Digitization and Long Term Archival of Photographic Collections: Recommendations of the Swiss Federal Office for Civil Protection, Section Protection of Cultural Property

Rudolf Gschwind, Lukas Rosenthaler* and Rino Büchel***
** Imaging and Media Lab, University of Basel, Switzerland*
***Civil Protection, Section Protection of Cultural Property, Bern, Switzerland*

Abstract

The long term archival of analogue data, especially photographic images, has to minimize at least two risks: 1) the total destruction of the image e.g. through fire or water, and 2) the intrinsic decay of the images through the natural aging of the medium has to be slowed down as much as possible. Both risks can be minimized by appropriate storage (location of storage, fire protection, climate control etc.). As a consequence of optimal storage, access to the archived images will be very restricted and difficult. Further, the intrinsic decay of the images due to aging cannot be stopped totally but only be slowed down. Since all copying of analogue media such as photographs, film, video etc. always and inevitably imposes a decrease in quality, the original has to be preserved. It is usually possible to interpret analogue media such as photographs without the use of technical means. Metadata such as image description, photographer, timestamp etc. are often recorded together with the image on the same media (e.g. notes on the back of a photograph) and which therefore can not be separated from the image and the risk of loosing the connection of the metadata with the image is usually very low.

- With respect to long term archival and compared to analogue data, digital data has completely different properties:
- Given the proper procedures, digital data can be copied infinitely oft (infinite number of generations possible). The "original" and the "copy" are identical and cannot be distinguished. Therefore the notion of an "original" looses its sense in the digital domain.

There is no slow decay in the digital domain. Digital data either can be read completely, or there are errors reading the data that invalidates the whole data set. In order to recognize the occurrence of read errors, special algorithms ("checksums") have to be used.

What reasons may cause the loss of data within the digital domain? The following short overview is especially targeted to the problems of long term archival where not only the image object alone but also the information about the object, the metadata, has to be stored. For digital archives, another difficulty arises: digital media can only be read and interpreted by using technical devices. A computer tape looks the same to the human observer if it's empty or if it contains images, texts or other data. A vault with 10'000 CD-ROMs that are not labeled is almost without value if there is no other information available. The consequence is that improper archival strategies can lead to a fatal loss of information at many different levels. If only at one of these levels information is lost, the whole archive will be virtually worthless or at least the value will be diminished severely.

Introduction

In 1962 Switzerland joined the "Hague Convention for the Protection of Cultural Property in the Event of Armed Conflict" of 1954 and the Swiss federal government committed itself to implement all necessary measures for protecting and respecting the cultural properties. It also provides support for the cantons, helping them to coordinate and implement the measures that have been planned. In particular the Confederation issues directives and guidelines for specialized training, and itself sees to the training of a cultural property protection management team. The federal government also provides subsidies for precautionary measures, to ensure the protection of cultural property including buildings and other types of property of national or regional importance. The competent authority in this case is the section in charge of the protection of cultural property within the Federal Office for Civil Protection.

As one of the first steps, an inventory of the cultural properties of national or regional importance has been compiled. When inventory or documentary work is done for objects belonging to this inventory, the Swiss federal government usually pays a financial contribution between 20

and 30 percent. Museums, archives libraries etc. can apply for such subsidies at the Section of Protection of the Cultural Property (SPCP) of the Federal Office for Civil Protection.

Usually such inventories and documentations are being microfilmed with the goal of long term archival. Up to now, about 49'000 microfilms are stored in protected shelters.

In the past years however an increasing number of institutions applied for support and subsidies for the digitization of image data. These demands raised the question of whether the goal of preserving the cultural heritage could also be achieved using digital technology. The SPCP of the Federal Office for Civil Protection therefore launched the study presented here establishing recommendation for the digitization and the long term archival of photographic collections as digital data.

It is important to note that often not safeguarding but improving access to the collections is the primary goal that involve the digitization of complete photographic collections. But in this case the quality requirements for the digitized images are relatively low. For safeguarding, however, the image quality must be much higher. Two questions arise in this context: a) which quality for the digitization is needed and b) having the whole collection of photographs digitized: is it possible to use the digital image data for safeguarding purposes?

Therefore the SPCP intends to financially support digitization projects in the future. However, financial support will only given to projects where strict quality standards are guaranteed. This paper will present the foundation onto which these recommendations for quality assessment are based.

Part I: Digitization

With the goal of preservation of cultural heritage, the digital image has to be a complete substitute for the original image relating to the visual content:

The "digital facsimile" must contain at least the same amount of visual information that could be extracted by conventional photographic reproduction techniques.

