The Obsolescence of Migration: Long-Term-Storage of Digital Code on Stable Optical Media

David Gubler, Lukas Rosenthaler and Peter Fornaro; Imaging and Media Lab of the University of Basel, Basel, Switzerland

Abstract

Future accessibility to information requires on one hand a stable media that reliably stores the information content and that, one the other hand offers accessibility with the tools available in the future. Today, institutions have to rely on proprietary technology. Most of them, hard-drives, magnetic tapes, or DVDs, do not fulfill the requirements of long-term stability and accessibility. Thus, an ideal storage solution for archival purposes should meet the following requirements: it should offer high stability, high data density, low costs per GByte, should be easy to handle, undemanding regarding to the infrastructure, and technology independent..

An example of a technology independent and reliable method to store data is human readable symbolic code, such as written text. In this case, the interface to access the data is reduced to the eye of the observer. Considering this fact, we propose to investigate a data carrier based on photographic material of high stability, on which digital information of any kind can be stored as visible digital barcode. Thanks to the visibility of the data, it is possible to recover the information with any digital scanning or camera device of appropriate quality. As photographic material is usually based on more than one layer of dyes, each layer can be used as a separate data channel. The quantization depth of each channel leads to a digital code over a higher than binary alphabet. Fragmenting and distributing the data on the film can increase resistance against local deterioration.

The available technology at the time of the future digitization and interpretation of the archived information is unknown. Thus, the complexity of information retrieval will be simplified by a sophisticated encoding procedure, an asymmetric approach common in data compression. Such a sophisticated asymmetric codec requires an optimized interaction of information carrier and digital code. We will analyze state of the art photographic material to evaluate its properties. Taking into account the characteristic features of the data carrier, the digital code will be optimized regarding data density, robustness and simplicity. Furthermore, highly sophisticated error correction will be applied to the data to enhance security. The presented approach combines the positive aspects of digital data and an ultra stable medium like microfilm (estimated lifetime up to 500 years at room temperature) and leads to an archival media with an expected storage capacity of up to 700 MByte per sheet (104x148mm2 color microfiche).

Philosophical Principles of anArchivist

Digital data of all kinds will be a fixed constituent of most archives. This new digital cultural heritage comprises no longer only documents on paper or film materials but will contain digital media in a wide variety of hardware- and software-formats of various brands etc. However, the long-term archival of digital data is still an unresolved issue.

Other than the preservation of paper or microfilm, today the preservation of digital information demands constant attention. This constant input of effort, time, and money to handle rapid technological and organizational advance is considered the main stumbling block for preserving digital information beyond a couple of years. Indeed, while we are still able to read our written heritage from several thousand years ago, the digital information created merely a decade ago is in serious danger of being lost.

Today digital information is stored on different media using various techniques. Typical systems are the very popular CD-R, different derivatives of the recordable DVD, magnetically working media like hard disks or magnetic tapes. Without going into the details of these technologies, it is important to focus on some features which make the named solutions unacceptable for archival use: Each of the named technologies is depending on the specific hard- and software to access the data. The compatibility of a certain technology like the CD-R is supported by second maybe third generation technologies of equipment but no longer. For example it is possible to read a CD-R with a modern DVD drive, but is debatable if such a disc can be read with a next generation optical drive. Taking into account the limited lifetime of a specific technology, the complete loss of data can occur after a quite short time. Magneto optical drives for example disappeared from the market within 10 years.

- Many data carriers like the CD-R or the DVD-R or any magnetic tape are not stable. The used dyes respectively the magnetization or even the carrier itself decays with time. If no hardware copy process is performed, the data is most likely lost after a time of 5 to 20 years. The decay can not be easily predicted [1, 2].
- The technological details of the hard- and software are not open to the public. It would be expensive to build an appropriate e.g. CD drive after the technology has disappeared from the market. With most technologies it is not possible to access the data without the proprietary hardware.
- There is little or no cross compatibility between different storage systems concerning hard- as well as software formats.

