Developing Specifications for Archival Digital Still Images

Franziska S. Frey
Image Permanence Institute, Rochester Institute of Technology
Rochester, New York
Stephen Chapman
Harvard University Library
Cambridge, Massachusetts

Abstract

Museums, archives, libraries, and commercial stock houses all around the world are busy converting their holdings into digital form. Digital imaging technology offers distinctive advantages to institutions in accessing their photographic collections. In the hands of expert operators, today's best digital imaging hardware is capable of representing almost any type of photograph with such visual quality that reference to the original material is unnecessary for most purposes, and the original materials can be stored away in an appropriate environment. However, the required investment for digital image conversion is tremendous. Furthermore, digital image conversion requires a deep and long-standing commitment to traditional preservation, the full integration of the technology into information management procedures and processes, and significant leadership in developing appropriate definitions and standards for digital preservation.

Introduction

Despite all the possibilities for manipulating digital images, quality choices made when files are first created have the same ‘finality” that they have in conventional photography. Quality choices include image quality, usability, and functionality of images. They will have a profound effect on project cost, the value of the final project to the users, and the long-term usability of the digital images. Requirements for all of these aspects therefore have to be established carefully before a digitization project starts. Quality control tools to check imaging systems, digital masters, and derivatives have to be created and established in the field to ensure the long-term value of the digitized files.

A major challenge in creating digital collections that will survive for a longtime is to build systems defined broadly as ‘digital repositories” that maintain functionality and quality intrinsic to images. One management strategy, migration, proposes to preserve image data by copying files to new formats at designated intervals. The premise that underlies migration is the same as that which informs new concepts of preservation: digital technologies offer the unprecedented opportunity to preserve content without any loss of information from generation to generation.

Cultural institutions have been focusing primarily on defining descriptive metadata for the purpose of discovery and identification, and comparatively little work has been done to codify technical attributes of digital images and their production. Technical metadata is necessary to support two fundamental goals: to document image provenance and history through production metadata, and to ensure that image data will be rendered accurately on output to screen, print, or film. Ongoing management of these core functions will require the development of applications to validate, process, refresh, and migrate image data against criteria encoded as technical metadata.

Setting Goals

Some of the following findings and suggestions are based on work done within the Harvard College Library Digital lmaging Group, later referred to as HCL DIG.[1]

The best-managed conversion projects have clear goals.[2] Brainstorming, the first phase of project management, is the time to talk about outcomes. Too often there is a tendency to dive right into questions of technology—e.g., which scanner should I buy?—before articulating the purposes that digital reformatting must serve. Setting goals is a process of thinking about things from several angles before writing project plans. What are the possible outcomes for the collections? What are the potential benefits to the users, to collection managers, and to the institution? What is a reasonable price—in time and money—to invest in new procedures, systems, and services? Is self-publishing a good idea, or are partnerships a better course to follow? Is this the right time to begin digitizing a collection?

Archival Versus Deliverable

The principles of secure preservation for digital data are fundamentally different from those for traditional analogue data. First, while in traditional preservation there is a more or less slow decay of image quality, the digital image can either be read accurately or cannot be read at all. Secondly, every analogue duplication process results in a slight deterioration of the quality of the copy. The duplication of the digital image data is possible without any loss at all. In a traditional image archive the images should be stored under optimal climatic conditions and ideally never be touched again. As a consequence, access to the images is severely hindered while the decay is only slowed down. A digital archive has to follow a fundamentally different strategy. The safe keeping of digital information requires an active and regular maintenance of the data. The data have to be copied to new media before they become unreadable. Since information technology is evolving rapidly, the lifetime of both software and hardware formats is generally less than the lifetime of the recording media.

It is imperative that the involved parties are clear about the difference between "archival" and "deliverable." An archival file has a very low risk factor, meaning that we are confident that neither its integrity nor its functionality will be lost when the format must be migrated in order to remain compatible with image processing applications. A deliverable file can have the same image quality, but depending on file format and compression choices, there is a higher risk of obsolescence, but not total loss if an archival version has also been created and saved.

Putting Things Into Institutional Context

The following points are part of a document describing a necessary framework to start a digitization project within an institution.[3] It is imperative to keep in mind that, besides defining the technical parameters, there are other issues that have to be looked at in a broader perspective of the whole institution where the digitization effort takes place.

