documents such as paper surrogates in the form of PDF documents. In many cases the digital content of such a document cannot be entirely recovered from the image representation by OCR scanning. The AIIM poll (Frappaolo, 2008) showed, that on less than 10% of all scanned image content OCR can be used by 67% of all participants. However, if the image representation is archived alongside with a digital representation of the document or data on microfilm, it is still possible to recover the actual digital file while having an analogue images as stable preview. The technology used in the experiments reported in this paper is similar to the Computer Output on Microfilm/Microfiche (COM) technique (Gavitt, 2002). Most projects, however, so far used this technique solely to create visual representations of digital information, i.e., a print-out to microfilm, such as e.g., in the Digital to Microfilm Project, where scanned images of brittle books were transferred to 35 mm microfilm using COM technology (Kenney, 2002).

The need to use error correction on barcode based data storage is know since many years and was used as early as 1966–1968 in the IBM 1360 Photo-Digital Storage System (Griffith, 1969; Oldham et al., 1968). In order to achieve a 'fail safe' implementation a new method has to be designed in order to have multiple levels of error correction, redundancy and to increase the systems tolerance of disturbances. Nowadays approaches which work with every micro metre on the microfilm can lead to problems recovering the data due to imperfect scanning from the microfilm (Amir et al., 2008; Hofmann, 2008; Voges et al., 2008). Additionally, the use of colour in such a system can add complexity as film material tends to change the colour when stored over a long time.

3 Film-based archiving set-up

This study uses standard microfilm technologies for long-term archiving of digital data on microfilm. The devices used are a Kodak Digital Archive Writer (DAW) i9610 and a microfilm scanner Kodak 3000 DSV (a.k.a. Minolta ES-7000). The systems use standard software (PowerFilm and AWIS/KIWI). The microfilm used is a 16 mm black and white microfilm, which is 30.5 m long. It is developed by the Kodak Prostar development unit. This technology is successfully deployed in the long-term archiving market since many years.

The standard 16 mm, high-quality microfilm used is the Kodak Reference Archive Media which is based on polyester (PET) and meets ISO requirements for safety film (ISO 18906:2000, 2000). The material has a life expectancy rating of LE-500 and is certified to remain readable for 500 years when properly processed and stored under controlled conditions (ISO 18911:2000, 2000). In addition the microfilm meets (ISO 18901:2002, 2002) specifications for stability. COM equipment has not been available for the experiment and to ease the scanning of the microfilm an automated microfilm scanner has been favoured. This is widely available equipment which allows an automated scanning of a microfilm, delivering the input for the system which decodes the images back to the preserved format.

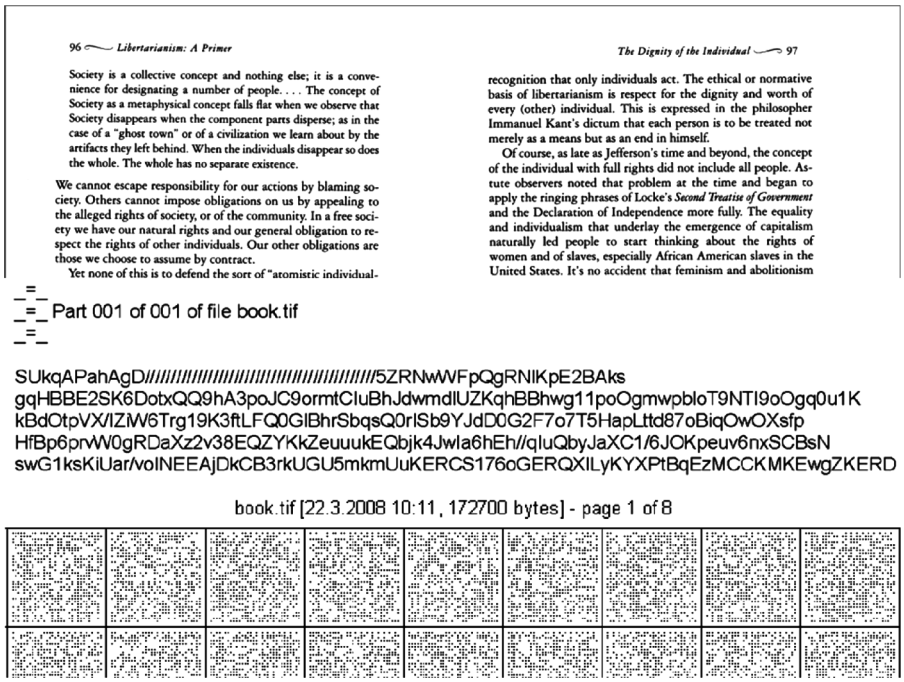
The content of the digital file is not relevant as the entire file (e.g., audio, video, images, text, raw data, source code, byte code/programs, colour, black and white) will be encoded and recorded to microfilm. When the data has to be read, i.e., converted back to a digital file, the images which contain the data in encoded format will be scanned from the microfilm and decoded in order to create the original digital file.

