Guides to Quality in Visual Resource Imaging

5. File Formats for Digital Masters

Franziska Frey

1.0 Introduction

The purpose of this guide is to identify how to contain and formally describe the qualities of digital masters as discussed in Guide 4 in this series. This guide focuses on the features of the containers—the **file formats**—that affect the performance of the digital master and the ability of the custodian of the master to ensure that it persists over time as technology changes. Some parts of this discussion may identify technical requirements for product features that need to be, or may already be, the subject of further research and

development, and may not yet be commercially available in a suitable form.

Because file formats are evolving, this guide generally does not include concrete specifications of file formats for different applications; instead, it gives guidance in what to look for when choosing a file format for the digital images. Before deciding on a specific format, the user will have to check the most recent information on file formats. A list of sources that can help with that search appears at the end of the series.

Choosing a file format is one of many decisions that have to be made when undertaking a digital project. It is a decision that comes relatively late in the process. Especially for the digital master, issues such as openness of the format and longevity are in the forefront. Only a few formats actually comply with archival standards.

Besides longevity issues, several interdependent technical considerations have to be looked at, including quality; flexibility; efficiency of computation, storage, or transmission; and support by existing programs.

Design goals may be conflicting, as indicated in the following list, which is based on one authoritative source (Brown and Sheperd 1995):

- **Memory.** Some formats emphasize the presence or absence of memory. For example, the TIFF format has a design goal of modifying a raster image in place, without reading the entire image into memory. This can be a concern with master files, which are by definition large.
- **Accuracy.** The accuracy of images is sometimes critical and sometimes negotiable. In some cases, the accuracy of the data can be decreased to save memory, e.g., through compression. In the case of a digital master, however, accuracy is very important.
- **Speed.** The ability to access or display a data set at a certain speed is critical to certain applications. In the case of the master file, this is not a concern.
- **Device independence.** This is a very important aspect for master files because they will be most likely used on various systems.
- **Robustness.** Some formats are concerned with transportation accuracy for the data. A robust format contains several layers of defense against corruption, so that if part of the data is corrupted, the data set does not completely fail to produce an image. This is most widely used in television signals or data on audio CDs.
- **Extendibility.** The simplest definition states that a data format can be modified to allow for new types of data and features in the future. A more stringent definition states that a format can be extended without invalidating previously implemented readers of the format.
- **Compatibility.** Compatibility with prior versions of data set definitions is often needed for access and migration considerations. However, compatibility carries a

penalty in performance and in increased size of both code and data.
- **Modularity.** A modular data set definition is designed to allow some of its functionality to be upgraded or enhanced without having to propagate changes through all parts of the data set.
- **Plugability.** Plugability is related to modularity. It permits the user of an implementation of a data set reader or writer to replace a module with private code.
- **Openness.** A standard is designed to be implemented by multiple providers and employed by a large number of users. A truly open standard also has an open process for development and enhancement.
- **Scalability.** The design should be applicable both to small and large data sets and to small and large hardware systems. Thus, access time should increase no more than linearly with increasing data set size.

All of these are good goals for design, and it would be extremely convenient if they could all be satisfied by a single data format. Given current technology, however, it is not possible, for example, to have both real-time access to data and device-independent data. Real-time access requires special and specific hardware devices. Therefore, the result of this set of conflicting goals is a large set of diverse data formats to meet different needs.

The goals of extendibility and robustness are very similar. Both require that a data reader recognize and skip unknown data and resume reading data at some subsequent location in the data stream, when the reader recognizes valid data. It is important to keep in mind, however, that the more complicated the format, the higher the chances of future errors due to bytes that cannot be read anymore.

Functionality is key to any data format. One must be able to store application-dependent data that meet a user's needs. Application data should be separated from graphical data, e. g., in the form of extended file headers.

2.0 Attributes Associated with Performance

2.1 Ability of Files to Open Easily on Any Platform

Most file formats were designed to work best with a particular platform. All computer operating systems (e.g., Unix, DOS, Windows, Macintosh) have strict rules governing file names. The maximum length of a file name and the type of characters it will accept are matters that confront all operating systems.

2.1.1 Encoding Byte Order

Binary encoding refers to a broad range of encoding schemes that are machine-dependent. For files to be readable on different machines, the byte order must be known. The

National Archives and Records Administration (NARA) guidelines therefore state, for example, that file formats should be uncompressed TIFF files with Intel byte order and header version 6 (Puglia and Roginski 1998).