- Digitization is a tool, not an end in itself. Selection of photographic materials for digitization should be based on a thorough understanding of the nature and potential use of the collection.
- A digital project starts well before the scanning of the first picture. Investments made in careful planning to define the aims, priorities, technical requirements, procedures, and future use are essential for an efficient workflow and a result that meets expectations.
- Digitization of photographic collections differs fundamentally from digitization of text or line art. The creation of a digital image requires photographic expertise with ethical judgment. Even with the best equipment, capturing the essence of photographs in a digital format is a sophisticated activity and can never be a routine job like the production of photocopies.
- Digital images of photographs constitute active collections that require regular maintenance. Provisions to upgrade digital collections to keep pace with the changing computer infrastructure should be made at the start of a project. This is necessary to avoid digital collections created at considerable cost becoming inaccessible over time.
- Digitization of photographs should not be the sole responsibility of one department. A good digitization project is conceived as teamwork, combining expertise in imaging, collection management, information technology, conservation, descriptive methods, and preservation strategies.
- In every project for digitization of photographs the input of specialists in photographic preservation is essential. Their advice is required for the best selection of materials. They should be consulted on how to integrate preservation measures in the workflow, on how to handle fragile materials, and on the equipment used to avoid damage to the originals.
- Preservation specialists should be trained to advise on strategies for management of digital assets that are in line with the overall preservation policy of the institution.
- Museums, archives, and libraries have a strong interest in the development of international standards on which a strategy for the preservation of digital collections has to be built. Their active involvement is essential to ensure that the long-term view of heritage institutions is represented in groups working on standards.

In Harvard's libraries and archives, digitization workflows and imaging specifications are project-specific. Handling requirements, functional requirements for digital reproductions, and budgets vary according to materials, users, and funders/owning libraries, respectively. Bringing curators and stakeholders into the production environment to evaluate test scans early in a project is one of the great advantages of having an in-house operation. All parties benefit. Curators have an opportunity to assess all potential risks to the source materials. Designated users of the digital surrogates have an opportunity to see which specifications best meet their needs. For collection digitization projects in this category, defining scanning specifications is a group process based upon a review of sample images. These discussions tend to focus upon the preferred size (resolution) of delivery images.

Findings from Harvard projects show that it is best for the owning library or archive to create descriptive metadata, while scanning can be outsourced successfully. The cost and quality benefits from an in-house operation became apparent when accounting for the significant time and skill needed for other activities: materials preparation and transfer, digital image processing and quality control, structural and administrative metadata creation, file management, and storage. The goal is to enable curators and project managers to work with a single, flexible organization that can offer a full range of preservation and reformatting services for printed and visual materials.

Setting Parameters and Ensuring Quality Control

Judging image quality is a complex task. The viewer has to know what he or she is looking for. The visual literacy required for looking at conventional images has to be translated for digital images. A great deal of research must be done before it will be possible to fully understand image quality of complex images.

It is helpful to ask users whether their expectations are met when comparing the digital master with the photographic original. In the best of cases, there should be no differences in the appearance of the two.

To achieve this goal, one must control the viewing environment. Systems should be setup and calibrated carefully. This is often not done properly, and problems ensue. Moreover, even when systems are calibrated, measurements may not be taken correctly.

Image Representation[4]

One of the basic things to understand is that a great deal of image quality can be lost by processing the images at the moment they are captured. Consequently, a high-quality digital master image should be archived as raw data. Any further transformations are dependent on current engineering practices and knowledge, which might be improved in the future. Or, to quote B. Fraser:

> Without a doubt, color management will improve in the future as we gain a deeper understanding of the human visual system. If we want to be able to capitalize on future improvements, we should make sure that we archive the raw captures, along with as much information as possible about the way the image was captured, including, if at all possible, not only the best available device profile for the capture device, but also a record of the spectral power distribution of the light source used for the capture, and a record of the spectral responses of the filter set used by the capture device. Both records should be obtainable from the respective vendors, albeit with some prodding required. The purpose of this archive is to avoid the need to subject originals to the capture process in the future.[5]

When a scene or original is captured, either by a scanner or by a digital camera, its first representation is device- and scene-specific, defined by illumination, sensor, and filters. In the case of scanners, the illumination should be constant for each image. With digital cameras, the illumination can vary from scene to scene, and even within a scene. When images are archived in sensor space, camera or scanner characterization data, such as device spectral sensitivities, illumination, system MTF, and linearization data (opto-electronic conversion function OECF), have to be maintained so that further color and image processing is possible.