The method can be used as one long-term archiving component for the Archival Storage part of the ISO 14721:2003 (2003) reference model. The Manage Storage Hierarchy and the Disaster Recovery parts of this section can use this type of archiving of digital data (as backup media or as part of the storage hierarchy). Using such an approach, the need to migrate data from one media to another is reduced or not necessary at all, especially when the media is stored properly. In such cases bit rot of the media is not an issue.

In the study we use files in a range of different formats, specifically Text (TXT or CSV), TIFF, PDF and PDF/A (ISO 19005-1:2005, 2005). Two running examples will be used throughout this paper: The *book.tif* sample file is a DIN A4 scan (ISO 216:2007, 2007, 28.0 cm × 21.6 cm, landscape) of two pages of a book. The size is 168 KB. The image has 2 bits per pixel (two colours) and a resolution of 6.600 × 5.100 pixel. Another sample file is a PDF file called *TIFFPM6.pdf* with 17 pages (1995, TIFF Technical Notes) in DIN A4 (ISO 216:2007, 2007,

21.6 cm × 28.0 cm) and a size of 47 KB. Figure 1 (top) shows an excerpt of the TIFF image. Note, that the actual type of file does not have any impact on the study. They are merely chosen as a representative sample of the type of content present specifically in e-Government settings with its strong requirements on the long-term retention of the original, unmodified objects. Any type of digital data, be it documents, measurement data, or any kind of software application can be archived using the methods analysed below. The main principle is to take binary data and to transform it into a representation that can be stored on an analogue carrier such as microfilm. This could be simply the sequence of 0's and 1's underlying all kind of traditional data storage. However, the encodings detailed below have shown to be more efficient for storing data on carriers such as microfilm or even paper. For each of the source sample files HASH values (SHA-256) are calculated to check if the de-coded data is identical to the original files. The files are then converted into text-based formats and barcode formats (2D barcode Images) as described below. To write these files on microfilm the encoded representations are converted to TIFF images. The resulting images are written on the microfilm which is then developed. For the tests about 2.700 images in DIN/ISO A3 (ISO 216:2007, 2007) landscape or Tabloid landscape (ANSI/ASME Y14.1-2005, 2006) were written, resulting in approximately 1.5 microfilms.

Figure 1 Encodings of a TIFF image (excerpts): (top) analogue representation, (middle) Base64 representation and (bottom) barcode representation



The next step is to use the microfilm scanner to digitise the images from the microfilm into a digital image. These images where transformed back into a digital form by de-coding the textual or barcode representations using Open Source software. The resulting files were checked for identity with the original objects using

the Hash code. It is important to keep in mind that writing the images on microfilm always involves a scaling factor in order to project, i.e., write, the larger images on the microfilm. By scanning the images from the film, a magnification takes place by the optical system. This has negative effects on the quality of the resulting images, which is the main reason for the loss in data quality, leading to errors in the de-coded files.

3.1 Text-based encodings

Since the dawn of the internet, text-based formats have been used to transport digital data, e.g., in e-mails or web services. Using these encodings, binary data is transformed into the subset of printable ASCII characters. In this study, the formats Base64 (Josefsson, 2006), BinHex and UUencode are used. The sample files can be converted using any of the many available open source converters. An excerpt of the UUencoded representation of the TIFF image is given in Figure 1 (middle).

The advantage of text-based encodings lies with the fact that these can in principle be read without additional software, allowing for manual error correction in the scanning/OCR process. While in principle binary numbers (e.g., octets of 0/1) could be written onto the film, these encodings offer a higher compression by using e.g., 4 characters to encode 3 bytes of information, rather than 24 characters. However, not all character encodings are suitable for the purpose of reading and de-coding via OCR. Specifically, BinHex uses a large alphabet of characters to represent the binary data, including all kind of special characters, apostrophes, quotes, commas etc., which cannot reliably be read using standard OCR software. We thus focus for the subsequent experiments on UUencoded representation of binary data. These files are then printed using a TIFF printer driver (Microsoft Office Document Imaging Writer) as TIFF images. Different fonts (Arial, OCR-A, Courier) and font sizes (22pt, 26pt) and different resolutions (100, 200 and 300 DPI) were used for the print out to evaluate the impact of these parameters. The text-based formats where formatted with different fonts and font sizes in order to best utilise the available space on the page, e.g., DIN A4 (ISO 216:2007, 2007) landscape. The page numbers required for encoding the two sample files in different forms are given in Table 1.