Binary encoded data can use different byte orders and, in the case of bitmap images, can use different bit orders as well. Where there is more than one bit per pixel in a bitmap image, several variations are possible. Each pixel's multi-bit value could be recorded at once, such as a 16-bit value being recorded as two consecutive bytes, or the image could be broken into 16 one-bit-deep bit planes (Kay and Levine 1995).

Byte order usually depends on the processor used in the computer. It, too, can vary. One variation is the ordering of least and most significant bytes (LSB and MSB, respectively) for numbers of more than one byte. In the memory (and generally in the files) of personal computers and other Intel CPU computers, the LSB comes first; in Motorola CPU systems such as Macintosh, the MSB comes first. Data stored in files on these systems usually reflect the native order of the machine (i.e., the order in which data are stored in memory).

Because of these differences, a graphics file format must record some information about the bit and byte order if it is to be used by computers from different manufacturers. A file format developed for a particular computer will generally lack byte-order information; however, if the CPU of that system is known, the correct order can be implied.

Before selecting a file format, therefore, one must know whether the byte order in which the file is written makes the file readable on different systems. For example, when saving TIFF files, Photoshop asks whether the byte order for Intel (PC) or Motorola (MAC) should be used (Blatner and Fraser 1999).

2.1.2 File Naming

Another important issue is the file naming system. To quote Michael Ester (1996):

> "To take a simple example, consider giving an image a name. This can begin as a straightforward task of giving each image file the same name or identifier as the image from which it was scanned. This assumes, of course, that the name will fit in a filename. Next, there are perhaps three or four smaller images created from the master image, or alternatively, versions of the image in different formats. More names needed. The image variations will probably not be stored in the same place, so now we need to come up with names for the places we put them and what goes into these different areas so we don't mix them up. Finally, we need to record all of the names somewhere so that other people and computer programs can find the images and versions of images they need. What starts out as giving one file name to

> an image grows to a many-sided production step, and names are only one characteristic of the image we need to track."

There are two approaches to file naming. One is to use a numbering scheme that reflects numbers already used in an existing cataloging system; the other is to use meaningful file names. Both approaches are valid, and the best fit for a certain environment or collection has to be chosen.

When developing a file-naming scheme, one must have a good understanding of the whole project. What kind of derivatives will be needed? How many images will be scanned? Will they be stored in different places? Will the files be integrated in an existing system (e.g., a library catalog), or will it be a new, stand-alone image base? Because some issues will be system-specific, all these questions have to be discussed with the people running the computer system on which the files will be put.

Windows 98, Macintosh, and UNIX systems allow both longer file names and spaces within the file name than other systems do. Each operating system has its own rules and preferences. The use of special characters, in particular, has to be checked before a naming scheme is chosen.

Pre-Windows 95 Applications do not support long file names. They used an 8.3 system, meaning the file name was eight characters long, followed by a three-character extension (i.e., the characters after the dot). Long file names can be up to 255 characters long and contain spaces. The long file names, the aliases, and the DOS file names can contain the following special characters: $, %, ', -, _, @, ~, `, !, (,), ^, #, and &. Windows 98 creates two file names when a file is not named within the old DOS naming convention. The system automatically generates an alias file name consisting of the first six characters of the long file name, followed by a tilde (~) and a number. In this case, it is important that files can be distinguishable by the first six characters. This might be needed in a networked environment, where many PCs continue to run Windows 3.1x and DOS. Therefore, it has to be taken into consideration whether files will have to be shared over a network with someone who cannot read long file names or who uses the file in a pre-Windows 95 application.

Many extensions have standard meanings and are employed widely. Care has to be taken when dealing with nonstandard extensions because a file extension does not always refer to a certain file format. A list of the most common extensions can be found at Webopedia, an online encyclopedia related to computer technology, http://webopedia.internet.com/TERM/f/file_extension.html. The extension is assigned by the application that created the file. The Macintosh hides its file extensions with invisible type and creator codes, and the operating system deals with them.

Another issue that has to be taken care of is the metadata associated with the different file formats. All the metadata inherent to a specific file format have to be read and made usable for an application or for conversion into another file format.

Since the digital master files might be used on different systems, the following cross-platform issues have to be looked at carefully.

2.2 Scope (Size of Containers)

The operating system and applications determine the maximum file size. At this time, maximum file sizes (usually expressed in gigabytes) exceed by far the file sizes that will be produced for a digital archive.

2.3 Pros and Cons of Various Formats

File formats must be compared in terms of ability to contain:

- Detail (maximum number of pixels)
- Tone (maximum number of bits per pixel)
- Color (color space associated with format)
- Administrative metadata (number of file headers)

Some of these aspects are more pertinent for derivative files. File formats that do not make it possible to keep all the information inherently available in the digital master are not considered valid candidates for digital masters.