If this is not possible, an unrendered image space should be chosen to contain a colorimetric estimate of the original. An unrendered space maintains the relative dynamic range and gamut of the scene or original. The advantage of unrendered image spaces, especially if the images are encoded in higher bit-depth, is that they can always be tone- and color-processed for all kinds of different rendering intents and output devices at a later date. There are several color spaces that can accommodate unrendered data: CIEXYZ, CIELAB, Photo YCC, ISO RGB, and RIMM RGB. Unrendered images will need to go through additional transforms to make them viewable or printable.[6]

Rendered image spaces are color spaces based on the colorimetry of real or virtual output characteristics. The transforms are usually image-specific and nonreversible, as some information of the original scene encoding is discarded or compressed to fit the dynamic range and gamut of the output. For example, an image that has been pictorially rendered for preferred reproduction cannot be re-transformed into a colorimetric reproduction of the original without knowledge of the rendering transform used. One therefore has to be careful when selecting this approach in an archival environment. There are several color spaces that can be used to encode rendered images. The choice is usually device- and/or application-dependent. From an archival point of view, the color space needs to be well defined to allow for future output rendering. The current method is to include an ICC (International Color Consortium) profile with the image that contains gamut mapping information from the sensor color space to the profile connection space.[7] There are still some inconsistencies in the ICC profile specification, and backward compatibility is not guaranteed.[8] From an archival point of view, it is therefore preferable to encode images in a well-defined color space such as sRGB, CIELAB, or ROMM RGB and create profiles "on the fly" when needed.[6]

Encoding

Linear encoding (in intensity) is acceptable when high bit-depth information can be retained and file size doesn't matter. In most cases, a nonlinear, perceptually compact encoding (nonlinear in intensity, but linear in lightness or brightness) is preferable, since the visual artifacts due to image processing would be equally visible across the tone scale.

Depending on the color space used to archive the image, 8-bit encoding might not be enough. Banding effects can appear, depending on image color distribution, editing, and/or color space conversion, especially if the color space is large or unlimited. However, 16-bit/component RGB is not widely supported yet in either applications or file formats. Images archived in sensor or unrendered representations will go through extensive image processing and color-space conversions and should be encoded in higher bit depth.

File Format

Ideally, archived images should be saved in a standard file format whose source code is readily available. 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.

Of 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 master (archival copy). It is also important to use standard formats for file storage, e.g., tar and ISO 9660. This format is dependent on the type of media. Open and nonproprietary 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. Checksums are also routinely used to ensure that data are not lost due to incomplete transfer in routine data management tasks or due to "decay" when stored on media not routinely accessed.

Compression must be judged as part of the format when looking at file longevity. It needs to be reversible when a file is opened years later. This has proven to be problematic, since the compression code can be lost when even a couple of bytes get lost.

Metadata

Metadata, literally "data about data," has become a ubiquitous term that is understood in different ways by many different professional communities. As these communities, and also the repositories and computer systems, come together to make the information age a reality, it is essential that we understand the critical roles that different types of metadata can play in the development of effective, authoritative, flexible, scalable, and robust image database systems.[9]

Traditionally, cultural heritage and information professionals such as museum registrars, library cataloguers, and archivists have used the term "metadata" to refer to cataloguing or indexing information that they create to arrange, describe, and otherwise enhance access to an information object. But there is more metadata than description. Repositories also create metadata relating to the administration, accessioning, preservation, and use of collections. All of these perspectives on metadata become important in the development of networked digital information systems, but they lead to a very broad conception of metadata.

The metadata information must be in a form that is easily readable. Metadata of digital information can be stored in three ways:

- In the header of the digital file
- In the file name and directory structure
- In a separate database

The storage of information in so-called "medium independent" format is considered to be the only way to ensure that data can be used in the future and be easily exchanged between different databases and applications. Although it is very labor-intensive, structuring of data in SGML format is considered to be the best way to store information. It is predicted that XML will replace HTML as the main mark-up language in the near future. It is no surprise that encoding and structure standards such as the Encoded Archival Description (EAD) and Dublin Core (DC) use SGML or XML.

There are various standardization efforts for technical metadata currently under way.[10, 11]

Implementations by HCL DIG

Guided by ISO 3664 recommendations, HCL DIG calibrates its monitors to 5500ºK, gamma 2.0 for an optimized white point for soft proofing.12 Archival images are saved as RGB files. Adobe RGB 1998 is used as the working color space when inspecting archival images in Adobe PhotoShop®. To date, all archival images have been saved as uncompressed 24-bit TIFF files. The Kodak gray scale Q13 has been included in-frame with all manuscripts and images photographed with the Leaf camera. Including this target—a common practice in traditional studio photography—is intended to provide an objective tone and color reference. When photographing a collection of items with shadow and highlight densities that fall within the tonal range characterized by the target, HCL DIG photographers set the tonal response of the digital camera system once per photography session by using the known density values of the target's grayscale patches. By using objective references as guides, it is the hope to create archival images that, to the extent possible, accurately represent the original items photographed, and not subjective, aesthetic biases.

