Digital Archiving Without Preservation Is Just Storage: Education Is The First Step To Achieving Preservation Goals

Sue Kriegsman and Lee Mandell
Harvard University
Cambridge, Massachusetts, USA

Abstract

When imaging science and technology comes up for discussion the topic of digital preservation should be right in the mix. Unfortunately digital preservation is not something enough people have considered as an aspect of emerging imaging technology. Not all data, or digital images in this particular instance, should be saved far into the future but those worth maintaining need more consideration than they are getting now. The problems of digital preservation will have an impact on the development of imaging science and technology, and learning about digital preservation is the first step in being able to achieve preservation goals.

There are four primary communities who will be most affected by digital preservation: industry, digital labs, home consumers and collection managers. Acknowledging the relationships and points of interaction between the communities will facilitate preservation activities. There are basic concepts that can be put into practice now to help maintain digital images and foster ongoing research. All four communities must recognize the pivotal roll education plays in long term digital preservation.

Introduction

Although this is a technology conference, hence the T in IS&T, the issues of digital archiving, preservation, and its related education components, are critical to the success of the technology of digital preservation. Without educating people about digital preservation all attempts for effective related technology will fail.

Long term preservation goes beyond simply burning CDs, storing them off site, and preserving the bits. From the time of the Civil War, photography has been the primary way for visually documenting our lives. Ensuring that the next generation can access our images the way we access our grandparents' photos is essential to continuing this tradition. Analog photographs can be viewed with our eyes and light. Digital still images require a known file format, stable physical media, a mechanism to read the media, an operating system, an application, and compatible hardware. Long term preservation requires that we address all of these issues. The field becomes even more complex with formats other than digital images, such as digital audio, digital video and the "holy grail" of being able to preserve interactive web sites as functioning units.

The first step to addressing these issues is education. Digital still images will be used as a representative example in this discussion. Without education about digital preservation, consumers will not be aware of the need for related products, and producers will not create products to be able to easily access digital images far into the future.

Four Communities

There are four basic groups that will be involved with some form of digital preservation from the creation of digital objects today to accessing them in the future:

1. Industry
2. Digital labs
3. Home consumers
4. Collection managers

Group 1

Industry, in this case, is comprised of hardware companies creating products associated with the proliferation of original digital images such as digital cameras, scanners, and small storage devices. Software companies, such as Adobe, are also a strong player in this category.

Group 2

Digital labs specialize in the conversion of analog (traditional) photographs and images to digital format. Software companies offering on-line image viewing and printing services also fall into this category. Another critical part of this group, which is often overlooked, is the conversion of digital to digital. This type of conversion is necessary for format migration and therefore fundamental to long term preservation.

Group 3

Home consumers do just that: consume. These are the buyers and users of home imaging equipment. This group has high expectations for the long term accessibility of their

images. Most people do not intentionally plan to discard all of their images within 5 or 10 years and therefore expect that their digital images will be ready and waiting for them in the future. This group assumes that having digital images today means they will have the same digital images tomorrow.

Group 4

Collection managers hold a wide array of materials created by industry, digital labs, and consumers. A collection can contain anything such as financial records, death certificates, conference proceedings, personal letters, as well as images. This is the group who are professionally responsible for saving everything. Well, not absolutely everything, but other people *think* they are going to save everything – collection managers are charged with preserving material in perpetuity.

An Example

So why do these four groups need to be educated about digital preservation? It's not because everyone should be sentimentalists and demand the ability to save all images forever. Preservation of personal, professional, and cultural heritage materials is important but some of the need for education comes down to money as a primary driving force. There is no such thing as passive preservation in the digital world and active preservation costs money. Digital materials can not be tucked away in a shoebox in the attic and expected to be available in 20 years. Some would argue that under these same conditions they will not be accessible even in 5 years. In order to maintain digital materials someone (industry or digital labs) is going to have to offer preservation, creation, and format migration services. But these services will not be created or demanded unless everyone involved knows they are necessary for long term preservation.

As mentioned before, preservation is going to cost money. It is far less expensive to do it today than it is to do it in the future. If preservation is ignored now, digital archeologists will be required to undo the damage from neglect. An excellent example of this was The Domesday Project in England.[1] The project conducted in 1986 was established to conceptually replicate and honor the 900th anniversary of the Domesday Book commissioned by King William I (William The Conqueror) in 1086. The goal of the 1986 project was to collect still images, maps, statistics and text of how Britain looked that year. The project involved contributions from BBC project staff, communities, and school children. It was all stored on videodisc. In 2002, the 916 year old Domesday Book was intact but the data on the discs from the 1986 project were in danger of being lost because the brief era of the videodisc had ended. In short, the entire project had to be reverse engineered including looking at hexadecimal editors and data structure books from the 1980s to retrieve and revive the materials collected for the project. The capture and reinstatement of the project data, as well as the look and feel, was a success but it was achieved at great time and expense to those who participated.[2]

This is exactly the kind of work that can be avoided if proper preservation measures are taken at the onset of a project or new technology.