From the currently available formats TIFF is the one that can be considered most "archival." It is a very versatile, platform-independent, and open file format, and it is being used in most digitizing projects as the format of choice for the digital masters.

It is also important to have a closer look at the format in which the data is written onto the media, e.g., tar and ISO 9660. This format is dependent on the type of media. Open and non-proprietary formats are also in this case a must in an archival environment. The format that has been used to write the data has to be documented.

2.4 Image Processing (Viability to Create Deliverable Images from Digital Masters)

Choosing the wrong file format or data encoding scheme in form of color space for the digital master can make it impossible to create specific deliverable images (Süsstrunk, Buckley, and Swen 1999; Frey and Süsstrunk 1997). Depending on the specifics chosen,

there are different reasons for this, which are discussed below. Rendering images for a specific purpose might be very useful; however, it has to be done correctly, and the information about the rendering process has to be kept with the image.

2.4.1 Open Versus Closed Formats

When choosing a format, one must consider the software/format combination in the future archiving environment. Will the software packages allow the chosen format and all the additional information, such as tags, to be read? Are certain proprietary transformations needed to open and display the file format? Installing the right filters can solve this, but this makes a file format vendor-specific and requires the future user to have the right reading filters.

On the other hand, it must be ascertained that a software application supports writing files in the chosen file format. With a proprietary vendor-specific file format, this possibility might not exist.

Most vendors realize that it is important that their systems be open. They offer filters to write and read files in the most common file formats. However, saving the files in a format other than the system's native format might make it impossible to use certain special features of the system.

2.4.2 Color Spaces Natively Optimized for the Screen

Color spaces may be optimized for images being viewed on a monitor. The sRGB format, for example, has been developed as an average monitor space for the World Wide Web. It is a useful output space for images that will be displayed on monitors of unknown characteristics. No additional transform is necessary to view the images.

Images in sRGB display reasonably well even on uncalibrated monitors. The format is sufficiently large to accommodate most photographic reproduction intents (see Guide 4). However, sRGB is currently designed only for 24-bit data and leaves no bits for modifying the image or mapping it to another output device. There is a serious mismatch between the sRGB gamut and the CMYK gamut needed for printing.

All in all, sRGB is a valid format for access images that are used only on the monitor. Future repurposing possibilities of this format are very limited.

2.4.3 Color Spaces Natively Optimized for Print

Certain CMYK color spaces (e.g., CMYK SWOP) are used for images that are ready to be printed. These color spaces are used often in the prepress industry. The color separations

that have been created from the RGB files are fine-tuned to match the color gamut of the chosen output device.

Because of differences in color gamut, it is often impossible to repurpose these images into an RGB color space. CMYK images are rarely found in an image archive and should generally not be used for the digital master.

2.4.4 Telescoping ("Russian Doll") Formats such as FlashPix

Tiling formats like FlashPix [http://www.kodak.com/US/en/digital/flashPix] have the advantage of very fast transmission, because only the part of the image that is needed is transferred to the client.

FlashPix has other useful features such as the ability to store viewing parameters (e.g., crop, rotate, contrast, and sharpen) within the file without affecting the original pixel values. The image can also be adjusted for the chosen output device without affecting the original pixel values. All image processing can, therefore, be done in one step before the final rendering of the image. This ensures that no quality is lost due to multiple image processing steps.

Support for FlashPix was strong in the beginning, especially in the consumer market. However, FlashPix did not gain the sustained reception that was predicted and never really caught on. New formats will carry forward the tiling aspect of FlashPix while adding other aspects such as lossless compression and higher bit-depth per channel. New data encoding schemes discussed in JPEG2000 will most likely adopt some of the FlashPix features.

3.0 Attributes Associated with Persistence

3.1 Standards

Standards are an essential basis for sharing information, both over current networks and in the future. Standards will ensure the protection of the long-term value of digital data. Several types of standards are being used. Industry standards are available and are often given stamps of approval from official standards organizations. Postscript (EPS) and TIFF are examples of such industry standards.

3.1.1 Who Maintains Standards

Standards are developed by various standards committees. There are international and national standards groups. The following list contains the names of some of the more

important groups in the imaging field (Brown and Sheperd 1995).