Table 1 Encoding sizes for two sample objects using Base64 and UUencode

book.tif (168 KB)	*TIFFPM6.pdf* (47 KB)
Base64	*Base64*
77 characters/line	77 characters/line
3035 lines in total	844 lines in total
102 pages, Arial 20pt	52 pages, Arial 20pt
92 pages, OCR A, 20pt	26 pages, OCR A, 20pt
UUencode	*UUencode*
61 characters/line	61 characters/line
3846 lines in total	1071 lines in total
153 pages, Arial, 24pt	45 pages, Arial 24pt
117 pages, OCR A, 28pt	45 pages, OCR A, 28pt

3.2 Barcode encodings

Barcode representations transform the digital data into large 2D barcodes. We evaluate the open source tool *Paperback* (Yuschuk, 2007) that was initially created to store digital data on paper. In the context of this feasibility study such a format is evaluated for the first time for its usability for microfilm-based storage of digital data.

The Paperback software (Yuschuk, 2007) is designed to print the 2D barcode via a printer on paper. In addition the output can be saved as BMP image file. This BMP image is converted into a TIFF image using a printer driver (Microsoft Office Document Imaging Writer, at 100, 200 and 300 DPI), which can then be written by the DAW on microfilm. Paperback is using a Reed-Solomon Error Correction, c.f. Reed and Solomon (1960). In addition, a factor for a redundancy can be chosen. During the evaluation, the redundancy factors of 1:5 and 1:7 have been used. A redundancy of 1:5 represents a redundancy for five blocks of data. This allows to recover the lost data (e.g., because it is unreadable) if one out of five blocks cannot be decoded. Other parameters of Paperback, which can be adjusted, are the dot size, white space and compression.

The *book.tif* file is converted into 13 pages of barcode images when using a 1:5 redundancy (with 200 DPI, 70% Dot Size) and into 12 pages when using a 1:7 redundancy (with 200 DPI, 70% Dot Size). The first page of the barcode representation of the TIFF image is shown in Figure 1 (bottom).

3.3 Writing on microfilm

The images created from Paperback or the TIFF print output of the text-based formats are DIN/ISO A3 landscape (ISO 216:2007, 2007, 297 mm × 420 mm) and Tabloid landscape (ANSI/ASME Y14.1-2005, 2006, 11″ × 17″) 'prints'. These formats have been chosen to use the most available space on the 16 mm microfilm, taking the scanning capabilities of the microfilm scanner into account. The maximum page size the microfilm scanner can read is a little bit larger than tabloid landscape. As the DAW can write a maximum of 3888 pixel horizontally (for the 17 inch of the tabloid landscape page) an image can be vertically about 2700 pixels (the 11 inch dimension of the paper).

The images are written using the standard DAW software, which automatically chooses the best reduction to utilise most of the available space on the 16 mm microfilm resulting in a scaling factor of around 40. The resulting images were written in sequentially order. The DAW uses a LED line that writes 3888 points per line on the 16 mm microfilm. It writes black and white pictures on the microfilm which are inverted before writing.

3.4 Reading from microfilm

The digitisation of the images from microfilm is done using the Kodak microfilm scanner 3000 DSV. The magnification of the images and the focus is adjusted before scanning. As the magnification is adjusted mechanically, it is not possible to know the exact magnification factor of the optical system. The Powerfilm software controls the microfilm scanner and scans the images from the optical system and stores the images as TIFF. The images are inverted during the scan to get the

original image. The microfilm scanner was set to 600 DPI and grey scale. This reflects the recommendation of the authors of the Paperback software (Yuschuk, 2007). Subsequently, the images are converted into bitmap (BMP) format for subsequent de-coding by the Paperback software.

For the text-based encodings, the images were converted using OCR software bundled with Microsoft Office which is based on the OmniPage engine, and subsequently de-coded using UUdecode, BinHex and Base64 decoders.

4 Evaluation of storage/recovery quality

4.1 Recovering text-encoded data

The different DPI resolutions and font sizes (e.g., Arial 22pt or 26pt) used show, that the results improve when using larger fonts so that the characters are easier to differentiate for the OCR engine. For paper based scans a test showed that the OCR accuracy does not improve beyond a scanning resolution of 300–400 DPI, cf. Rice et al. (1996). Studies show that a higher resolution is necessary for scanning images from microfilm because of the downscaling/magnification process, cf. Hofmann (2008).