However, if managed correctly digital data can be kept forever without any loss. The following features of digital data explain why this provocative statement is true:

- Digital information can be replicated, i.e. copied, without information loss if the replication procedure is correctly applied.

- It is possible to introduce local redundancy. Local redundancy reduces the risk of loosing data if small parts of the media the data is written on are damaged.
- Non-local redundancy can be achieved by storing several identical digital versions of the same analog original. If one of these digital copies were destroyed, the information would still be available. In order to minimize the risk, the different copies should be stored in various geographical locations.

Taking this the three statements into account it is straight forward that:

Digital data must be migrated to survive technological changes. The best case would be a system that migrates data by itself. The University of Basel is working on this topic in the project Distarnet [3].

The data carrier must be as technology independent as possible and should have as little decay as possible. In other words: From the point of view of a perfect digital medium for archival use, the solution would be to store digital information with intrinsic infinite lifetime on a medium with as little decay as possible.

The second approach is interesting, because it is similar to the classic challenge of archiving material with analog content, eg photographs or paintings.

Visible Data on Photographic Material

Microfilm is a well-known stable medium with very slow decay and it is widely used and accepted in archives.

Today for archival purposes, documents are captured on this high-resolution film and stored under optimal conditions. This technique is standard in the community of archives and museums for photographic or art collections. Beside its stability, microfilm has other positive features for archival use. Stored information can be read (looked at) without a high-tech machinery, it is visible. From a psychological point of view, this characteristic attribute of visibility is very important for the use in an archive. One of the major difficulties of nearly every device that can store digital data is the high degree of technological dependence. As already mentioned, the technological change has as much an influence to the lifetime of any digital storage system as the decay of the actual data carrier. Photographic material combines some important advantages, which make it unique for the use as a digital storage medium:

Stability: Photographic material is highly stable even under suboptimal conditions. A well stored Ilfochrome microfilm for example can last as long as 500 years.

Visibility: Photographic material has either one or three layers of dyes with absorption spectra located in the visible range. Such material can be read without proprietary hardware. In fact any digital scanning device or camera can be used to read the film.

Data density: Photographic material can have a very high data density at a reasonable low price. A state-of-the- art microfiche (104x148mm2) can have a resolution of up to 300 pairs of lines per millimeter leading to a theoretical pixel count of 62’400 x 88’800 pixels.

Costs: Photographic material is reasonably inexpensive. Even though the initial costs are relatively high (€uro 10.- for one microfiche), the accumulated costs over time are low because of the lack of revolving costssuch as unnecessary migrations.

Another important fact is the technological advance of the imaging industry. Digital cameras as well as scanners have improved in quality and became much cheaper within the last few years. Considering a normal further development of imaging technologies one can suggest that if it is possible to read data from film today, it is very likely that it will be even simpler to do it in the future [Fig. 1].

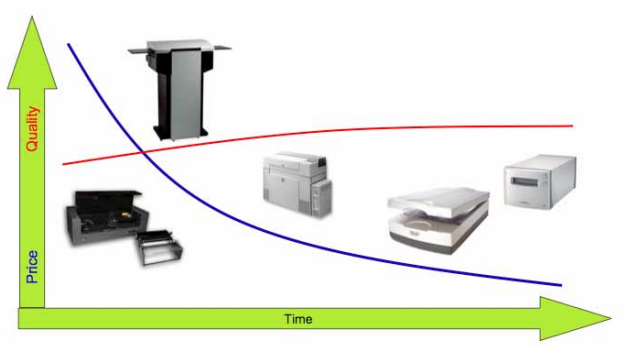

Figure 1: *Scanners are part of the IT base technology. Within the last years scanners became much cheaper without any loss of quality.*

Technological Challenges

The main goal of the project is the development of a digital data storage solution for archival purposes with extended lifetime (>100y) based on standard photographic material as well as high-resolution microfilm. The digital information (original domain) is to be saved as 2D barcode on the media (media domain) with a high data density. Depending on the film used, we are dealing with pixels of one or three layers of dye (grey-scale or color). Each layer can have different density levels, allowing a code over a higher alphabet than binary. The number of bits per pixel will be maximized regarding signal to noise ratio. In addition, state of the art error correction coding will be applied to the original domain values to increase storage redundancy. Furthermore neighbor original domain values will be fragmented and distributed in the media domain, leading to high reliability. Beside the digital barcode, the film can also hold a preview image and text (meta-data) describing the content [Fig. 2]. Part of the project is the evaluation of different media to define their qualities as data carrier. To prove our concepts the read-back process for the digital data will be implemented as well.