However, a proper digitization alone is not enough to comply with the conditions of long term archival. There are three areas where special quality measures have to be taken:

Digitization

The digitization process bridges the analogous world of classical photography with the digital domain and it determines the upper limit of the quality of the digital images. If this first process does not truthfully reproduce the visual content of the original image, there is no way to improve the image quality later on (except by a new digitization).

Meta Data

Available information about the image – the meta data – must be recorded and stored in a reasonable way together with the image data. An image without meta data (such as time, location, photographer, etc.) is for the purpose of cultural heritage preservation virtually worthless.

Quality Control

The digitization process has to be accompanied by a permanent quality control. Both in the digitization process and the recording of metadata errors can be introduced by either technical problems or (more often) human error.

During the actual scanning process (digitization), the following parameters have to be chosen carefully and permanently monitored:

Spatial Resolution

The spatial resolution has to be adapted to the photographic original, i.e. the information content. Important parameters are the spatial resolution of the photographic emulsion, but also the quality of the optical system used to make the original image.

Photometric Resolution (Gray Scale Reproduction)

The brightness range (contrast) has to be reproduced completely. In the digital domain, the numbers of gray values (or, in case of color images, the number of values per primary color) determine the degree of accuracy. Within 8 Bit, 28 = 256 distinct levels can be represented, with 12 Bit there are 4096 levels and 16 Bit allow for 65536 distinct gray levels. The number of bits required is determined by the contrast of the original. However these values alone are almost worthless without a physical interpretation: it could be transmission/reflection (a linear scale), optical density (a logarithmic scale), a visual brightness (CIE L*) or uncalibrated "values". Therefore a proper photometric calibration of the scanner is necessary, and the meaning of the digital gray values has to be recorded with the image. The properties of the different photographic material leads to the following recommendation for minimal photometric resolution:

transparent positive (slide)	≥12 Bit
transparent negative	≥14 Bit
reflection print, linear scale	≥10 Bit
reflection print, logarithmic scale	≥ 8 Bit

For practical reasons (computer architecture) the storage of the data has to be done either in 8 Bit or in 16 Bit format (for color images it is 3 x 8 Bit = 24 Bit or 3 x 16 Bit = 48 Bit):

negative	16 bit
slide	16 bit
reflection print	8 bit

For reflection prints, it is possible to reduce the 10-12 Bit internal representation with a logarithmic transformation or a "gamma"-correction to 8 Bit. In this case, the transformation curve has to be recorded. The calibration for gray value images can be done with gray wedges, whereas the calibration for color images requires the IT 8.7 scanner

calibration standard that exists both for transmission and reflection.

Hardware Calibration (Scanner)

Scanners have 3 properties that require special attention: The so-called "dark current" gives a signal even if the sensor is in complete darkness. Noise introduces small random modifications to the resulting brightness values. Especially in dark areas, noise can effectively destroy fine details. In addition, the illumination may be inhomogeneous and may vary with time. These properties that usually are not constant during the lifetime of a scanner have to be taken into account through proper calibration and regular quality monitoring.

Digitization for long term archival requires a strict quality control. On one hand, the calibration of hardware as described above has to be repeated regularly. On the other hand, a complete visual control of each digitized image in full resolution is required: The following properties have to be monitored through this visual control:

Sharpness

Mechanical wear, vibrations, etc. may change the geometry of the optical system of the scanner and introduce systematic unsharpness.

Dust and Dirt

The scanner (the glass plate) may get dirty through dust and residues from originals.

Geometry

Are all images scanned in the correct way or mirrored?

Scan Errors

Is the image as expected?

This monitoring should accompany permanently the digitization process in order to recognize systematic errors as early as possible.

Another aspect is the completeness and integrity of the digitized collection. Photographic collections that are valuable enough to be preserved form an ensemble that has to remain complete. For large collections, where the digitization process lasts for a long time, the completeness has to be carefully checked, as e.g. an image may be forgotten in the process or a file name may be used twice etc.

Part II: Digital Long Term Archival

The "digital archive" has to preserve the digital facsimile and guarantee the access and readability at least as long as a conventional archive using microfilm would offer.