International standards groups	
ISO	International Organization for Standardization
	ISO TC42: Technical Committee Photography
	ISO TC130: Technical Committee Graphic Arts
IEC	International Electrotechnical Commission
ITU	International Telecommunications Union
CIE	Commission Internationale de l'Eclairage
IPA	International Prepress Association
CEN	European Committee for Standardization

National standards groups and associations interested in standards	
ANSI	American National Standards Institute
NIST	National Institute of Standards and Technology, US Department of Commerce
PIMA	Photographic and Imaging Manufacturers Association
SMPTE	Society of Motion Picture and Television Engineers

3.1.2 Opportunities for Participation in Standards Development

Standards are not developed in a laboratory and presented to the world only after they are finished. Standards development is a team effort, and it is possible for individuals to join a standards group. Rules and regulations for participation can be obtained from the group secretary or chair.

Standards represent a consensus of the best thinking of those who are most active in the field as well as of other individuals from widely different backgrounds, training, and interests. This ensures that the views of one particular discipline do not predominate. Standards can resolve differences of opinion and remove the problem of deciding which expert to believe. Consequently, standards are a convenient source of unbiased information.

3.1.3 Criteria to Judge Openness

The openness of a file format can be judged in different ways. One way is to learn whether a particular file format is being used by other institutions. It is usually helpful to look at what large institutions are using, because they commonly spend a share of their large resources on finding the best solution. It can also help to see how many software packages and systems are supporting the chosen file format. If several software packages are allowing the use of a file format, the format will probably be in existence for quite a while. A third idea is to see how easy it is to find full documentation on a specific format that is readily available for every user. On-line resources about standards should be maintained and up to date.

3.2 Archival Format for Digital Photography: Myth or Reality?

The efficient storage and transfer of image data have been sources of difficulty throughout the history of electronic imaging. Also of importance is the need for interchange of metadata associated with images. More sophisticated image processing routines require more information about image capture and display. These factors have resulted in a growing number of image file formats, many of which are incompatible. The ISO42-WG18 Committees (Electronic Still Photography; http://www.pima.net/it10a.htm) reviewed a variety of file formats in an attempt to find one with the following attributes:

- efficient storage of image data in both compressed and uncompressed forms;
- ability to accept image data in a wide variety of formats;
- capability for standardized storage of image attribute information to facilitate image processing;
- extendibility for the accommodation of future needs; and
- reasonably wide levels of current acceptance.

No formats completely satisfied these requirements, and work was done to devise new formats using several popular existing formats as a base. This has led to the work on ISO 12234-2, which is described in the following paragraph (Holm 1996b; ISO 12234-2/DIS, ISO/TC42 1998).

3.2.1 TIFF/EPS (ISO 12234-2)

TIFF 6.0 was taken as the base format. A major goal of the IT10 Committee in designing Tag Image File Format/Electronic Photography (TIFF/EPS) was to identify possibilities for data storage forms and attribute information. TIFF/EPS is also meant to be as compatible as possible with existing desktop software packages. Most of the problems that may arise through the use of TIFF/EPS are a result either of features of TIFF 6.0 that could not be changed without affecting compatibility or of the large number of optional

tags. The IT10 Committee recognizes that increasing the number of formats at this time may not help with compatibility issues, and until all the features desired by users are available in a single format, multiple formats will be necessary. The ISO 12234 and TIFF/EPS standards should facilitate the use of electronic cameras for image capture and make other types of electronic imaging easier to deal with from a photographic standpoint. As this process occurs, currently available file formats may begin to merge.

Time will show whether industry is buying into this standard and making applications compatible. Because most of the large imaging companies have been involved in creating TIFF/EPS, the chances that this will happen are good. TIFF for Imaging Technology (TIFF/IT) is another version of TIFF that has been developed within the graphic arts; however, it does not address a number of points relevant for photography. TIFF/IT and TIFF/EPS will probably merge in the future, perhaps under the umbrella of TIFF 7.0.

3.2.2 Universal Preservation Format

The Universal Preservation Format (UPF) grew out of an initiative of WGBH Educational Foundation. As described by its developer, "The UPF is based on the model of a compound document, which is a file format that contains more than one data type. The Universal Preservation Format is a data file mechanism that uses a container or wrapper structure. Its framework incorporates metadata that identifies its contents within a registry of standard data types and serves as the source code for mapping or translating binary composition into accessible or useable forms. The UPF is designed to be independent of the computer applications that created them, independent of the operating system from which these applications originated, and independent of the physical media upon which it is stored." (WGBH 1999)

Before UPF was defined, information was collected in several ways to find out all the requirements of a digital archiving system for users. Technical requirements for UPF have been set and stated. However, without a clear buy-in from the industry, this standard will not be born.