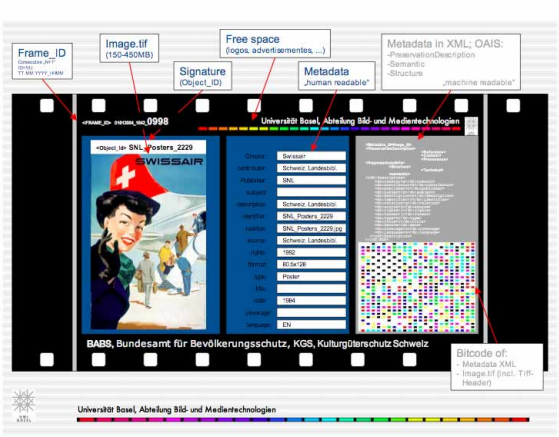

Figure 2: *A possible dataset exposed to film. The frame is divided into parts with different content. Besides the digital bar code,meta information is stored as well as an image of the content.*

As photographic material is highly non-linear, an important part of the research will be a detailed quantified analysis of current film material and its aging behavior. Taking into account this knowledge, an asymmetric coding process (complex writing, simple reading) will be developed [Fig. 3].

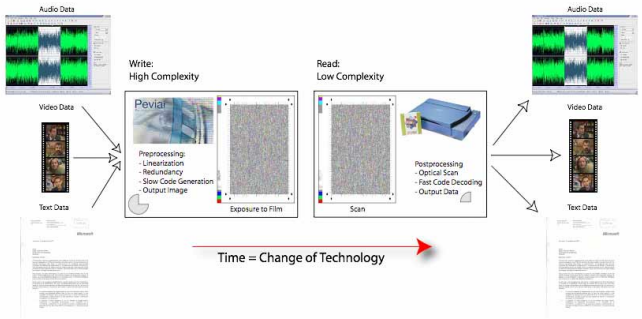

Figure 3: *Peviar uses a asymmetric processing to simplify the future read-back of data.*

For example, the encoder will compensate the non-linearity of the media itself. In addition to the digital information, visually correct images of the objects represented by the digital data will be recorded on film. Within an archival workflow, the visual impression of data can help to identify content at a glance. Further metadata containing technical, self-explaining information about the read-back procedure, descriptive information about the content and custom information will be stored on film. In case of a reliable digital data storage solution for archival purposes, we must take into account that the digitalization and interpretation of the content will happen without the knowledge at the time the data was archived. To simplify the future readability, read-back instructions will be written on the media as well as an error correction description to explain the recovery of the archived content. If necessary, it is possible to write the source code of the read-back program on film as human readable note. One aim of the project is to evaluate what kind of, and how describing information has to be stored on film to guarantee the read-back in future.

The research can be separated in two main parts: Coding theory and photographic theory. In addition further research will be done in the field of metadata.

Coding based on Channel Characteristics: On studying the characteristics of the photographic film such as how the film quality decays over time, the interactions between pixels, and other such media impairments, the coding approaches will be adapted to suit the photographic film media on which information is written.

A common impairment observed in magnetic recording channels is inter-symbol interference (ISI), i.e., during read-back of data, the value of the symbol read-back is affected by the values of the symbols written adjacent to it. In photographic films if the nature of pixel to pixel interaction is studied, then this information can be incorporated either during encoding or during decoding of data. The ISI can be handled at the time of writing data by using a recorder, that essentially does an inverse of what an ISI channel is expected to do, and thereby nulls out the ISI. Alternatively, the ISI can be handled at the decoder wherein the ISI can be modeled as a separate channel and the turbo principle can be used to exchange information between the ISI decoder and the channel decoder (LDPC/product-code decoder), as implemented in [4]. Furthermore, to protect some portion of the data more than the rest, the portion of data that is vital will be first encoded using simple linear codes and then re-encoded with the remaining information using the above methods. Alternatively, unequal error protection codes may be considered to prioritize the protection according to the level of importance.