HCL DIG also has incorporated the use of system targets for quality control. Since every scanner and camera introduces biases, generates noise, and fails to reproduce details in certain tonal regions, technical targets document the performance of a system in a given configuration. HCL DIG uses a variety of targets to characterize the camera setup. As long as the lights and camera position remain stationary, the RGB data in the scanned target images accurately characterizes the output of that setup for its associated batch of images.

Production archival images are sometimes saved in addition to the archival images that include in-frame targets. These cropped, high-resolution images are color corrected using an ICC profile to convert the raw image from the camera into the Adobe RGB 1998 color space. Derivative images are then sharpened and saved with an embedded Adobe RGB 1998 profile. In creating delivery images, either in batches or individually, one does rely upon subjective aesthetic assessments of images on screen to guide the decisions.

Longevity of Digital Files

When building digital repositories, various issues ruling the longevity of digital files have to be taken into account. It is also important that users and builders of the repository agree on certain definitions of terms when talking about the repository.

First and foremost, only data that one knows how to manage should be accepted into the digital repository

Metadata information has to be incorporated to ensure the longevity and usability of the files. Targets can be part of every scanned image, or the target information can be put into the image header. Putting a target into every scan might be particularly appropriate for very high-quality scans of a limited number of images. However, the target area will make the file size bigger. For large collections, a better approach might be to characterize the scanner well and to include this information in the file header. In this case, the images can be batch-scanned and processed later.[13]

Information about the use of targets must be well documented. Target measurements, numbers, and types need to be linked to the files that have been scanned with them, preferably by putting that information into the file header.

Software and Hardware Issues

Archival issues have never been a prime concern in the computing world. The rapid pace of development in the computer industry has led to extremely short product cycles. Generations are being counted in months and not years or decades. New versions of software products have a lifetime of at most a year, and computer hardware seems to become outdated in the time between ordering and receiving the hardware.

Hardware Compatibility

The longevity of digital recording media is only part of the story. While the coded information may be physically readable for a long time, the device to read it has to exist and the formatting code has to be known. Because of rapid developments in storage technology, the reader/writer of a certain type of digital storage medium is likely to be produced for only a short time. Then, a new enhanced device will be put on the market. For all known media types, a new generation can be expected about every two years. Generally, a new generation is able to read and write to media of the previous generation, and only read the media of the generation before that. All earlier generations are incompatible. As a result, media has a useful lifetime of about five years. After that, there usually exist no more devices to read the media, and support for these older devices may no longer be available.

Software Compatibility

Even if the digital data can be transferred to the computer, the data still has to be interpreted to be useful. Thus, the longevity of digital data is also determined by the capability of software to read “old” formats. Obsolescence occurs when current image processing applications cannot parse data in its stored format. Trained staff who monitor industry trends in standards and software development are envisioned to be integral to the trustworthy archiving of digital images. Some of the metadata that is stored with images to interpret the data will be consulted by managers to determine if images should remain in their native format or be transformed programmatically—“ migrated”— to formats supported by new or soon–to–emerge applications. Generating these transformations *with no loss of quality or functionality* is the key to digital preservation. The digital representation of an image makes sense only if metadata, like the image size in pixels and the meaning of each number, is known. For convenience, this metadata is often stored together with the actual image data. Depending on the header structure of the file format, software applications might not be able to interpret that data correctly.

Digital Storage

All digital data is recorded in the form of binary numbers. However, plain binary representation is very rarely used. Instead, 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.

Digital information usually contains a certain degree of semantic or syntactical redundancy. In order to efficiently store binary data, data compression algorithms are applied. However, compression is mainly an issue for transferring data over networks. Compression in an archival environment has to be evaluated very carefully. However, the user might not have a choice, since pictures from digital cameras are often already compressed in the hardware to allow for faster download.

Error detection and data compression are very often used in combination. On a hardware level, digital storage devices use error correction in order to guarantee a certain level of data quality.[14] 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.

Future Digital Asset Management Systems (DAMs) should include features to monitor the error rate and use appropriate error correction systems.