3.2.3 Implications of Version (e.g., TIFF 6.0 vs. TIFF 5.0)

It is important to state what version of a file format is being used. In the case of TIFF, specifications of both versions have been published and can be found on the Web. However, it is usually advisable to use the newest version of a file format (in this case TIFF 6.0). Issues of incompatibility can arise when using the different versions of TIFF; i.e., a file created in one version might not be readable with a certain type of software. Attention also needs to be paid to the format version used by a scanner or software application for writing the files.

3.3 Color Representation

3.3.1 How to Calibrate Images for a Long-term Repository Rather than to a Specified Device Profile

It is important to know how to calibrate images for a long-term repository rather than to a specified device profile (Süsstrunk, Buckley, and Swen 1999). Closed-production environments, in which skilled operators manage color, are disappearing. The current International Color Consortium (ICC) architecture is not applicable for some professional applications, even in an archival environment. ICC profiles are changing; they contain vendor-specific, proprietary tags; there is no guaranteed backward compatibility; and they are not easily updated.

The need for good, unambiguous color has resulted in the development of new, "standard" RGB color spaces that are being used as interchange spaces to communicate color, or as working spaces in software applications, or as both. Discussion continues about color representation, in the prepress area and for image databases. Various new standards have been developed during the past years. Currently, there is no "one-size- fits-all" approach; i. e., no single color space representation is ideal for archiving, communicating, compressing, and viewing color images. Standard color spaces can facilitate color communication: if an image is a known RGB, the user, application, or device can unambiguously understand the color of the image and further manage it if necessary.

The general color flow of a digital image has to be defined before individual color spaces can be examined. After an image is captured into source space, it may first be transformed into an unrendered image space, which is a standard color space describing the original's colorimetry. In most workflows, however, the image, being the color space of some real or virtual output, is directly transformed into a rendered image space.

When a scanner or digital camera captures an original, its first color space representation is device- and scene-specific, defined by illumination, sensor, and filters. In the case of scanners, the illumination is more or less constant for each image. When images are archived in this space, scanner or camera characterization data (e.g., device spectral sensitivities, illumination, and linearization data) have to be maintained.

The purpose of an unrendered image color space is to represent an estimate of the original's colorimetry and to maintain the relative dynamic range and gamut of the original. XYZ, LAB, and YCC are examples of this type of color space. Work is under way on defining an unrendered RGB space, called "ISO RGB." Unrendered images need to be transformed to be viewable or printable. Unrendered image spaces can be used for archiving images when it is important that the original colorimetry be preserved so that a facsimile can be created later. The advantage of unrendered image spaces, especially if the

images are kept in high-bit depth, is that they can later be tone- and color-processed for different rendering intents and output devices. However, reversing to source space might be impossible later because the transformations used to get to the rendered space are nonlinear or the transformations are not known.

Images can be transformed into rendered spaces like sRGB, ROMM RGB, and Adobe RGB 98 from either source or unrendered image spaces. The transforms are usually nonreversible, because some information from the original scene encoding is discarded or compressed to fit the dynamic range and gamut of the output. Rendered image spaces are usually designed to closely resemble some output device conditions to ensure that there is minimal future loss when converting to the output-specific space. Some rendered RGB color spaces, such as sRGB, are designed so that no additional transform is necessary to view the images.

With the exception of the graphic arts, images are rarely archived in output space, such as device- and media-specific RGB, CMY, or CMYK spaces.

The best thing to do at this time is to characterize and calibrate the systems and scan into a well-described, known RGB color space. In this way, it will be possible to update just one profile if another space is chosen later.

3.4 Compression

Compression is mainly an issue for transferring data over networks. Image compression in an archival environment must be evaluated carefully. At present, most institutions store master files uncompressed. Instead of adapting the files to fit current limitations associated with bandwidth and viewing devices, the digital masters remain information-rich, ready to be migrated when new, perhaps better, file formats or compression schemes become available. Visually lossless compression might be used for certain types of originals, such as documents, where legibility is the main issue.

Compressions may be numerically or visually lossless. Numerically lossless means that no data are lost because of compression; all data are recoverable to their original precision. Visually lossless (but numerically lossy) means that some data are lost during compression but the loss is not discernible by the human eye (Brown and Sheperd 1995).

A position on lossy versus lossless compression is a matter of weighing different issues (Dale 1999). Compression decisions must be combined and weighed with others in the imaging chain. However, it has to be kept in mind that *if one crucial bit is lost, all of the file information might be lost*, even in the case of lossless compression.