In general, photographic film material can be regarded as carrier of information. The amount of information, which can be stored, is determined by:

Characteristic curve (sensitometry): The characteristic curve [5] defines the dynamic range for each colorant. Taking into account the intrinsic noise of film, the possible bit-depth or photometric resolution will be experimentally researched. Various photographic materials will be analyzed micro densitometrically [6] to define the maximum recordable amount of data per pixel. To optimize the re-degitalization process, the distance between two consecutive optical density levels must be maximized (highest differentiation). For a given number of densities, this is achieved when the levels are equidistant, i.e. linearly distributed in the density domain. The necessary exposure levels to generate this ideal distribution can only be computed if the characteristic curve is well-known. In an experiment a grey scale will be exposed to film and sensitometrically measured.

Grain (SNR)/Resolution: The spatial resolution [7] as well as the maximum dynamic range is limited by the silver halide and the coupled dyes of a film. Both properties are of high relevance to achieve high data density. Modern microfilm can reproduce up to 300 line-pairs per mm (dotsiz e = 1.66µm) leading to a theoretical data density of 135M Byte/cm2 at a quantization depth of 10 density levels per layer. The physical limitations of the laser film recorder used to write the data decreases the maximum obtainable data density. To optimize it, the interaction of film and film recorder will be evaluated. To get the RMS granularity of the material, homogenous fields of a defined optical density will be

exposed and analyzed with a microdensitometer. The resolution (modulation transfer function) can be defined by exposing a point to film and measuring the sharpness of its edge. Alternatively a sinus function is exposed and the cut-off frequency of the material is measured with a micro-densitometer.

Beside the sensitometric and noise aspects, photographic color film material is a highly non-linear material. Each layer of dye, causes various distorting effects, namely:

Side-absorption of the dyes: The characteristic spectral absorptions of the dyes overlap. If light of a particular wavelength is exposed to film, supposed to address a single dye, the side absorption causes the neighboring layers to be activated as well. The optical density measured at a certain wavelength, e.g. at about 450 nm for blue, results not only from the yellow dye but as well from the side absorption of the cyan and magenta dyes [5]. This effect is well known in photography and has a strong influence on color reproduction. If color film is used to write data, each layer can be used as a separate data channel. Do the spectral sensitivities overlap, the channels can not be addressed autonomously, leading to a loss of data capacity. The effect will be researched to optimize the data density. The side-absorbtion must quantitated by measuring the color-dyes spectrophotometrically.

Aggregation of dyes: Certain colorants in photographic material (especially the azo-dyes in silver dye bleach material, which are the most promising for archiving purposes) aggregate in high concentration. This means that the spectral absorption maximum is shifting at different density levels. Quantitative knowledge of the effect allows to linearize the density of the film, leading to a simplified read-back process [8].

Inter-image effects: Inter-Image-Effects [9] occur in all colour materials. In the widest sense the word Inter-Image-Effect comprises all the effects within a certain layer of the film that are produced by the exposure of the other two layers. The reasons for Inter-Image-Effects are mostly chemical in nature (vertical part of the diffusion of chemical reactants between layers). Inter-Image-Effects are distort the data reproduction as well and have to be measured and compensated.

Neighboring effects (chemical diffusion and optical scattering): There are two neighboring effects in photographic material:

A) The photographic film is composed of different layers with a certain thickness (some µm). Light penetrating in the film is scattered by the silver halide, resulting in an unsharp image (point spread function),where the effect depends on the thickness of the layers. Pixels next to each other will be distorted by scattered light.

B) Neighboring effects can also be caused by diffusion of chemicals [10] that are used widely in color film material as they enhance the sharpness.