If the images have a sharper or 'larger' look, the OCR results are better. With small fonts and low resolution more than 10 OCR errors per line of text have been experienced. Images with high resolution and larger fonts resulted in only 1–2 errors per line of text. Nagy et al. (1999) has mentioned similar results on OCR of text on paper. Analysing the type of errors encountered may help with devising suitable error correction techniques for each type of encoding. We thus provide a detailed review of errors found within the OCR result of a sample UUencoded page with 21 lines with 61 characters each in Table 2. Each line has one character for the length and 60 characters for the 45 byte data. In the setting used for this experiment one page can store 945 bytes of data.

The scan from the microfilm shows some grey-ish colour and scratch lines which harmed the recognition of characters by the OCR software. On a few lines the OCR engine recognised some characters which have nor been on the image. The error for line 2 which accounts for 56 missing or not recognised characters can be avoided by using a better OCR engine and special OCR fonts. The main OCR errors are a confusion of the capital character *O* or lower-case character *o* instead of the number 0 or vice versa (35 times) and the number 1 instead of lower case character *l* or vice versa (8 times). Other common recognition errors for UUencode have been similar looking characters like: I, i, 1,], j, J, etc. Xxencode for example does not use such special characters so that these OCR recognition errors could be avoided.

Extra spaces are recognised but can be easily filtered because they are not used by UUencode. A mis-recognised or missing linefeed can be repaired by inserting it if there are more then 61 characters in a line. The number of wrong characters account for approximately 6.25% of all characters. These results were further validated by OCR-ing directly from the created TIFF images, i.e., without writing to and scanning from the microfilm, to evaluate the impact of the optical distortion (see Table 3).

Table 2 Analysis of OCR errors encountered decoding 1 sample page of UUencoded data, printed in Arial font size 24pt, scanned from microfilm

Line	*Extra space*	*Wrong character*	*Missing character*	*Additional character*
1	5	8		
2	1		56	
3	4	3	2	1
4	2	2	3	
5	10	7	1	
6	5	2	1	1
7	2	4	3	1
8	5	4		1
9	5	3	3	
10	1	4	1	
11	2	5	1	
12	6	6	1	1
13	3	1	5	
14	1	3	3	1
15	2	7		
16	3	2	1	
17	3	2	4	
18	2	3		
19	5	4	3	
20	3	5		
21	2	4		
Totals	72	80	86	4

None of the files could be correctly recovered from the OCRed encoding. Also here the main OCR errors are a confusion of the capital character *O* or lower-case character *o* instead of the number 0 or vice versa (26 times) and the number 1 instead of lower-case character *l* or vice versa (3 times). Other common recognition errors for UUencode have been similar looking characters like: I, i, 1, j, J, etc. The number of wrong characters account in this example for approximately 4.84% of all characters.

As these text-based formats do not include error correction and redundancy, they do not deliver a result that can be used for long-term archiving of digital data. Different errors of the OCR engine like character confusion, substitution, split or misinterpretation are the main reasons for such problems, cf. Alkula and Pieskä (1994), Nagy et al. (1999) and Rice (1996) (Note, that the commonly reported high quality OCR results of 99.9% correctness are achieved for standard texts using dictionaries to correct for the most likely mis-recognitions. Such a correction approach is not possible for ASCII-based encodings of binary data with virtually random/unlimited character sequences.).

These problems could be reduced by using a limited character set and designing a special text-based format, which includes mechanisms for error correction and provides redundancy on the line and page level. The character set selected should, for example, avoid characters that are easily confused, substituted or misinterpreted. Still, a clear advantage of text-based formats is the possibility to let humans read and type the characters. Thus, further investigations are required to develop encoding

Table 3 Analysis of OCR errors encountered decoding 1 sample page/image of UUencoded data, printed in Arial font size 24pt

Line	*Extra space*	*Wrong character*	*Missing character*	*Additional character*
1	2	6		1
2	1	6	3	
3	6	3	2	
4	1	1	4	
5	4	5	2	
6	5	1	3	
7	2	3		
8	5	2		
9	5	2	2	
10	3	4	1	
11	2	3	2	
12	2	5	2	
13	5	2	3	1
14	1	3		1
15	4	4	3	
16		1	2	3
17	3		3	1
18	3		3	1
19	5	5		1
20	2	4		
21	2	2		
Totals	63	62	35	9

schemas that are suitable for high-quality OCR recovery in order for this approach to be feasible for long-term data storage.