3.4.1 How to Avoid Proprietary Traps

Some institutions have decided to use a specific proprietary file format that includes a form of compression. While this might be a good solution for the moment, any proprietary solution may cause problems later. If a company decides to discontinue support of a certain system, there might be no way, or only an expensive way, out of this trap. A contract should state that if the company does not support a certain system anymore, it will provide the full description of the system. In most cases, companies will not be willing to sign such a contract. Thus, a proprietary system is not a valid solution in an archive. Another problem with proprietary solutions is that it is never 100 percent clear to the user what is being done to the bits and bytes of the image file.

3.4.2 File Sizes

The full benefit of compression, i.e., having much smaller files to store, comes from very high compression ratios-something not to be considered with master files because of the loss of quality. A lossless compression scheme reduces the file size by about a factor of two.

3.4.3 Speed of Decompression

The length of time algorithms take to compress and decompress high-quality, high-resolution images must be kept in mind. This time is much longer at higher-quality settings (lower compression ratios), which are appropriate for master image files, than at lower-quality settings. Longer compression and decompression times affect processing time and users, and minimize the potential benefit of file compression. Higher compression ratios (lower image quality) are more appropriate for use with lower-resolution access images, where users can derive greater benefit because of the faster compression or decompression times and where loss of image quality is of less concern.

3.5 Storage

The long-term preservation of visual resources is very demanding. The principles of secure preservation of digital data are fundamentally different from those of analog data. First, in traditional preservation there is a more or less slow decay of image quality, whereas digital image data either can be read accurately or cannot be read at all. Second, every analog duplication process results in a deterioration of the quality of the copy. Repeated duplication of the digital image data, by contrast, is possible without any loss.

In an idealized traditional image archive, the images would be stored under optimal climatic conditions and never touched again. Consequently, access to the images would be severely hindered while the decay would be only slowed. A digital archive has to follow a

fundamentally different strategy. The safekeeping of digital data requires active, regular maintenance. The data have to be copied to new media before they become unreadable. Because information technology is evolving rapidly, the lifetime of both software and hardware formats is generally less than that of the recording media.

The fundamental difference between a traditional archive and a digital archive is that the former is a passive one, with images being touched as little as possible, while the latter is an active one, with images regularly used. However, this often works only in theory. If a document is known to be available, it is likely to be used. Therefore, in practice, original documents are frequently handled as soon as they are available in digital form. The future will show whether a high-quality digitization can satisfy some of the increased demand for the original.

The digital archive needs an active approach in which the data, and the media it is recorded on, are monitored continually. This constant monitoring and copying can be achieved with a very high degree of automation (e.g., using databases or media robots), and is quite cost-effective.

Professionals in photographic preservation rarely differentiate between visual information and the information carrier itself. In traditional photography, such a distinction is superfluous, because visual information cannot be separated from its support. Therefore, the conservation of an image always implies the conservation of the carrier material. The situation is quite different with digital images, because the medium can be replaced.

The distinction between information and support leads to new reflections about the conservation of digital information. The interpretation of numeric data requires that the medium on which the data are recorded be intact, that the reading machine be working, and that the format in which the data are recorded be well known. If any of these prerequisites is not met, the data are lost.

One of the major obstacles to the long-term preservation of electronic media is the lack of standards (Adelstein and Frey 1998). Existing industry standards tend to be distillations of vendor responses to the imperatives of the marketplace. Preservation is seldom a priority.

3.5.1 Choice of Media

Much has been written about the stability of digital storage media, and several reports and test results are available. However, standardization in this area is still lacking, and it is often difficult to compare the results of different tests. Media stability is just one of the issues that must be evaluated. Having a migration plan that considers all the issues is difficult but it is the only way to ensure that data will survive.

It is also important to consider the hardware/media combination for writing the data. Hardware is often optimized for media from a certain manufacturer.

Finally, storage conditions for the media need to be chosen correctly.

3.5.2 Magnetic Media

The ANSI group that worked on magnetics recognized that the critical physical properties are binder cohesion, binder-base adhesion, friction, clogging of magnetic heads, dropouts, and binder hydrolysis (Smith 1991). Magnetic properties of interest are coercivity and remanence. The group agreed on test procedures for adhesion, friction, and hydrolysis, but cohesion and dropouts are difficult to measure and require extensive development work. They are also very system-dependent.

The consumer is left without a recognized specification to use in comparing tape products, and the manufacturer without a standardized procedure to use in evaluating tape life. For the user, the only option is to purchase tape with recognizable brand names.

It is well accepted that good storage conditions prolong the life of tapes (Van Bogart 1995). The accepted recommendations have been incorporated into an ANSI document on the storage of polyester base magnetic tape that was published in 1998 as IT9.23 (ANSI/ PIMA IT9.23-1998).