A Food Analogy

"It's easier to bake a pie from a recipe, than it is to make a recipe from a pie. So write it down." – Sam Farber, Founder of OXO Good Grips.

Activities and Decisions

The four communities each have a series of decisions and related actions to take that will lead up to the challenges of preserving digital materials. A portion of the activities and decisions are outlined here. Try to envision where the communities intersect. Although it would be easier to see the connections if this outline was in three-dimensions.

Group 1: Industry

1. Develop software for labs, consumers, managers
    a. Image management systems
    b. Digital repositories
    c. Web publishing
    d. Format migration tools
2. Design hardware for labs, consumers, managers
    a. Digital cameras
    b. Scanners
    c. Printers
    d. Storage media
3. Provide preservation services for everyone
    a. Digital format registry[3]
    b. Migration strategies and planning

Group 2: Digital Labs

1. Acquire digital images from consumers
    a. From analog sources
        i. Scanning
            1. file format
            2. File size
    b. From digital sources
        i. Memory cards
        ii. FTP/web submission
        iii. Email
        iv. CD/DVDs
2. Digital to digital format transfer for consumers
    a. For sharing with other consumers
    b. For long term preservation
3. Make digital images available to consumers
    a. Digital
        i. CD/DVDs
        ii. FTP/web site downloads
        iii. Email
    b. Analog
        i. Snapshots
        ii. Proof sheets
        iii. Archival quality prints
4. Provide preservation services for everyone

a. Image management
b. Storage
c. Format migration
d. Emulation

Group 3: Home Consumers

1. Take pictures
    a. Digital camera
        i. Camera quality
            1. Cost
            2. Image capture size
            3. Lens quality
        ii. Memory cards
        iii. File formats and file size
    b. Film camera
        i. Print film
        ii. Slide film
2. Transfer images to a computer
    a. From a digital camera
        i. Card reader
        ii. Direct cable connection
    b. From a film camera
        i. Home slide or film scanner
            1. File format and size
        ii. Service bureau
            1. File format and size
            2. Delivered by CD/DVDs
            3. Delivered by FTP/web download
3. Organize the collection
    a. Selection – what to keep and what to discard
    b. Image management software
4. Use of digital images
    a. Print – snapshot vs. archival quality
        i. From desktop computer
        ii. Kiosks that accept memory cards
        iii. Service bureau
    b. Share with others
        i. Email
        ii. Hard copy
        iii. CD/DVDs
        iv. Web sites
5. Storage and long term preservation
    a. File formats
    b. File size
    c. Metadata
    d. Media

Group 4: Collections Managers

1. Image management and cataloging
2. Selection: what to keep and what to discard
3. Preservation decisions

Oh What A Tangled Web We Weave[4] or Relationships Between the Identified Groups

In a museum setting the consumers are curators, art historians and the general public. In a family the collections manager will be an energetic family member who wants to see the family's history preserved. In the corporate world there will be librarians, records managers and archivists on staff trying to preserve information. Regardless of who is fulfilling these roles these decisions and activities start to show the connections between these groups. Preservation activities will involve all of the groups to one degree or another.

- Acquiring digital images will involve consumers taking pictures, collection managers cataloging and selecting, and digital labs reformatting, providing storage and producing hard copy.
- A consumer wanting access to collections will use on-line catalogs and image management software to determine what they want. Digital labs will produce, either on demand or ahead of time, digital and analog copies.
- A child will go to the family photo web site, whose content will have been carefully chosen by the family historian and order high quality prints from a digital lab.
- Research into a corporation's history will involve searching catalogs of digital materials and retrieving objects from the company's digital repository.
- As a digital format becomes obsolete collection managers will look through the catalog of materials to determine what is at risk and assess what is worth preserving. Digital labs will provide reformatting services.

Of course all of this will only be possible with the support of hardware, software and services provided by industry.

Nine Things to Learn

Here are a few of the key components that all four identified groups should learn about in order to help preserve their digital objects.