4.2 Recovering barcode encoded data

Depending on the parameters used for Paperback (redundancy and DPI resolution) for the barcode encoding, the data could be perfectly decoded without any remaining errors after the error correction and redundancy information had been used to cope with errors. This dramatically changed once the small range for redundancy and DPI resolution for the feasible parameters was left, leading to complete loss of decodability of data.

The decoding of the sample file *book.tif* results in a 100% decoded and error free file after applying the built-in error correction. By using the error correction the Paperback software could correct 1.100 bytes which showed an error in the 2.293 good blocks of the 2D barcode (1 : 5 redundancy). When using the 1 : 7 redundancy the number of bytes which had to be corrected was 720 out of 2.209 blocks of the 2D barcode (see Table 4). Paperback 2D barcodes created with higher resolutions resulted in higher error rates so that the software could not recover the file by using the error correction and the redundancy. Reasons are the loss of quality and resolution due to the re-scaling of the image when writing it on the microfilm. This loss of data did not leave enough data for a recognition of the barcode once magnified during scanning.

Table 4 Results of de-coding a Paperback 2D Barcode from a microfilm scan

File	*Redundancy*	*No. of good blocks*	*No. of bad blocks*
book.tif (40 DPI)	1:5	2.293	1
book.tif (40 DPI)	1:7	2.209	0
book.tif (100 DPI)	1:5	1.433	766
File	*File size*	*Bytes corrected*	*% Error*
book.tif (40 DPI)	172.700	1.100	0.64
book.tif (40 DPI)	172.700	720	0.42
book.tif (100 DPI)	172.700	15847	>9.17

The column % Error show the percentage of bytes from the whole file which have been corrupted and corrected by theredundancy and the error correction functionality of Paperback. The scanning of the Paperback barcode pages works only properly when a low DPI resolution (e.g., 40/80 DPI) was chosen for the creation for the output of Paperback. Using higher DPI setting (e.g., >100 DPI) the 2D barcode images suffered a big loss in quality. The system could either not identify the frames or recognise the barcode data in a recognised frame properly which led to an abortion of the operation. This is because more information is being compressed into a more fine-granular image, which is distorted by the reduction and magnification when writing and scanning the microfilm. Even when using higher resolutions of the microfilm scanner the results could not be decoded. Even the redundancy and the error correction could not help to improve the decoding of the barcode into the digital form. In addition the gray-ish look and the scratch marks or lint on the microfilm or lens resulted in poor quality of the scan.

4.3 Cost issues

With the DAW usually two DIN A4 pages are written side by side on a microfilm (LE 500 – life expectation of 500 years, produced according to ISO 18901:2002, 2002). This results in a capacity of a 30.5 m microfilm (deducting the necessary lead in and lead out of the microfilm) of approximately 7.200 DIN A4 images. Given the cost of microfilm of approximately 10 Euro the price per image stored is 0.14 cent. For the storage of text-encoded or barcode encoded data in DIN A3/Tabloid landscape, 3.600 images fit on a 30.5 m microfilm. For the 168 KB file *book.tif* 92 pages with OCR A/Base64, 117 pages with OCR A/UUencode and 13 pages of Paperback 2D barcode were generated. The Base64 representation thus can hold approximately 1.8 KB per page, the UUencode representation can hold 1.4 KB per page and the Paperback 2D barcode can hold approximately 12.9 KB. By using these numbers a single microfilm can store 6.32 MB coded in Base64, 4.92 MB coded in UUencode, and 45.32 MB coded as Paperback 2D barcode. This would result in a cost per MB for Base64 of 1.58 Euro, for UUencode of 2.03 Euro and for the Paperback 2D barcode of 0.22 Euro (all above see Table 5). Considering the long-term perspective of the medium this cost seems to be more efficient than other storage systems, especially when considering the cost for regular media migration every 5–7 years required with other storage forms, as well as higher maintenance cost for these.

Table 5 Storage capacity on microfilm

Encoding	*KB/Page*	*MB/Microfilm*	*Cost Euro/MB*
Base64	1.8	6.32	1.58
UUencode	1.4	4.92	2.03
Paperback 2D Barcode	12.9	45.32	0.22

The numbers above represent the costs for the plain storage media. These do not seem to compare favourably with the costs per MB for other storage media. Current hard discs cost around 100–300 EUR per TB, or 10–30 cents per GB. However, these technologies usually carry a huge burden in terms of operational costs, as massive amounts of discs need to be kept operational (i.e., actively spinning) in order to ensure that they remain operational. This adds a significant amount of cost in terms of power supply, infrastructure and maintenance over the relatively short lifetime of such a storage system, cf. Linden et al. (2005).