The long term archival of analogue data, especially photographic images, has to minimize at least two risks: 1) the total destruction of the image e.g. through fire or water, and 2) the intrinsic decay of the images through the natural aging of the medium has to be slowed down as much as possible. Both risks can be minimized by appropriate storage (location of storage, fire protection, climate control etc.). As a consequence of optimal storage, access to the archived images will be very restricted and difficult. Further, the intrinsic decay of the images due to aging cannot be stopped totally but only be slowed down. Since all copying of analogue media such as photographs, film, video etc. always and inevitably imposes a decrease in quality, the original has to be preserved. It is usually possible to interpret analogue media such as photographs without the use of technical means. Metadata such as image description, photographer, timestamp etc. are often recorded together with the image on the same media (e.g. notes on the back of a photograph) and which therefore can not be separated from the image and the risk of loosing the connection of the metadata with the image is usually very low.

With respect to long term archival and compared to analogue data, digital data has completely different properties:

- Given the proper procedures, digital data can be copied infinitely oft (infinite number of generations possible). The "original" and the "copy" are identical and cannot be distinguished. Therefore the notion of an "original" looses its sense in the digital domain.
- There is no slow decay in the digital domain. Digital data either can be read completely, or there are errors reading the data that invalidates the whole data set. In order to recognize the occurrence of read errors, special algorithms ("checksums") have to be used.[*]

What reasons may cause the loss of data within the digital domain? The following short overview is especially targeted to the problems of long term archival where not only the image object alone but also the information about the object, the metadata, has to be stored. For digital archives, another difficulty arises: digital media can only be read and interpreted by using technical devices. A computer tape looks the same to the human observer if it's empty or if it contains images, texts or other data. A vault with 10,000 CD-ROMs that are not labeled is almost without value if there is no other information available. The consequence is that improper archival strategies can lead to a fatal loss of information at many different levels. If only at one of these levels information is lost, the whole archive will be virtually worthless or at least the value will be diminished severely.

There are 6 levels where a loss of information can occur:

- **Storage data**: The information (e.g. the digital image) cannot be found anymore because the knowledge about where and how the image was stored is lost. We define storage data as all necessary data to find and read the archived digital data. This includes location, media type, formats etc. In some sense, these are the metadata of the

[*] Modern storage technologies most often integrate the use of checksums into hardware. In case of non-recoverable read errors the whole data set is declared non-readable. Therefore it is correct to say that in the digital domain there are only 2 possibilities: the data can either be read correctly or it cannot be read at all.

archive itself (and not the metadata of the archive content).
- **Meta data**: the digital data of the archived objects are readable, but the metadata has been lost (such as images, where the image description is lost).
- **File formats**: the files can be read, but the format is no longer known or supported (e.g. no more software to interpret the files)
- **Media formatting**: the media can no longer be read because the formatting of the media is not known. Most types of storage media need some form of low level formatting which is often dependent of the software used for recording the data on the media. For one type of storage media there may exist many different, incompatible formatting methods. For example, DAT tapes may be formatted as NT-backup tapes, as tar-tapes (common for open source systems) or as ANSI labeled tape, each formatting method being incompatible with the others. Other examples are CD-R's, which can be written as ISO9660, Joliet, UDF DirectCD, and in Mac format.
- **Reading devices**: the storage media cannot be read anymore because there are no more working or supported reading devices (e.g. tape drives) available.
- **Storage media**: the storage media can no longer be read because of aging, damage, handling errors etc.

For all these levels precautions have to be taken in order to guarantee the long-term preservation of digital data. In the following the basic criteria will be described.

Procedural Criteria

A digital long term archive requires a total change of paradigm: Whereas the "classical" archive tries to lock away the objects at a secure place touching/using them as little as possible, a digital archive requires the stored (digital) objects to be touched as often as possible: only permanent checking and copying/reformatting of the digital long term archive guarantees the longevity of the stored data. These processes have to follow very strict quality measures to preclude all loss of data:

Redundancy

On all levels, the probability of loss of data has to be virtually zero. Since digital data can be cloned (copied) without loss, the risk of data loss can be increased dramatically through redundancy:

Redundancy of Media

Use of multiple redundant media. The data is stored on several identical sets of media.

Geographical Distribution

The media sets should be stored at different locations in order to minimize the risk of loss through catastrophes such as earth quakes, fires, ethnical conflicts, riots etc.).

Migration

Since the periodical migration of the data due to changing hard- and software is necessary, the migration procedures have to be taken into account throughout the whole design of the digital long-term archive. The migration strategy has to obey the following principles:

Moment of Migration

The migration has to be completed early enough, before a loss of data due to aging or changing technologies can arise.

Periodical Proofreading

The media has to be proofread periodically in order to have early indications about aging or other problems.[†]

Zero Fault Tolerance

All copy processes must be accomplished without errors. This can be achieved by an immediate comparison of the "original" and the "copy" for correctness.