1. Select what should be preserved. With today's digital cameras people are producing more images than ever. There is no hope for preserving all of the images: it would be too expensive and many of the images are not worth persevering. Instead focus preservation efforts and attention on the materials that meet the needs and requirements of the collection. Allow the proper attention to be paid to selected images that should be saved.
2. All media is ephemeral. Think back to paper tape, 8″ floppy disks, 88mb Syquest cartridges and many other forms of digital media that are now coasters at best. It is not hard to imagine that in 20 years CDs and DVDs

could face the same fate. Careful choices for storage media and carefully planned migration off of old media, onto new appropriate media are essential parts of a long term preservation plan.

3. Plan for format obsolescence and migration. Parallel to ephemeral media is the obsolescence of formats. Not many people can still read XYWrite files and even Microsoft Word 1.0 files can't be read by current Microsoft products. Keeping technical information about the file formats and having format migration strategies is also a critical component to preservation.
4. The expectations for quality will change over time. The quality of images continues to increase as technology progresses. A high quality image today will seem like a mere low-res thumbnail tomorrow. A 640x480 image 5 years ago was too big for many computer screens to display. Today that same image looks quite small on a new monitor. Preserving the highest quality image appropriate and staying away from lossy compression schemes is another important consideration for digital preservation.
5. Look around now and then predict the future. Before creating or migrating an image, consider its potential future uses. How materials will be used at a later date will effect how they should be preserved now. There is a difference between images for a web site, images used to print a billboard, or images used to study the grain of the paper in an original photograph. The requirements for each use will determine what is captured now.
6. Don't forget the "M" word: metadata. It's critical to capture as much information about the image in addition to the image itself. Metadata has to be preserved along with the image.
   a. Administrative Metadata covers all of the technical information about images such as file format, machine(s) on which the image was created, image enhancements, color profile, and the source for the digital version. Administrative Metadata has to be preserved along with the image.
   b. Descriptive Metadata is the information about the subject of the object. Descriptive Metadata has to be preserved along with the image.
   c. Structural Metadata defines an image's relationship to other images. For example, a picture in a photo album has to be placed within the album in order to tell a complete story. Structural Metadata has to be preserved along with the image.
7. Open vs. proprietary formats must be considered for digital preservation. Open formats allow for a greater chance of being able to preserve images over a long period of time. The images, and how they are stored, should not be dependent on any one company or organization. Home movies on Betamax tape would have a greater chance of being watched today if they had been shot on 8mm film. But that's not to say that there still isn't a profitable business model in offering open source preservation services.
8. Lossy compression actually degrades digital images. Using lossy compression on images lowers the quality of images because information is discarded from the file and it cannot be retrieved.
9. Specialized repositories are needed for preservation. As the title states, "digital archiving without preservation is just storage." Storing digital objects and images without taking all of these other components into consideration will not help ensure that digital images will be accessible in the future. Digital preservation is not just about storing bits, it is about storing a living object with a lifecycle.

Conclusion

Education both within these four communities, and among these communities, is the first step to the long term preservation of digital objects. Without knowledge about digital preservation, industry will not create the necessary products essential for preservation, digital labs will not offer appropriate services for preservation, and consumers and collection managers will be left with a bucket of undifferentiated bits. We as technology leaders must provide the necessary education to achieve preservation goals.

References

1. Andy Finney, The Domesday Project – November 1986 http://www.domesday.org.uk/
2. Jeffrey Darlington, Andy Finney and Adrian Pearce, Domesday Redux: The rescue of the BBC Domesday Project videodiscs, Airadne Issue 35 (2003) http://www.ariadne.ac.uk/issue36/tna/
3. Stephen L. Abrams and David Seaman, Global Digital Format Registry IS&T Archiving Conference (2004)
4. "Marmion" "O, what a tangled web we weave…" Sir Walter Scott, Marmion, cto. 6, st. (1808)

Biographies

Sue Kriegsman received a B.A. from Alfred University in 1992 and a M.L.S from Simmons College in 1996. She is active with the Society of American Archivists and the current Chair for the Visual Materials Section. Sue is the Digital Library Projects Manager at Harvard University Library where she works with librarians, curators, and archivists as part of the Library Digital Initiative.

Lee Mandell received a B.S. in Computer Science from Northeastern University in 1987. He has spent 15 years working with museums and libraries focusing on visual and archival collections management systems. He also spent 7 years working in the pre-press industry helping develop image manipulation tools as well as managing digital work flows. Currently he is a programmer/analyst for the Harvard University Library, focusing on systems for visual and archival collections. He is also a visual artist focusing on photography and sculpture.