The ANSI document covers relative humidity, temperature, and air purity. Two types of storage conditions are specified, one for medium-term storage and the other for extended-term storage. The former is for a life expectancy of 10 years, and the latter is for materials of long-term value. More rigid temperature and humidity conditions are specified for extended-term storage; such conditions should extend the useful life of most tapes to 50 years.

Table 2. Storage conditions for magnetic media

	Temperature (ºC)	Relative Humidity (%)
Medium term	23	50
Extended term	23	20
	17	30
	11	50

The lowest recommended temperature is 8° C, because subfreezing storage may create problems. Another storage concern is the avoidance of external magnetic fields.

Work has been started on recommendations for the care and handling of magnetic tape. Topics to be covered include cleaning, transportation, use environment, disaster procedures, inspection, and staff training. Work in the different standards groups has to be followed closely.

3.5.3 Optical Disc Materials

Unlike magnetic media, optical discs are not only manufactured in a variety of sizes but also can be composed of very different materials. The most common substrates are polycarbonate and glass. The image-recording layer features various organic or inorganic coatings; since the discs operate by several different mechanisms, many different coatings can be found. For example, write-once discs can record information by ablation of a thin metallic layer or a dye/polymer coating, by phase change, by metal coalescence, or by change in the surface texture. Read-only discs have the surface modulated by molding of the polycarbonate substrate. Erasable discs are based on magneto-optical or phase-change properties. Despite this vast dissimilarity in composition, optical discs have an important advantage over magnetic materials, namely, that their life expectancy is more certain. Optical discs are recorded and read by light and do not come into contact with moving or stationary parts of equipment. Therefore, their useful life is mainly determined by the properties of the material itself; physical wear and tear is less of an issue than it is with magnetic tape. As a result, several laboratories use the Arrhenius method to predict the longevity of discs. A method for testing the life expectancy of CD-ROMs has been published (ANSI/NAPM IT9.21-1996).

Optical discs can fail by a number of different mechanisms, such as relaxation of the substrate, which cause warping; corrosive changes in the reflecting layer; cracking or pinholes; changes in the reflection of any dye layers by light, pressure, or crystallization; or breakdown of the disc laminate by adhesion failure and layer separation. Of particular interest to the consumer is how long optical discs will last. Various tests reported that their life expectancy ranges from 5 to more than 100 years, depending on the product.

Appropriate storage conditions can prolong the life of optical discs, regardless of the inherent stability of the material. The recommended environmental conditions are 23 °C and a relative humidity between 20 and 50 percent. Lower levels of temperature and RH provide increased stability, with the lowest specified conditions being -10 °C and 5 percent RH. The standard document (ANSI-PIMA IT9.25-1998) dealing with the storage of optical discs also covers magnetic fields, enclosures, labeling, housing, storage rooms, and acclimatization. Particular care should be given to maintaining a low-dust and low-dirt environment. Another important consideration is to avoid large temperature and

humidity variations. Protection from light is vital for many writable CDs.

3.5.4 Accessibility to Data (Off-line Versus Near-line or Online Access)

There are several approaches to image access. For the digital master, file size and security issues drive the decisions on the storage type chosen. Many institutions keep their digital masters off line, (i.e., on tapes or other storage media). In any case, it is advisable to have at least two back-up copies of the master files stored off-line, in different locations, and under recommended storage conditions. Often one version of the master file is stored near-line (e.g., on a tape robot system). If a file is needed, it is retrieved from the tape; this usually can be done in minutes. For the user, it appears that the image is stored online. Sometimes images are not made available for the outside world but are kept behind a firewall to prevent abuse. This might not be necessary in the future, when safer watermarking techniques are available. Online and near-line access is definitely the future for image databases to more rapidly get images to the customers, but solutions such as this are still a few years away. Faster networks with broader bandwidths and solutions to security issues are among the things that are still needed.

It is also important to keep a number of back-ups in different places. Sharing images among various institutions will likewise ensure that images will survive. This idea suggests that new approaches to data security and new copyright laws are needed.

Theoretically, a digital master might not be accessed until the next time the archive is copied onto the next-generation storage medium in a new file format. This will probably occur within five to seven years. At that point, a source code to interpret the data needs to be readily available.