Digital Media

Much has been written about the stability of digital storage media.[15] As mentioned above, media stability is just one factor in determining permanence. A migration plan that accounts for these issues is ultimately the only way to ensure that data will survive. Another important factor is the hardware/media combination for writing data. Often hardware is optimized for media from a certain manufacturer.

As with all materials, improvements in storage conditions will result in improved life.[16, 17]

It is also important to keep a number of back-ups in different places if images are to survive. While this seems like a common-sense issue, experience has shown that users, even professional archives, often neglect to back up their data properly.

Reality Check/Cost Models

Often, project planning and budgeting stops after the creation of the digital assets. However, in specifying requirements for archival digital still images and building digital repositories, a budget including costs for maintenance of the images over time is mandatory. There will be certified archival repositories that will be able to guarantee the storage of a file for a certain number of years for a certain amount of money.

There are various possible scenarios to be looked at, the main two being payment by file size and payment by traffic/access of the files. While in the first case file size and storage space are not an issue, file size is the main issue in the second case.

Access time also has a big influence on costs. If things need to be available within a few seconds, or on-line, costs will be higher than if access time can be a few minutes or hours.

The proposed cost model for the Harvard University Library Digital Repository (storage will be billed by gigabytes annually) favor approaches that minimize file sizes.

Conclusion

Those responsible for some of the big digital reformatting projects report the same problem: rapid changes in the technology make it difficult to choose the best time to set up a reformatting policy that will not be outdated tomorrow. The lack of communication between the technical field and institutions remains a formidable obstacle. It cannot be emphasized enough that if institutions fail to communicate their needs to the hardware and software industries, they will not get the tools they need for their special applications.

References

1. S. Chapman and W. Comstock, Digital Imaging Production Services at the Harvard College Library, *RLGDigiNews*, **4** (6), <www.rlg.org/preserv/diginews/diginews4-6.html>.
2. *Handbook for Digital Projects; A Management Tool for Preservation and Access*, NEDCC, Andover, MA, September 2000, pp. 21.
3. Recommendations from Safeguarding European Photographic Images for Access (SEPIA) Project, <www.knaw.nl/ecpa/sepia/events/expert2.html>.
4. F. Frey and S. Süsstrunk, Digital Photography—How Long Will It Last? *Proceedings of IEEE ISACAS 2000*, V-113, May 2000.
5. B. Frazer, *Spectra*, **26** (2), pp. 63, (2000).
6. S. Süsstrunk, R. Buckley, and S. Swen, Standard RGB Color Spaces, *Proceedings IS&T/SID 7th Color Imaging Conference*, pp. 127-136 (1999).
7. International Color Consortium (ICC) Specification, <www.color.org>.
8. T. Kohler, ICC Achievements and Challenges, *Proceedings International Colour Management Forum*, University of Derby, pp. 1-6, (1999).
9. *Introduction to Metadata*, Getty Information Institute, Los Angeles, 1998, pp. 1-3.
10. *Data Dictionary—Technical Metadata for Digital Still Images*, NISO Standard Working Draft, July 2000, <www.niso.org/PRImagemeta.html>.
11. DIG35 Specification, Metadata for digital images, <www.digitalimaging.org/i_dig35.html>, (2000).
12. ISO 3664:2000 *Viewing Conditions – Graphic Technology and Photography*.
13. F. Frey, Measuring Image Quality of Digital Masters, and File Formats for Digital Masters, in: *Guides to Quality in Visual Resource Imaging*, <www.rlg.org/visguides>, July 2000.
14. L. Rosenthaler and R. Gschwind, Long Term Preservation and Computers—a Contradiction? *Proceedings ICPS International Congress on Imaging Science*, Antwerp, Belgium, September 7-11, 1998.
15. P. Adelstein, and F. Frey, New Developments on Standards for the Permanence of Electronic Media, *Proceedings PICs*, pp. 127-132 (1998).
16. ANSI/PIMA IT9.23-1998, *Polyester Base Magnetic Tape—Storage Practices.*
17. ANSI/PIMA IT9.25-1998, *Optical Disc Media—Storage*.

Biography

Franziska Frey received her Ph.D. degree in Natural Sciences (Concentration: Imaging Science) from the Swiss Federal Institute of Technology in Zurich, Switzerland in1994. Since 1994 she has worked as a research scientist at the Image Permanence Institute at the Rochester Institute of Technology in Rochester, NY. Her work has primarily focused on establishing guidelines for viewing, scanning, quality control, and archiving of digital images. Her main research topics include the digital reconstruction of photographic images, digital imaging in the visual arts, color photography, electronic photography, and digital archiving.