Solid State Discs (SSDs), though significantly more expensive, are already available at costs of 2–3 EUR per GB. Similar to microfilm based storage, these do not require constant power supply and operation in order to maintain their data and are thus seen as a suitable alternative long-term storage medium. Thus, in terms of access speed and convenience SSDs outperform storage on microfilm technology. One advantage of the latter, though, is the independence from complex technology and control logic, as briefly reviewed in the recovery scenario below.

4.4 Recovery scenario

Contrary to other storage media such as hard discs, flash memory or SSDs, tapes, CDs or DVDs, the technology to read microfilm is first of all rather simple, and secondly very generic. While specific interfaces and rather sophisticated technology is required to recover data from any of the complex storage technologies in use today (optical drives with laser technology, high-precision placement of reading devices on magnetic carriers, controller software and hardware to select the appropriate locations, etc.) only simple optical devices are needed to actually read data from a microfilm. This effectively de-couples the storage media from the IT systems in use to read them. While it may be rather complex to get both a by then outdated tape drive, DVD drive or USB port connected to a future computer system, for microfilm it will be sufficient to connect any then contemporary optical imaging device, be it a scanner, a camera or other device. In the case of textual encoding, even mere human reading and re-entering of the encoded data stream into a contemporary IT system would be possible (although only from a theoretical point of view, as it obviously does not scale to allow large-scale recovery in a purely manual way) as long as our current character system is still in use.

Once the encoded data has been read into a contemporary IT system, it needs to be decoded to recover the original binary data stream. In the case of textual encodings, which have a long tradition and are widely documented (Base64 is for example documented as RFC 4648, cf. Josefsson (2006), this can be achieved via a simple translation table that these encodings are made up of. Programs to perform these are rather simple to write, as the only input (that needs to be archived along with the encoded microfilms) is the visual representation of the translation tables. Again, even

this step could – theoretically – be handled entirely manually. For barcode data, the decoding routine is somewhat more complex, but again rather simple to document and re-implement on any future computing system. This applies for barcodes which are defined as a standard (e.g., PDF 417, ISO 15438:2006, 2006).

It should be noted, that the resulting data formats may not be readable by that time as the software to read the specific file formats may not be available, or – in case the software has been archived along with the data as well – the hardware to run it may not be available. This aspect of digital preservation has to be and is being dealt with in another stream of DP research focusing on logical preservation of data, which builds on top o the physical preservation. This is commonly achieved by performing either regular migration of data to recent file formats or by offering emulation services, to name the two most common strategies currently deployed. The long-term storage of files as described above is motivated primarily by the desire of many institutions to always archive a copy of the very original digital object received in addition to the archival copy kept for continuous use. A complete recovery would, at least with current technology, not be a viable options in terms of scale. If using a microfilm based digital archive as backup media it has to be clear that this archive is not a fast random access storage but it is rather maintenance free for a long time so saving media migration tasks and costs. In case of a disaster recovery case for a digital archive a scan on demand approach should be chosen to recover only the necessary and needed data from the microfilms.

5 Conclusions and future work

This study shows that it is possible to store digital data (i.e., files or documents) on microfilm and to convert them back to a digital format. This offers a viable solution for the permanent storage of the original data streams as defined in most archival regimes in e-Government settings and long-term archives. Advantages of long-term archiving on microfilm include specifically the lack of maintenance required for maintaining the data, as opposed to most other digital storage alternatives which require regular media replacement or handling. A further advantage is the simplicity in terms of technological requirements for reading and decoding the data, as standard scanning/imaging devices coupled with OCR software are sufficient to decode the data, as opposed to more complex systems required for spinning disc type storage media or specific tape drives. Last, but not least, due to the analogue storage medium hybrid storage of dual analogue and digital views of the data side-by-side are possible, allowing easy inspection of data prior to reading and decoding into its original form.

Two encodings were evaluated for the analogue storage of digital data. The experiment using text-based formats like Base64, BinHex, and UUencode fails because of the poor quality of the OCR results of the digitised images from the microfilm. The lower quality of microfilm scans compared to a scan from paper is a result of the downscaling and magnification process. This was evaluated to be about a 10% difference comparing the source materials paper and microfilm, cf. Arlitsch and Herbert (2004), stemming primarily from the difference in the scanning quality. For the storage of digital data on microfilm in order to compensate this recognition difference precautionary measures have to be taken. For textual encodings to

become a viable alternative, several improvements such as embedded error detection and correction codes are required. By adding checksums to the grid of text-encoded data error correction could sufficiently improve the results. Furthermore, specifically designed encoding schemas avoiding likely OCR mismatches need to be devised. Additionally the use of special fonts like OCR fonts could increase the recognition quality of the scanned microfilm images.