Another aspect to take into account for compensation is the aging of the used photographic material that causes a change in density in each dye. Aging of film material at room temperature can be simulated. Film ages faster In an environment of high temperature and constant humidity. Arrhenius describes the relation between aging and temperature. To quantify aging, a gray scale is exposed to film and measured periodically while the film is kept on a defined temperature. The knowledge of linearization and aging will be used to optimize the data density.

A Foresight

We finally will give a brief outlook on the research and the industrial activities in this field, namely the PEVIAR Project of University of Basel and NOAH, a cooperation project of Fraunhofer IPM and MicroArchive Systems GmbH (MAS). The UniBas PEVIAR is a research project, dedicated to analyzing and quantifying film material under the aspects of archival purposes. This projects is endorsed by various institutions like the Swiss National Library, the Association for the Preservation of the Audiovisual Heritage of Switzerland (Memoriav), the Swiss Federal Department for the Protection of Cultural Property and the Swiss Federal Institute of Technology Zuerich. The NOAH project of Fraunhofer and MicroArchive Systems on the other hand is an industrial development, having presented its first prototype of a closed storage workflow at the 2006 CeBIT'06 fair in Hannover, Germany.

References

[1] OliverSlattery, Richang Lu, Jian Zheng, Fred Byers, and Xiao Tang
Stability Comparison of Recordable
Optical DiscsAStudy of Error Rates in Harsh Conditions
J. Res. Natl. Inst. Stand. Technol. 109, 517-524 (2004)

[2] AES Standard for audio preservation and restorationMethod for
estimating life expectancy of compact discs
(CD-ROM), based on effects of temperature and relative humidity
Reference number: AES28-1997 (1997).

[3] Lukas Rosenthaler, Rudolf Gschwind
DISTARNET A Distributed Archival network
IS&Ts 2004 Archiving Conference, San Antonio, April 2004. IS&T: The Society for Imaging Science and Technology, 7003 Kilworth Lane, Springfield, Virginia 22151 USA, p. 242-248, ISBN: 0-89208-251-8

[4] A. Dholakia, E. Eleftheriou, M. Fossorier, and T. Mittelholzer,
"Capacity-Approaching Codes for the Magnetic Recoding Channel",
IBM Research Report, available at http://whitepapers.zdnet.co.uk/ .

[5] W.T.Hanson, C.A.Horton,
Subtractive Color Reproduction Interimage Effects; Journal of the Optical Society of America, 42, p.663, 1952

[6] P.Kowaliski
Theorie photographique applique Masson et CieParis, 1972

[7] T.H.James
The Theory of the Photographic Process
4th Ed.; Macmillan Publishing Co., Inc., New York, 1977

[8] R. Gschwind, A. Rosselet, H.J. Buser and E. Baumann
Investigation and quantification of inter-image effects
The Journal of Photographic Science, 41, 86, 1993

[9] G.G.Attridge, R.E.Jacobson et al.;
Observations Concerning Inter-Image Effects in a Colour Printing Paper
The Journal of Photographic Science, 31, p.197, 1983

[10] J.M. Sturge, V. Walworth and A. Shepp (Editors)
Imaging Processes and Materials
Neblette's 8th edition, Van Nostrand Reinhold, New York, 1989

Author Biography

Dr. Lukas Rosenthaler is a full time staff member of the Imaging & Media Lab of the University of Basel, Switzerland. The main research topics are the long-term preservation of digital data and the restoration of movie.

He also leads the project team of Distarnet, a P2P based archival system for long-term archival.

Dr. Peter Fornaro is staff member at the Imaging & Media Lab of the University of Basel, Switzerland. Fornaro is working in the field of color management and the long-term preservation of digital data.

David Gubler is a research assistant at the Imaging & Media Lab of the University of Basel, Switzerland. He is member of the advisory board for the MicroArchive Systems and the Fraunhofer IPM Joint Project "NOAH". He is a founding member of the Swiss Mikrosave® Fachlabor Gubler AG, a service company for Digital Imaging and Archiving.

.