3.5.5 Is Automatic Error Correction and Detection Viable for all Formats?

All digital data is recorded in the form of binary numbers. However, plain binary representation is very rarely used. Rather error correction and data compression are applied, often in combination. The misinterpretation of an on/off switch always leads to a significant error in the interpretation of data. In order to cope with this inherent error, redundant information is added to the plain digital data. A simple form is the parity bit. This method has been replaced by very elaborate error correction codes like the cyclic redundancy check (CRC). CRC not only allows detection of errors but also allows correcting them, if not too many bits within a group have been misinterpreted. The principle of all error correction schemes is to add redundant information. On a hardware level, digital storage devices use error correction in order to guarantee a certain level of data quality. This error correction is performed automatically without the user being aware of it. However, the internal error correction rates (and their development with time) may be an interesting estimate of the quality of a storage medium. Furthermore, it has to be

kept in mind, that the error correction scheme needs to be known and documented, if the data is to be readable in the future.

4.0 Documentation

4.1 File Headers

Michael Ester (1996) has written that:

> Digital images are beginning to stack up like cordwood as museums and archives plunge into more and more electronic application. Yet from a management standpoint, it is not at all clear that collections of digital files are any easier to manage than collections of film and prints. At least with photographic materials, institutions have had decades to develop filing and recording methods. This is not the question of how to describe the contents of an image, which is an entirely separate discussion. Rather, production and management data answer the questions of: what is the source of this image; how was it created, what are its characteristics, where can it be found, and what is it called. Pointers within each record maintain the interconnections among versions of images and where they reside. There is no one time or place where all of this information is acquired: reproduction data are entered when the file is received; image capture variables are recorded during scanning; image characteristics and linking references are added in a subsequent production step.

Production and management data consist of all the technical information needed for further processing of an image. One place to put all that information is the file header.

It still has to be decided what image characteristics should be included in the header. Participants at the NISO/CLIR/RLG Technical Metadata Elements for Images workshop that took place in Washington, D.C. in the spring of 1999 started to address this issue (Bearman 1999). Using targets and including this information in the file will be the best method to go for successful data migration.

A tremendous amount of data are often lost when storing 24-bit data instead of raw data. The higher the quality of the digital master, the more technical data will have to be kept with the images and, probably, stored in the file header. This leads to the conclusion that the file format must be able to store a variable amount of different data in the header.

4.1.1 Advisability of Embedding Gray Step/color Patch Values

Gray step and other target values used during image creation need to be stored with the

image file. One way to do this is to keep these values in a separate database and to have a pointer in the file header point to the particular record. However, this approach bears the danger of losing the connection between the values and the image file, since two databases have to be updated and kept over time. In the end, it seems advisable to use a file format that can put these values into the header. It is also advisable to automate this process as much as possible in order to keep things consistent and to keep errors to a minimum while populating the header. Much work remains to be done to make this possible.

Other document information, such as the document name and page numbers, will also have to be recorded. Embedding them into the file header minimizes the chances of losing that information over time. It also makes it easier to distribute images, because only one file, instead of an image file and the database with all the information, has to be shared.

This approach, however, requires that software read all the header information. Standardization will help to do so.

4.1.2 Fitting Documentation into the Workflow

All documentation must be part of the workflow, from the beginning of a project. What is not documented today will probably never be documented.

4.1.3 Populating Headers During Scanning, Post-scanning, or Both?

Headers will have to be populated at several stages during the workflow. Some data, such as details about the scanning process and technical data about the software and hardware being used, will be put there during scanning. After the files are scanned and evaluated, other information, such as processing details, will have to be added.

4.1.4 Pros and Cons of Creating Documentation Outside the File Header

Creating the documentation outside the file header might be necessary for workflow purposes, (e.g., if this information is also needed in other parts of the working environment). This means that a separate database is often being created apart from the database that contains the images. To facilitate migration, it is always advisable to put the information into the file header as well.

4.1.5 Risks of Losing Information in File Headers

The process of populating the file header must be well documented; for example, one must state which field or tag contains what information, and ensure that this is done consistently over time. Doing so will also make it possible to prevent information loss during migration. Images, as well as file header information have to be checked when

migrating files. Part of this process might be automatic, but it is always advisable to open some file headers and see whether the information has been transferred correctly.

4.1.6 Controls that Can be Used to Mitigate Potential for Loss During Migration

One way to prevent accidental loss prescribed for TIFF/EPS files is to store them in read-only mode. This will prevent accidental loss of important TIFF/EPS tag-value information if the image is edited by a non-TIFF/EPS-compliant application. TIFF editors generally remove unknown tags when saving or updating an image file to maintain the integrity of the TIFF file, since the unknown tags might not apply to the edited image.

GUIDES HOME | **DLF HOME** | **RLG HOME** | **CLIR HOME**