Still, textual encodings offer an interesting alternative due to the fact that the encoded data is human readable, holding potential for better error recovery. With error-correcting codes added, correct de-coding seems feasible, which in turn may also enable higher compression rates by using smaller font sizes to increase the storage capacity of a single microfilm. Thus, one of the major downsides of textual encodings, namely the relatively low storage density, may be resolved (see Table 6).

Table 6 Comparism of encodings

Encoding	*Advantages*	*Disadvantages*
Textual Encoding	Human-readable	Low storage density
	Visually correctable	Low error-resilience
	Simple technology	
	Simple de-coding Algorithm	
2D Barcode Encoding	Higher storage density	More complex En/decoding
	Error-correction	Non-human
	Embedded efficiently	Readable/correctable

With the tested text-based formats, a 170 KB file was transformed to 191 pages, whereas the 2D-barcode needed only 12/13 pages depending on the chosen redundancy. Barcode based representations allow almost perfect recovery of digital data from the scanned microfilm images even with current standard state-of-the-art techniques. Thus, this representation offers itself for a stable storage representation of digital data on microfilm. Yet, the downside of the barcode representation is the need to rely on more complex decoding algorithms to re-create the original file, while for text-based encodings, manual correction and inspection may be applied.

A specialised 2D-barcode with an optimised usage of redundancy and error correction can further improve its viability for a long-term archiving system based on microfilm. The clear advantage is that a microfilm based solution in combination with long-term archiving data formats reduces the need for media migration and is thus very cost efficient also in terms of maintenance and media handling. No spinning discs are needed and the only precautions necessary for the storage of the microfilm is a dry, temperature controlled and locked closet. In addition, the standard DAW can write two microfilms at once so that one microfilm can be stored off-site. The advantage of such a method is clearly the 'don't care' factor for the media migration. With a life time of more than 100 years a lot of media migration projects (usually necessary every 3–7 years) can be saved.

Thus, microfilm-based storage using state-of-the-art technology offers itself as a feasible alternative for permanent storage of the original data objects, as well as

for permanent back-up storage of format migrated versions of digital objects when desired.

References

Alkula, R. and Pieskä, K. (1994) *Optical Character Recognition in Microfilmed Newspaper Library Collections: A Feasibility Study*, Tech. Report Espoo Research Notes 1592, Technical Research Centre of Finland, VTT Tiedotteita, Meddelanden.

Amir, A., Müller, F., Fornaro, P., Gschwind, R., Rosenthal, J. and Rosenthaler, J. (2008) 'Towards a channel model for microfilm', *Archiving 2008*, Vol. 5, pp.207–211.

ANSI/ASME Y14.1-2005 (2006) *Decimal Inch Drawing Sheet Size and Format American Society of Mechanical Engineers*, USA, ISBN 0791829901.

Arlitsch, K. and Herbert, J. (2004) 'Microfilm, paper, and OCR: issues in newspaper digitization at the Utah digital newspapers program', *Microform & Imaging Review*, Spring, Vol. 33, pp.59–67.

Frappaolo, C. (2008) 'Content creation and delivery – the on-ramps and off-ramps of ECM', *Market IQ Intelligence Quarterly Q4 2008 – Market Insights based on the Opinions and Experiences of 198 AIIM Members and Industry Associates AIIM – The ECM Association*, Silver Spring, MD, USA.

Gavitt, S. (2002) *Computer Output Microfilm (COM)*, New York State Archives, Report 52, 14 March.

Griffith, R.L. (1969) 'Data recovery in a photo-digital storage system', *IBM Journal of Research and Development*, Vol. 13, p.456.

Hofmann, A. (2008) 'Der ARCHE Farbmikrofilmbelichter – Digital speichern – filmbasiert archivieren?', *Presentation at the Workshop of the Deutschen Gesellschaft für Photographie (DGPh) at Photokina*, Slides are available at http://www.dgph.de/node/82, http://www.dgph.de/content/sektionen/wissenschaft_technik/symposium08/Hofmann_%28FhG%29-Vortrag.pdf

ISO 14721:2003 (2003) *Space Data and Information Transfer Systems – Open Archival Information System – Reference Model ISO*, Geneva, Switzerland.