Interleaved Migration

The continuous development of storage technology requires the periodical migration of the data to new media types. In order to increase the redundancy and minimize the risk of data loss, the new technology has to be introduced while the "old" technology is still supported. Therefore, in the long run, all the data should be stored on at least 2 storage technologies, where one of these should be a proven technology.

Quality Assessment of Media

The quality of storage media has to be assessed using the following criteria:

Detection of Recoverable Errors

All known digital recording methods have to deal with recording errors that are inevitable due to small media defects. These defects are detected and completely corrected by built-in hardware and software ("error correction", "recoverable errors"). If a recording/reading device allows determining of the number of corrected errors, it is possible by periodical proofreading to determine the number of correctable errors. This number is a very effective indicator of the aging storage media.

Media Assessment

The quality of storage media can change from manufacturer to manufacturer, and can even change from lot to lot from the same manufacturer. Therefore each lot of storage media should be checked for defects.

[†] Such a periodical proofreading may consist of the bit-by-bit comparison of several redundant storage medias. There should be no differences. If any difference occurs, a third copy is used to determine the proper bit pattern. Therefore at least 3 sets of identical storage media have to be available for proofreading.

Storage Conditions

Digital storage media normally have a limited lifetime. In order to maximize the lifetime, the storage conditions should be held in the optimal range. E.g. for magnetic tape media, this would be 15ºC ±2ºC with a rel. humidity of 20-40% ±5 (SMTPE, RE 103 or ANSI/AES).

Handling Risks

A major risk of loss of data is human error. A storage media such as a tape or a CD may fall to the floor, coffee may be poured over it, it can be lost or the media may be erased by accident etc. In order to minimize this kind of errors, all steps in handling the media have to be planned carefully, have to be documented and very careful handling is required. Even a small error like a wrong label may have a big impact. Therefore most processes should be automated as much as possible, and a strict quality management has to be implemented.

Technical Criteria

On all levels mentioned in the introduction, certain technical criteria have to be fulfilled in order to minimize the risk of data loss:

Storage Data

Storage data is used to organize and retrieve a certain piece of information. In a simple case, it's a list where (on which storage unit) a certain archived object can be found. Storage data is necessary to data retrieval and for proper migration.

Labeling

The storage medias have to be labeled properly, best in a human readable manner, with enough data that the content can be determined without other means. E.g. the label "Museum of fine arts Basel, Collection Sarasin, images 1-250, 12. Dec. 1995" contains much more information than for example tape "nr. 47035466b". If in addition a machine-readable label is attached, it may facilitate the process of migration and increase the efficiency of media handling dramatically.

Bookkeeping/Logging

All handling of the media must be recorded and documented. Such a strict documentation reduces the risk of lost of a media, and in case of problems the history of the media is known.

Storage Forms

The storage data should be recorded in addition to electronic forms also in a traditional analogue form (paper printouts). Since storage data are usually very compact and small, this imposes little overhead and secured future interpretation of the electronic media.

Metadata

Metadata of objects is often stored separated from the objects (image description, date of photograph, photographer, copyright etc.). However, the metadata is an integral part of the object. Therefore the following rules apply:

- The Metadata should be (in addition to storage separated from the object) directly connected to the object data, e.g. a basic set of metadata should be recorded in the header of the object data, or the metadata may be an integrated part of the image (e.g. the title of an image is also scanned as part of the image).
- Metadata are often highly structured and maintained in a hierarchical or relational database. For long-term storage, this database should be transferred into a flat, unstructured form, which is independent of any software or hardware technology (e.g. an ASCII-file). If properly converted, such a flat format can easily be converted back into a future structured format (if necessary for efficient access).
- The metadata have to be stored with the same redundancy and precaution like the actual object data.
- The use of checksums guarantees the integrity of the meta data.

File Formats

File formats used for long term archival of images should be chosen using the following criteria:

Open Standard

The archival format has to be open and completely documented. It should be possible, that at any point in future it is possible to develop appropriate software to read the format using the available documentation. Many standard formats such as the TIFF Format, which is fully documented there, exists open source software to read and write the image files. These formats fulfill this requirement of following an open standard in an optimal way.

Dissemination

The file format should be widely used. Given a high dissemination, the probability that the format will be supported over a long period of time is very high, and many open source and commercial software will support it.

Flexibility

The file format should have a high flexibility to integrate some meta data, calibration data etc. directly within the image file.

Fault Tolerance

Given proper archival procedures, the probability of a bit error within an image file is very low. To even decrease the risk of loss of an image, the chosen file format should have a high tolerance against bit errors. All formats that use loss less or even lossy compression do so by reducing the intrinsic redundancy of an image. If in such a compressed image a bit error occurs, the result can be devastating. The

following image (figure 1) is on the left an uncompressed TIFF image with a bit error, which hardly can be seen. On the right, the same image with an identical bit error, but this time using the JPEG format.