ISO 15438:2006 (2006) *Information Technology – Automatic Identification and Data Capture Techniques – PDF417 Bar Code Symbology Specification ISO*, Geneva, Switzerland.

ISO 18901:2002 (2002) *Imaging Materials – Processed Silver-gelatin Type Black-and-White Films – Specifications for Stability ISO*, Geneva, Switzerland.

ISO 18906:2000 (2000) *Imaging Materials – Photographic Films – Specifications for Safety Film ISO*, Geneva, Switzerland.

ISO 18911:2000 (2000) *Imaging Materials – Processed Safety Photographic Films – Storage Practices ISO*, Geneva, Switzerland.

ISO 19005-1:2005 (2005) *Document Management – Electronic Document File Format for Long-term Preservation – Part 1: Use of PDF 1.4 (PDF/A-1) ISO*, Geneva, Switzerland.

ISO 216:2007 (2007) *Writing Paper and Certain Classes of Printed Matter – Trimmed Sizes – A and B Series, and Indication of Machine Direction ISO*, Geneva, Switzerland.

Josefsson, S. (2006) *The Base16, Base32, and Base64 Data Encodings Request for Comments: 4648*, Network Working Group, The Internet Society, http://tools.ietf.org/rfc/rfc4648.txt

Kenney, A.R. (2002) *National Endowment for the Humanities PS-20781-94. Digital to Microfilm Conversion: A Demonstration Project 1994–1996*, Cornell University Library, Department of Preservation and Conservation Final Report, http://www.library.cornell.edu/preservation/com/com_n.html

Linden, J., Martin, S., Masters, R. and Parker, R. (2005) *The Large-Scale Archival Storage of Digital Objects*, DPC Technology Watch Series Report 04-03, February.

Müller, F., Fornaro, P. and Gschwind, R. (2008) 'Paper and digital encoding: toward self-explaining codes', *Proc. of EVA 2008*, 12–14 November, Berlin, pp.31–36.

Nagy, G., Nartker, T.A. and Rice, S.V. (1999) 'Optical character recognition: an illustrated guide to the frontier', *Proc. of Document Recognition and Retrieval VII*, SPIE, Kluwer, Norwell, MA, USA, Vol. 3967, pp.58–69.

Oldham, I.B., Chien, R.T. and Tang, D.T. (1968) 'Error detection and correction in a photo-digital storage system', *IBM Journal of Research and Development*, Danvers, MA, USA, Vol. 12, No. 6, pp.422–430.

Reed, I.S. and Solomon, G. (1960) 'Polynomial codes over certain finite field', *Journal of the Society for Industrial and Applied Mathematics*, Vol. 8, pp.300–304.

Rice, S.V. (1996) *Measuring the Accuracy of Page-Reading Systems*, Doctoral Dissertation, Dept. of Computer Science, Univ. of Nevada, December, Las Vegas, The dissertation is online available at http://www.isri.unlv.edu/publications/isripub/rice-dissertation.pdf.

Rice, S.V., Jenkins, F.R. and Nartker, T.A. (1996) *The Fifth Annual Test of OCR Accuracy*, Technical Report, Information Science Research Institute, TR-96-01, April, University of Nevada, Las Vegas, NV, USA.

Strodl, S., Becker, C., Neumayer, R. and Rauber, A. (2007) 'How to choose a digital preservation strategy: evaluating a preservation planning procedure', *Proceedings of the 7th ACM IEEE Joint Conference on Digital Libraries (JCDL'07)*, 18–23 June, ACM Press, New York, NY, USA, pp.29–38.

VDI VDE IT – InnoNet (2006) *Millenium: Bits on Film*, Project Description, http://www.vdivde-it.de/innonet/projekte/in_pp146_millenium.pdf

Voges, C., Märgner, V. and Fingscheidt, T. (2008) 'Digital data storage on microfilm-error correction and storage capacity issues', *Proceedings of IS&T Archiving Conference*, June, Bern, Switzerland, pp.212–215.

Wendel, K. and Schwitin, W. (2007) *SchluBbericht zum Verbundprojekt ARCHE: Entwicklung eines Farbmikrofilm-Laserbelichters zur Langzeitarchivierung digitaler bzw. digitalisierter Dokumente*, March, Universitätsbibliothek Stuttgart, Stuttgart, Germany, Online available at URN: urn:nbn:de:bsz:93-opus-30113, URL: http://elib.uni-stuttgart.de/opus/volltexte/2007/3011/

Yuschuk, O. (2008) *PAPERBACK v1.00*, http://www.ollydbg.de/Paperbak/index.html last accessed 19.3.2008.