Figure 1. Top, an uncompressed TIFF image with a one bit error. Bottom, the same image in the JPEG format, also with a one bit error.

Media Formatting

Open Standard

Most backup systems use their own, proprietary format for formatting the storage media. An example of a non-proprietary format is the "tar" originating from the Unix/Linux domain. However, for virtually all systems there exists software to read and write tapes in the tar format. Also a proprietary format could be used if it has an exceptional high dissemination (e.g. NT backup).

Hardware

Dissemination

The storage system used should have a high dissemination, and more than one manufacturer should exist. In today's market situation, even big manufacturers may vanish within a short time. If it is the only manufacturer of a storage system, support of the system will be very difficult and the archived data may be lost.

Wear

Since long term archival of digital data requires many cycles of reading (proof reading, copying etc.), wear of the media is a critical factor. With respect to wear, optical systems with contact free reading/writing mechanisms are optimal. For magnetic tape drives, linear systems are considered to have less wear on the media than helical scan systems.

Reliability

The "mean time between failure" which manufacturers data sheets should indicate are a good estimate of the reliability of a storage system.

Exchange

Media written on one drive should be readable on any other drive of the same system. In our experience, this is not always the case. CD-R's and magnetic tape using helical scan methods are notorious in this respect.

Compatibility Across Generations

Given the rapid development of storage technology, often whole families of related storage systems are being developed, where each successing generation has more capacity (e.g. DAT. DDS-1=2GB, DDS-2=8GB, DDS-3=24GB, DDS-4=40GB, DLT: DLT I – DLT IV). Since the live span of the system is often more restricting than the live span of the media itself, the chosen system should have a high degree of backwards compatibility.

Fault Tolerance

Small amounts of bit errors should be recognized and corrected automatically. Most manufacturers indicate the bit error rate (the probability that a bit error is not recognized and corrected).

Ease of Handling

The storage system should be robust and have an easy handling. The probability that a media is destroyed during reading/writing should be very low. There should be the possibility of automatic handling through robots.

Media

Robustness

The media itself should be robust and forgiving bad treatment (e.g. dropping to the floor).

Live Span

The live span should be well known and consistent. Most modern media have a live span that is much higher than the live span of the technology.

Quality Tests

There should be means to test the quality of media either individually or for the system (media – drive).

Dissemination

A high dissemination is of great advantage, as mentioned before.

Capacity

A high capacity reduces the amount of storage data that has to be archived and maintained. For example, for a 1TB archive, it's much easier to handle 10 LTO tapes than 1500 CD-R's.

Conclusion

The criteria described in here will form the basis of future funding of digitization projects with long term archival in mind. These recommendations will be submitted to a consultation to different archives and institutions on federal and cantonal level. It is planned that these recommendation will be officially adopted in the year 2004. More information about the Swiss Civil Protection, Section Protection of Cultural Property can be found on the website http://www.zivilschutz.ch/ (Navigation point: Protection of cultural property).

Some Abbreviations

ANSI/AES	American National Standards Institute/Audio Engineering Society
ASCII	American Standard Code for Information Interchange
DAT – DDS	Digital Audio Tape - Digital Data Storage
DLT	Digital Linear Tape
GB	Giga Byte
ISO	International Organization for Standardization
JPEG	Joint Photographic Experts Group
LTO	Linear Tape Open
SMPTE	Society of Motion Picture and Television Engineers
SPCP	Section of Protection of the Cultural Property (SPCP)
TB	Tera Byte
TIFF	Tagged Image File Format
UDF	Universal Disk Format

Biography

Prof. Dr. Rudolf Gschwind, born 1949, studied Chemistry at the University of Basel, and got 1979 a Ph.D. in Physical Chemistry in the field of Photochemistry. During his studies he got involved with scientific photography. Since 1980 he is the Head of Department of the "Scientific Photography Lab" at the university of Basel, which changed now its name to "Image and Media Lab". During 1985 he did a 1-year sabbatical industrial research at IlfordAG, Fribourg/Marly (Electronic Imaging) and between 1989 - 1999 he did additionally research and teaching at the Swiss Federal Institute (Zürich), Photography Group at the Institute of Physical Chemistry. During this period he also developed new methods for the digital reconstruction of faded color photographs. The main research topics are Image Processing and Analysis, Color Photography, Color Imaging, preservation of the audio-visual cultural heritage.